

The Future of the History of Chemical Information

ACS SYMPOSIUM SERIES **1164**

The Future of the History of Chemical Information

Leah R. McEwen, Editor

*Cornell University
Ithaca, New York*

Robert E. Buntrock, Editor

*Buntrock Associates
Orono, Maine*

Sponsored by the
ACS Division of Chemical Information



American Chemical Society, Washington, DC

Distributed in print by Oxford University Press



Library of Congress Cataloging-in-Publication Data

The future of the history of chemical information / Leah R. McEwen, editor, Cornell University, Ithaca, New York, Robert E. Buntrock, editor, Buntrock Associates, Orono, Maine ; sponsored by the ACS Division of Chemical Information.

pages cm. -- (ACS symposium series ; 1164)

Includes bibliographical references and index.

ISBN 978-0-8412-2945-7

1. Chemical literature. 2. Information storage and retrieval systems--Chemistry. I. McEwen, Leah Rae, editor. II. Buntrock, Robert E., editor. III. American Chemical Society. Division of Chemical Information.

QD8.5.F88 2014

025.06'54--dc23

2014026840

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Copyright © 2014 American Chemical Society

Distributed in print by Oxford University Press

All Rights Reserved. Reprographic copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Act is allowed for internal use only, provided that a per-chapter fee of \$40.25 plus \$0.75 per page is paid to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA. Reproduction or reproduction for sale of pages in this book is permitted only under license from ACS. Direct these and other permission requests to ACS Copyright Office, Publications Division, 1155 16th Street, N.W., Washington, DC 20036.

The citation of trade names and/or names of manufacturers in this publication is not to be construed as an endorsement or as approval by ACS of the commercial products or services referenced herein; nor should the mere reference herein to any drawing, specification, chemical process, or other data be regarded as a license or as a conveyance of any right or permission to the holder, reader, or any other person or corporation, to manufacture, reproduce, use, or sell any patented invention or copyrighted work that may in any way be related thereto. Registered names, trademarks, etc., used in this publication, even without specific indication thereof, are not to be considered unprotected by law.

PRINTED IN THE UNITED STATES OF AMERICA

Foreword

The ACS Symposium Series was first published in 1974 to provide a mechanism for publishing symposia quickly in book form. The purpose of the series is to publish timely, comprehensive books developed from the ACS sponsored symposia based on current scientific research. Occasionally, books are developed from symposia sponsored by other organizations when the topic is of keen interest to the chemistry audience.

Before agreeing to publish a book, the proposed table of contents is reviewed for appropriate and comprehensive coverage and for interest to the audience. Some papers may be excluded to better focus the book; others may be added to provide comprehensiveness. When appropriate, overview or introductory chapters are added. Drafts of chapters are peer-reviewed prior to final acceptance or rejection, and manuscripts are prepared in camera-ready format.

As a rule, only original research papers and original review papers are included in the volumes. Verbatim reproductions of previous published papers are not accepted.

ACS Books Department

Editors' Biographies

Leah R. McEwen

Leah McEwen has been the Chemistry Librarian at Cornell University since 1999. Her background is in biochemistry and library science and she is responsible for library resources and specialized services supporting the chemical sciences at Cornell. She has contributed to and served in advisory capacity for a number of information resources including the ACS Style Guide, the ACS CPT Guidelines for Bachelor's Degree Programs, Cornell's VIVO, and CAS' SciFinder. She is an active member of the Chemical Information Division of the American Chemical Society, most recently as Secretary. She is also a member of the ACS HIST and CHED Divisions, and collaborates frequently with the ACS Ethics and Publications Committees.

Presently she is on academic research leave to investigate issues in chemical representation and data management from disparate angles with an eye towards streamlining workflow issues of academic research chemists. She was on fellowship at the Chemical Heritage Foundation in Philadelphia, PA to review archives pertinent to history of machine documentation of chemical information. She is currently collaborating with the Royal Society of Chemistry (RSC) on deposit, discovery and reuse of research data with Cornell researchers. She is also involved in curriculum development for the Cheminformatics OLCC (<http://olcc.ccce.us>) online course delivery system for undergraduates.

Robert E. (Bob) Buntrock

Robert E. (Bob) Buntrock is President of Buntrock Associates. After being well mentored in both organic synthesis and chemical information, he spent 5 years in the lab in pesticide synthesis for Air Products and Chemicals and Amoco Oil. He then transferred to Amoco Corp. where he advanced to Research Associate in Research Information Services where he grew up with the online chemical information industry as a proactive user. In 1995, he and his wife Gloria formed Buntrock Associates providing technical information services for a wide variety of clients. In the last few years, Bob has concentrated on reviewing books for four publications, writing on chemical information topics, and mentoring high schools and college students on alternative careers in chemistry. He has three patents and more than 100 publications and presentations.

As an emeritus member of ACS and the Chemical Information Division, he has been active in both for decades, including as chair of both the Chicago and Maine Sections and CINF and Councilor from the Chicago Section. He is also a member of the CHED, ORGN, COMP, and PROF ACS Divisions.

Preface

Inspired by the opportunities and challenges presented by rapid advances in the fields of retrieval of chemical and other scientific information, several speakers presented at a symposium, The History of the Future of Chemical Information, on Aug. 20, 2012, at the 244th Meeting of the American Chemical Society in Philadelphia, PA. Storage and retrieval is of undeniable value to the conduct of chemical research. The participants believe that past practices in this field have not only contributed to the increasingly rapid evolution of the field but continue to do so, hence the somewhat unusual title. Even with archival access to several of the presentations, we presenters felt that broader access to this information is of value so that an ACS Symposium book would be valuable to chemists of all disciplines.

The past is a moving target depending on the vagaries of technology, economics, politics and how researchers and professionals choose to build on it. The aim of this collection is to critically examine trajectories in chemistry, information and communication as determined by the authors in the light of current and possible future practices of the chemical information profession. Along with some additional areas primarily related to present and future directions, this book contains most of the topics covered in the meeting symposium. Most of the original authors agreed to write chapters for this book. Much of the historical and even current material is scattered throughout the literature so the authors strived to gather this information into a discrete source. Faced with the rapid evolution of such aspects as mobile access to information, cloud computing, and public resource production, we hope that this book will be not only of interest but provide valuable insight to this rapidly evolving field not only to practitioners within the field of chemical information but also to chemists everywhere whose need for current and accurate information on chemistry and related fields is increasingly important.

The editors would like to thank all of the original speakers, the sponsoring technical divisions of the symposium, CINF and HIST, and our symposium co-organizer, Andrea Twiss-Brooks, for their contributions to the stimulating discussion that inspired this volume. Presentation titles, abstracts and slides are listed at: <http://bulletin.acscinf.org/node/347>. We would also like to acknowledge the patience and support of the ACS books staff throughout the project, as well as the many reviewers. We are especially grateful to the authors for their willingness to reflect on these collective issues of our profession beyond the regular course of their individual work for the benefit of the broader chemistry and scientific information audiences. In recognition of the sometimes personal nature of these

pieces, we have preserved the original spellings provided by the authors whenever possible.

On the cover: Cover design inspired by the original presentation “Historical Cantilevering”, given by Peter F. Rusch, who is an author of a chapter in this volume. The punch card image is from the original presentation by Engelbert Zass, also an author with a chapter in this volume. The phone image was taken from Chapter 14 by Alex M. Clark. Photo of the bridge by brewbooks; Cantilever bridge construction - Sound Transit; <https://www.flickr.com/photos/brewbooks/394851251/>; Creative Commons license <http://creativecommons.org/licenses/by/4.0/>.

Leah McEwen and Robert Buntrock, co-editors

Leah R. McEwen

Cornell University
283 Clark Hall
Ithaca, NY 14853
607-793-6217 (telephone)
lrml@cornell.edu (e-mail)

Robert E. Buntrock

Buntrock Associates
16 Willow Drive
Orono, ME 04473
207-866-7930 (telephone)
buntrock16@roadrunner.com (e-mail)

Chapter 1

Taking a Long View: Traverses of 21st Century Chemical Information Stewardship

Leah McEwen*

Cornell University, 283 Clark Hall, Ithaca, New York 14850

*E-mail: lrm1@cornell.edu

The introduction of the Internet into the publication environment has greatly increased the breadth of concerns around stewardship of information. Not only are research libraries dealing with an overall expansion of more traditional scholarly publication genres, an unprecedented number of other information venues are focusing attention on networking pre-published research data. In addition to communication of the latest ideas, significant value lies in appreciating both the super- and substructures emergent in the vast knowledge bank of chemical research. Happily we don't have to reinvent too many wheels to leverage this information as the discipline has constructed itself around systematic organizing principles. Translating the value of these structures into digital utilities and engaging the broader community of research chemists and students is the work of today's information stewards, much as it has been over the course of the chemical information profession. Now more than ever is the worth of such stewardship apparent in the wake of blossoming information opportunity and resource conservation.

Introduction

At the turn of the last century, when I started working as a chemistry librarian in 1999, I had a pretty good notion that most of my time would be spent online. Full text of current issues of journals was becoming available in critical mass,

adding the next step in the online information cycle to accompany searchable indices that had been available via federated systems for some decades. Directly accessible primary literature en masse at point of need for those who consume it reflected an increase in the storage media of the electronic information industry, and more impressively the opportunity for more end-user friendly interface development options in the advent of the Internet. This rather revolutionary convenience completed the cycle of online information transfer for most scientists, or at least seemed to meet the vast majority of the need.

Fast forward through a decade of growing pains on the part of scientists, publishers, libraries and information technologists, and we arrive at a point where most of the journals were available in user-friendly and manager-feasible electronic form over full publication history, and usage measured in full text downloads was skyrocketing. Monographs also began transitioning online and series, textbooks and data compilations were emerging via odd hybrids of content management systems and individual interfaces. All this, the promise of Google, and the rapidly growing and globalizing Internet could be had for a price. About this time, developed economies worldwide began to strain due to over-anticipated growth, tighter budgets and the consideration of information as a commodity. Research institutions responded by reducing investment in libraries and other support services.

The Internet has brought much more than convenience, turning upside-down the information transfer industry and with it the business practices of conducting the science that it supports. Access, copy-of-record, return-on-investment and information literacy are all concepts being rethought by stakeholders in the value chain of published scholarly literature from scientists, through publishers, system developers, libraries, and back to scientists. With freedom from old business models comes new responsibility, and this need holds true in an online environment as much as it did in a hardcopy environment. Can prior principles of scientific communication transfer? Have changes in the online environment impacted science to the extent that the information cycle has changed? Or, is it more the rules of business that have changed? More critically from the perspective of a scholarly information steward, what are the broad impacts of recent changes on the utility of the information and the experience of the users? Can we ascertain a trajectory of chemical literature and information practices by gauging the future against the past?

To gain some perspective on the challenges of today, I am interested in past challenges of the people involved in chemical information transfer. This interest prompted the organization of a symposium at the American Chemical Society Meeting in Philadelphia in 2012 on the Future of the History of Chemical Information. The experience represented by the speakers spanned the chemical information timeline from early implementation of computerized information systems through more recent opportunities and challenges of networked data. The sense of concern that arose out of the discussions there further prompted this symposium volume which pulls together an even broader range of perspectives from information professionals who are or have in their careers tackled difficult problems of translating the essential information of chemistry through technical revolutions. Studying the advent of machine documentation presents an

opportunity to understand dichotomies inherent between the online interfaces and the chemistry being covered (implicit/explicit, human/computer).

The Professionalization of Chemical Information

It would be an understatement to say that computing technologies have merely improved access to chemical information. The impact on chemical research has made a profession of chemical information management. The Division of Chemical Information of the American Chemical Society formed in 1948 as the Division of Chemical Literature, and has hosted many passionate debates over the years ranging from standards of coding to pedagogy (1). In his eloquent 2007 review of the previous 50 years of chem(o)informatics research, Willet references Herman Skolnik's criteria of what defines a discipline: active researchers, research forums, research journals and specialized education (2). A rigorous consideration of the development of this discipline along all of these dimensions can help us understand the core principles underlying its strengths and limitations.

Willet and others have documented the history of chemical information from an informatics perspective, generally as far back as the early 1960s with the founding of the Journal of Chemical Documentation (2–4); or from a teaching perspective with a few notable references before the founding of the Chemical Literature Group in the ACS Division of Chemical Education in 1943 (5). There is very little historical treatment of this problem that considers the impact of machine documentation on the experiences of *practicing chemists* during this transformational period. Most histories of chemistry literature and communication predate the computer era (6). Most of the historical considerations originating in the chemical information field understandably focus primarily on technical developments and highlight successes (7).

This focus on chemical information history distinct from the larger chemistry discipline is probably indicative of the professional interests of those substantially employed in either information or history studies. A comment in the introduction to a 2002 Conference on The History and Heritage of Scientific and Technological Information Systems at the Chemical Heritage Foundation emphasizes this sense of separation in consideration of the *technology* of information: "From being a kind of special tool used as an *adjunct* to the creative, substantive conduct of science, information technology and systems has assumed a central role in the actual constitution of a number of scientific disciplines that have been given such eponymic designations as biomedical informatics and chemical informatics" ((8), p. 6, emphases are mine).

What of the role of information technology in such an information-rich and interdependent discipline as chemistry? Have the systems really only recently assumed central roles, as recently as the modern computing machine, or do the principles of informatics trace back farther in chemistry? Certainly chemical information is inherent to the discipline and managing it in systematic and increasingly automatic ways has been part and parcel of the practice for time immemorial. Discipline-based development of systematic approaches that reflect

scientific inquiry date back centuries to debates over chemical nomenclature, symbolism and reactivity, were inspired by even longer histories of the empirically based data-generating methodologies of alchemy (9).

I don't think we have to look far from the current practices in managing chemical information to find a frame that distinguishes our history, and in doing so, shows that our opportunities in this field are firmly rooted within chemistry practice. Rayward, a noted information science scholar, struggles with an appropriate approach of study for information that is not quite a thing- "a word, a concept, encapsulated, represented, embodied"; "a process or a product... text or document... content... expression of meaning... process of symbol representation and manipulation of electronic machines..." ((8), p. 4). By his own admission he is unsatisfied with these articulations and settles on "the most useful *modus operandi* for the historical study of information, it seems to me, and what is implicitly or actually explicit in the discussions above, involves some notion of system, of the creation and use of what I call information infrastructure without which in its varied historical manifestations societies (or telecommunications engineers or neurophysiologists) could not function" ((8), p. 4).

There it is- systematic information infrastructure that allows the operators, human and machine, to take some action on the information. Who better to articulate the characteristics of this enabling infrastructure in chemical information than those operators whose professional work is to nuance and facilitate action upon it, from industrial R&D and expert systems development to academic discovery support and chemical education. While most chemical information professionals are not also professionally active in history studies, we can offer some unique perspectives intermediate between the evolving technologies of information and the scientific practices of chemistry. Chemical information infrastructure has evolved around two unique types of information, chemical structures and chemical reactions, which distinguish it from generalizable approaches in information science (7). Representation, notation, provenance, metadata and other documentation practices around the organizational motifs of chemical structure and reaction are recurrent themes throughout this volume.

Wither the History, Whither the Future of Chemical Information?

Driven by my own journey through the transition of hardcopy research library ecosystems into electronic information workflows, I applied for a sabbatical and embarked on a search for more conversations around human-machine communication in a chemical context (10). An exemplar conversation surfaced in the archives of the well-known organization that has been developing and promoting chemical standards for almost a century: the International Union of Pure and Applied Chemistry (IUPAC) (11). In the late 1960s it was becoming apparent to IUPAC functionaries that information critical to chemical research had expanded beyond human indexing and finding capability. An Interdivisional Committee on Machine Documentation was formed to pursue the holy grail of bridging human and machine processing – "a unique definition of

chemical structure which is understandable on the printed page and yet logical, unambiguous to a computer program... universally applicable and can be readily understood by both processors and users,” ((12), emphases are mine). The membership included several chemical information luminaries (13), who engaged in seemingly spirited and at times quite insightful debate along various axes of this problem. However, this committee generated no concrete solutions or rules such as those, however complex, dutifully produced by the nomenclature committees, and the effort was quickly wrapped up.

Committees are not always the best venues for accomplishing revolutionary movement and this attempt ultimately bogged down in the variability of implicit needs and explicit requirements among different systems, both machine and human focused. Interestingly, the tension was not over any particular human-machine differences. Challenges arose over interpretation of when and why these distinctions are important and the collective responsibilities of the fledgling chemical information field to coordinate, collaborate and communicate such requirements with the larger community of chemical researchers; in other words, the business parts of the problem. Recent IUPAC efforts have more successfully reframed the problem within the terms of more tractable projects, including the development of the “Graphical representation standards for chemical structure diagrams (IUPAC Recommendations 2008)” (14), and the “International Chemical Identifier (InChI)” (15). It is worth noting that neither of these published specifications is attempting to be comprehensive or in perfect alignment with machine and human-appreciable definitions and state their primary purposes accordingly.

Unambiguous representation of spatial structure and gathering spatially similar clusters of data are very important to chemistry research. They underlie both human and computer approaches to chemistry data management and require managing what is implicit vs. explicit carefully. Repetitive patterns are explicitly defined while iterative adjustments are implicitly determined. Computing machines excel at keeping track of things and automation is essentially a matter of scale and accuracy, i.e. tracking more consistently and quickly and relieving humans of the direct burden of repetitive and iterative actions over scale. Rules were spatially articulated with pre-electronic technologies such as punch card machines (16). Computer models are numerically articulated, both approaches well suited for topological expressions of chemical structure inherent in the language of chemistry (7, 17).

Variability in the development of the rules and criteria underlying explicit definitions and implicit determinations as well as how and when each are invoked in a system can enable different types of application, as advocated by the various members of the IUPAC Machine Documentation Committee. Specifications of patterns and adjustments in literature indexing databases such as those provided by the Chemical Abstracts Service (CAS) are based on repeatability and streamlining of earlier manual indexing techniques employing nomenclature rules (18). Structural motifs such as ring scaffolds are explicitly defined in these approaches by rules for numbering of atoms, with likely some variability among treatment of functional groups depending on what is to be highlighted. Specialized representations of chemical structures called Markush structures

incorporate generic functional group notation and are used in patents to implicitly cover large families of molecules (19). Systems built on predictive models such as DARC (Documentation and Automated Research of Correlations) explicitly define atom properties and implicitly determine molecular structure through correlation (20).

Variability in systems has indeed added possibilities for applying the principles of chemical structure representation, and has subsequently created myriad challenges downstream for chemists moving between systems due to lack of transparency in underlying rules and assumptions (21). Further tensions arise between systems managers over different approaches to crosswalks- should rules be systematized (explicit approach), or best practices developed for exceptions (i.e. approaches for implying most useful solutions). IUPAC has approached this by developing rules for “preferred IUPAC names” to support use of a common language in legal and regulatory scenarios, and principles (including unambiguity) for guiding use of alternatives for diverse applications in daily use (22). Additional factors impacting overall information processing include data sources, access, responsibility, and many other familiar non-technical issues juggled by information professionals daily. All of these concerns, technical and otherwise, require attention to enable a functional work environment for chemical research and should be captured and described in the provenance of data and metadata structures to support problem solving along the full cycle of information transfer.

These challenges are not exclusive to the advent of the computer in managing information. Conversations with historians of science studying the systematization of chemical nomenclature suggest similar ongoing tensions arising around the codification and adoption of nomenclature rules from the 1800s (23). The struggle appears to lie in the impact of systematization on the usability of representation for communication (as a type of use) or larger scale indexing (as a type of handling). Infrastructure and decisions associated with *handling* by information purveyors vs. *use* of information by practicing chemists is an additional layer of information management that needs to be considered (24). Given the importance of unambiguity, it is critical for the chemists using these systems to appreciate the underlying approaches for establishing canonical representation and streamlining automatable processes, whether manually created or automatically generated, or as Currano eloquently expresses to her students, “think like a database” (25).

Information Eras and the Continuum of Transition

The amount of chemical information and level of detail far exceeds the capacity of linear or chronological indices. Aside from bibliographic referencing, indexing by compound has long been of prime interest. Various approaches encode compounds systematically, focusing on specific rules such as nomenclature and atom-bond connectivity, or common motifs such as functional groups and topology, rather than on individual instances. Such methods stretch back to the

development of systematic nomenclature rules at the turn of 20th century for humans to better manage indexing of more substances reported after the advent of the dye industry (23). One hundred years of development suggests that this has been a fruitful idea. Chemical structural notation is an important motif in chemistry and has dominated the focus of chemical information development, both in indexing and informatics. As the workflow environment becomes more enabled for end-users, it is imperative to introduce the broader technical power of this language back into the hands of practicing research chemists.

Chemical information has figured prominently online since early days of machine documentation (1). Chemistry research was an early adopter in both searchable indices such as STN, and later full text journals (26). As a current member of CINF, I have been hearing at division meetings for some time that chemistry is ready for digital information, content management for computing purposes, not just human convenience. Electronic information involves computerized representations of human-readable information, able to be transferred electronically and reconstructed for other human readers. Digital information is constructed in a digital environment; its semantic content is computer-readable, discrete and explicit, and the computer can not only transfer this information but also aggregate it and analyze it in chemically meaningful ways beyond digital object management and basic statistical analysis of incidence. A recent example of this distinction might be the use of fax machines to electronically transport hardcopy where the computer only understands the document at the level of rasterized dots on a page versus current word processor documents where the native file format contains some level of markup in the form of words organized via punctuation into sentences, paragraphs, pages, etc. In chemical information, the different ends of the scale might be saving of sketches of a molecule as purely graphical images to cut and paste into manuscripts versus the underlying connection tables of atoms and bonds that are used internally by a drawing program, and encapsulate a significant amount of chemical meaning that can be transmitted to other chemically aware software (7).

To help frame the transition from electronic to digital information, it might be helpful to reflect on the transition of chemical information flow from print to electronic environments. In the print era, the push for chemical information was focused on documentation to keep track of what was done and what happened in long series' of shots in the dark and increasing optimization exercises. Broadly speaking, early publications clustered around topics, including such literature forms as treatises, textbooks, and encyclopedic chemical dictionaries. Primary publication of research results shifted to documenting research society and institutional transactions, and subsequently society and national journals with some eventually being absorbed by commercial publishing houses specializing in scientific communication. With increasing accumulation of reported substance characterization and methodology, research and teaching chemists further specialized indexing around the most relevant aspects, chemical structures and reactions. Secondary data compilations such as the Gmelin and Beilstein Handbooks and the Houben-Weyl standard methods reference became established as core tools of chemistry research practices to meet the need for high quality, repeatable, protocols and methods (6).

Such documentation of content was increasingly critical for successful research and training as the corpus of chemical information continued to expand. Hence the profession of literature chemists emerged who were trained as chemists and applied this knowledge to the production, aggregation, organization and distribution of chemical information. They produced abstracts and indices of abstracts. They acquired and reviewed volumes of research ephemera, primary published literature and secondary indexing sources. They taught research chemists about the organizational structure of these collective sources and developed further tertiary guides on how to maximize the utility of the ever-increasing magnitude and variety of information bits. Machine documentation of chemical information began in earnest after the Second World War and morphed into the electronic information era soon after. The push this time was access, towards more complete and more detailed access to the rapidly growing corpus of information. From the mid 1960s through the 1990s was a golden era for chemical information professionals, engaging in an ever-broadening array of activities related to translating hardcopy content into the automated processing environment of computers and back out (more or less) for the use of human chemists (27, 28).

With the advent of the Internet, where data can pass in native formats through computer and cloud networks directly, and simultaneously to humans not connected in physical space and time, we are entering the era of digital information. The push here is towards application, re-use and mash-up of empirical data in broader applications. This is the era of the semantic networking of data in order to facilitate better discovery of related data and to find linkages that result in a whole that is greater than the sum of its individual parts. Well-handled metadata and data structure can indicate that two packets of bits are related and enable them to be compared in a semantically meaningful way. The trick is delineating which data types and relationship attributes among the metadata need to be made explicit in order for the system to yield useful connections. The stewards of this movement are emerging under the banner of informaticians; magicians of parsing chemistry representation and machine-human translation.

The digital information era is in relatively early days and the long-term challenges for stewardship are not fully apparent yet. Indeed, the majority of chemical information professionals are involved with the still critical work of the previous eras. The scientific nature of the underlying information seems much the same and considerations of stewardship responsibilities can be initially based on several recurrent themes. What are the secondary structures needed to help manage, organize, find and reuse information for informatics activities, other computing analyses, researcher interpretations, etc.? What tasks will be necessary to ensure provenance and archiving of data collections increasingly captured and destined for re-use? What training will be necessary to convince research chemists to adopt practices in their workflow in order to facilitate direct re-use of data in an environment of increasingly diverse data and data-analysis tools? Finally, what pedagogical directions involving information will set the course of digitally enabled chemistry research itself going forward? Harkening back to earlier days of professionalizing documentation chemists, it could be argued that these responsibilities still center on documentation, but the focus of

documentation and validation has expanded beyond completed research. We are now focused on the entire life cycle and the myriad styles of data, establishing and documenting shared and transparent practices for stewardship that ranges from experimental planning and data capture to storage, access, and computational re-use.

Provenance: The Science and Poetry of Documentation

When communicating essential aspects of chemical information to librarian colleagues who work in other subject areas, three more general information science topics can be used as a framework: natural language processing, big data and the focus on education. Structural formulae are an essential language through which chemists communicate and evaluate their scientific claims (17). They are replete with ‘parts of speech’ and ‘rules of grammar’ in atoms, atom attributes, bonds, bond orders and other topological and geometrical notations. ‘Natural’ language processing on this language of chemistry has been going on for a very long time. In the early 1890s, prominent chemists engaged in codifying systematic organic nomenclature sought to ascertain patterns emerging from representations of chemical structures in order to classify and order molecules for further study (23). The next stop was punch card notation, again attempting to determine and translate useful patterns for classifying compounds. Subsequent efforts focused on developing binary connection tables and graph-based analysis that were then followed by statistical analyses of structure-activity relationships (7). The latest efforts have focused on semantic based mining of chemical structures, methodologies and processes (4, 40, 42). All of these approaches developed classification schema to facilitate inferences based on structural formulations, case-by-case as in individual lab or teaching scenarios and systematically as used for indexing and large-scale screening. It is worth noting that such approaches arose from within and have been taken seriously by the discipline itself long before computing machines, which have been extensively employed subsequently to ease the analysis and processing workload.

Also inherent to the practice of chemistry is the collection of data. The study of chemistry is focused on synthesis, analysis, and further application. All of these activities generate data of interest and early chemists and alchemists were meticulous note-takers (9). Quantitative, systematic and critically reviewed approaches to data collection have been in practice for some 200+ years on over 88+ million substances to date (29). The types of data range through extensive characterization, material properties and toxicity measurements. This is the stuff of big data, systematically collected, ordered and re-purposed within the normal course of discipline practice long before current approaches using correlative analysis. The metadata that has subsequently grown up around this data, originating primarily with practicing chemists in a variety of settings from teaching to industrial optimization and further systematized by reference compilation and standards development, is staggering. Over 300 separate fields were coded into the CrossFire version of the Gmelin Handbook of Inorganic Chemistry for example, in compilation for over 200 years, including such

specialized fields as multi-center ligand formulae and catalyst classifications (30). These metadata schema are complex and sophisticated, with layers of classification and dependencies, and require extensive guides and decision trees for navigating the printed version and command searching options for the online version (31). Metadata has become the language of meta – professionals, the literature chemists. When they consulted these reference works primarily in print, chemists had to be meta-experts, too. Understanding the manner in which their data was organized helped them understand their chemistry. Focusing on how digital metadata is created and functions could provide a means for re-engaging the meta-expert in every chemist.

A deeper knowledge and understanding of the context of chemical data enables the broadest re-use of this data, and the molecules to which it pertains, across the broad terrain of the chemical sciences, including areas with high social impact such as biomedicine, materials science, and environmental science. Cheminformatics techniques have exploited the metadata that already exists to articulate when there is an intriguing chemical story, the level of confidence of (parts of) methodologies, and the refinement of underlying assumptions through the application of scientific knowledge. Expanding the scope and depth of the documentation process furthers the potential of these approaches and the value of the source data and work of chemistry. At the most basic level, provenance documentation considers datasets as artifacts, supporting linking to associated publications for scientific context through data citation infrastructure (32). With a use-driven approach, provenance might take on the form of a family tree over the course of research associated with the data through subsequent re-uses (33). A full curriculum vita documenting the process of the original experiment and subsequent analyses could support better parsing of minimally reported methodology languishing in journal articles, and make it more mobile and amenable to aggregation along with the data (34). Purposeful capture of metadata and other notation can support richer scientific debates with as much empirically derived information as available to help adjudicate them. Without documenting and exploiting in-depth provenance, do we risk coming full circle to a sort of modern alchemy, clumsily trying to find gold in combinations of vastly increasing accumulations of common data?

In an abstract sense, expression of data within their original experimental contexts is medium agnostic: values can be recorded by hand in paper laboratory notebooks or streamed directly from instruments to networked systems. However, the nature of data is not measurement- method agnostic. A methodological context is needed to determine how much processing has been performed on the data: initial screen vs. analysis vs. publication and communication. The National Institute of Standards and Technology asserts that while “reliability may be the most fundamental concern encountered in any application of materials property data,” *purpose* and *use* are the critical dependencies and involve both quantitative and non-quantitative criteria to ascertain quality and establish validity (35). Add to this analysis the comparison of repeatability across scattered reports enabled by compilation of data, as long practiced in the chemical sciences and industry at great expense, and process documentation becomes an essential component of data capture. Just as we need to document data processing, we must also provide

documentation that enables readers to understand how much the data may have been altered from their original form. The absence of such documentation creates ambiguity that can degrade the value of chemical data to the point of uselessness for future re-use.

One of my favorite illustrations of the critical position that data compilation holds in chemical practice and the equally critical problem posed by inadequate representation is captured by the timeless film, the *Anatomy of Data* (36). The story bridges print and electronic information technologies as well as research and application perspectives, the work is daunting and one is left yearning for more data reporting standards, ideally digitally enabled. Data formats that incorporate several layers of metadata to track specific instrumentation, calculations, process conditions and even experimental rationale, can improve scientifically meaningful comparison by both automated and human processes. Such formats are becoming increasingly available, notably for crystallographic structures, various spectroscopic characterizations and thermodynamic properties. While some have been widely adopted, such as the crystallographic information file (CIF) data format and information framework (37), the use of other digital data formats such as the IUPAC JCAMP format for IR and NMR spectra have not captured the attention of practicing chemists in spite of the wide use of these characterization techniques (38).

Throughout the transition from handling print hardcopy to electronic chemical information, and now to digital data in just over a human career span, lack of transparency to researchers of the data manipulation and information decisions made by online systems is of increasing concern. The ‘more convenient’ the user interface, the harder it is for the scholar or professional to consider and intellectually examine what are really still open questions about their research (24, 25). When humans did the organizational processing work, the information systems had to make sense to chemistry-trained humans, e.g. literature chemists, and were extensively and explicitly documented. It was easier then for the human chemist and their support professionals to have a sense of how the system was structured and organized and how to leverage them, even if it required more legwork up front (27, 28). As more of the organization is handled through automated processes by a computer, it may not be as readily apparent for a human user to translate technical terms, especially if documentation is much less accessible and explicit. Imagine this proverbial scene of past, disassembling the intricate workings of a valuable pocket watch without sketching how it all fit together. Documenting what happens to data in automated systems is as important to the practice of craft-based trades such as chemistry as what happens to data expressed in figures of peer reviewed articles. Additional attention to curation of data capture, documentation and ordering principles can help address issues of due process and improve confidence in “black-box” information systems.

As a methods-based science, with strong discipline focus on communication, design, analysis and evaluation as discussed above, there is in tandem a strong emphasis on pedagogy. Undergraduates learn about such information and data management building blocks as language (structure and nomenclature), classification, stoichiometry and analysis. Graduate level mentorship generally runs towards refining methodology, technique and judgment, including leveraging

chemical metadata structures for experimental design. I have heard many expressions from students noting with interest the insights than can come with “rebooting the information structure” as they move between search systems. Without more transparency in digital information, generations of chemists using online systems will be emerging without the benefit of the experience that was gained previously from manual searches, which coincided with their (primarily) manual lab experience.

Tensions or Opportunities?

The question of more interactive engagement of chemists about digital information has garnered a great deal of discussion in several areas of the greater chemical information community among cheminformatics, resource providers, and librarians. How do we make our activities look like chemistry, building on the structures that chemists know, and show where and how these concerns fit into the research cycle beyond publishing articles? The more I study this debate, the more it appears to me as multiple facets of our long-term conversation about how to manage chemistry research data based on organizing principles that make logical sense in the chemical space, to humans and computers alike. When I think of particular strengths that we have built up professionally over our history, distant and recent, to manage these challenges, I think of the compilation of content characterizing chemical substances and reactivity and the organizing principles of topology and geometry. Looking forward, critical control points for chemists and professionals managing information are emerging through issues of accessibility, mobile workflows and the semantic web (39–42). These are fundamental questions in the chemical information profession that crop up among conversations of our peers and further in the pages of this book. Below are some introductory reflections from the perspective of an academic chemistry librarian on sabbatical. I have taken a purposely general and broad tone to let our colleagues in the chapters that follow speak more poignantly.

When I think of accessibility, I think of a community defined by responsibility for content. Compiling data; reviewing scholarly assembly (articles, dissertations, tenure dossiers); searching for, finding, parsing, and identifying connections; browsing and planning; mixing and matching methodologies; noting observations; and optimizing processes are all valued aspects of chemical synthesis research. Anyone who engages with scientifically derived content in any such fashion in a professional capacity is participating in the scientific enterprise and bears a measure of responsibility to the chemistry collective both to process data in some chemically intelligent manner and to leave them in at least as scientifically robust form as they found them. The long-standing community approach to meeting this responsibility is through documentation of process and subsequent publishing of this metadata, attested by the massive collections of scientific research libraries. While it is temptingly easy to concern oneself with greater potential for availability via the Internet and subsequently become that much more frustrated with barriers to current online systems, these challenges should

not distract us from our greater community responsibility to accessibility of and engagement in provenance documentation.

When I think of mobile workflows, I think of the power of modularity. Whatever our art or craft, method must manifest itself in discrete chronological tasks; so much the better in science where we are required to present and defend our process on scientific terms. Modularity reminds us to think through the merit of each step and capture the process. What could be considered limitations of mobile platforms- small on-board storage and small interaction screens- have forced increased transactions with cloud services, more discrete interactions with users and lower tolerance for sloppy data representation for both computer and humans, with the happy accident that users have more opportunity not only to know where they are in their overall process but also more opportunity to engage in step-wise decisions with their data, within or between apps. This ultimate usability model is compellingly packaged and data dutifully logged and captured to be potentially available for further transactions.

When I think of the semantic web, I think of the adaptive potential of pairing native human intellect and logic with the repetitive consistency of the computer. We are generating observations about our world at a rapid pace and are ever in need of catching up with useful interpretive structure. To date, the computer has primarily played a crucial but somewhat ancillary function of faithfully logging the communication of our research bounty in essentially flat-file form. The simple relationship format that underlies the networked structures of semantic approaches reflects a tractable compromise between implicit formulations of intelligent scientists and professionals, and explicit rule-driven computational systems. Divvying up roles between human interests and computer capabilities enables more flexible and extensible application to complex real-world problems. Formulating the core defining concepts of a scientific approach forces greater clarity of process on the part of the humans, putting the opportunity and responsibility of ascertaining meaningful relevance back in the hands of scientists and improving the overall documentation identified and captured by the computer.

When I think of pedagogy, I think of setting future courses of application and thus the methodology of a discipline. Framing structure around abstract ideas and articulating step-wise through complex methodologies built up over generations for the benefit of students causes us to reflect logically on habitual practices. This is one of the few situations where we must take the time to rework out loud within the greater knowledge of the day, the technical and scientific merit of our own underlying assumptions and those of the discipline as a whole. Through the explanations and scenarios we conceive and deposit with the students in the pipeline, educators are seeding the trajectory of further practice. Doing this through the efforts of individuals, teachers and their students, maintains maximum opportunity for creativity along the evolutionary course of a discipline (43). Research apprenticeships further engage the direct contribution of students in real time; jump starting them into the practice, community participation and collective responsibility of chemistry research. We have the opportunity to expand this model to documentation practices, encouraging more engagement of practicing scientists in establishing the provenance of their work and the data that they generate and use.

Stewardship & the Long View of Chemical Information

We are collectively responsible for the stewardship of the scholarly output of chemical research. Chemical information professionals encompass a richly diverse group of job duties and responsibilities across a variety of sectors. The diversity of this group of authors reminds me of the diversity within a chemistry department, and among chemistry related research labs across a university. The chemical information profession demands of its practitioners a multi-faceted appreciation of the underlying complex art and craft based science of chemistry. Our work has evolved around service and research focused on a context dependent discipline. And our stewardship concerns range across the people-users, the computing developments, and all the other issues in-between as well as the source material and formats. What do the data want? What do the computers want? What do the people want? As a profession we find ourselves in a veritable log-jam of ideas and activities. In the rush from the 'print' era to the 'electronic' and yet again to the 'digital' the object is to not fall too far back in a cycle of reinvention. We must refresh our technological bathwater without tossing out the baby of established organizational techniques that have facilitated innovative science. Our challenge is to bring our exploratory-based activities to a professional service level while still maintaining our appreciation of the scholarly perspective.

Chemists' documentation, the unifying and particular focus of the chemical information profession, is part and parcel of chemical practice. As stewards and scientists, we know that documentation of process is more than just good housekeeping and communication. Good notes enable immediate detailed focus on the chemical reactivity that defines the discipline. The methodological emphasis has cumulative value that impacts the overall quality of the practice at the inductive as well as deductive levels. A native data-driven approach codified into practice by 18th century scientists has resulted in precision measurement, rational nomenclature and stoichiometry, principles underlying chemical research and informatics today. Chemical information is valuable information indeed and one might consider it a boon to have such a diversity of attention paid to its care and feeding. This is certainly why librarians badger so much about quality, and why collectively the discipline needs provenance-badgers along with info-magicians to keep the data handling viable and robust as the scales and stakes continue their exponential growth. As quoted in the memorial of a respected colleague, "...In the past, trusting people might have been a necessary evil [of research]," Bradley said. "Today, it is a choice. Optimally, trust should have no place in science" (44).

Thus I close this missive with a moral for my valuable colleagues and myself in a venerable profession. If I have seemed in this reflection to obsess over what might at times be considered a secondary function of provenance tracking, I am reminded of the ancient saying – "the fox knows many things, but the hedgehog knows one big thing". For academic libraries it has always come down to accessibility of scholarly information in the broadest terms, from the past, in the present and for the future. The hedgehog's wisdom lies in passive resistance. To avoid a similar posture of perennial defense in the face of continuous change, it is the prerogative of chemistry librarians to steward the practice of documentation

into the digital era. Documenting provenance is essentially overseeing the documentation of stewardship in which we are all engaged professionally. In other words, practicing the good chemical information hygiene that we preach.

Acknowledgments

The author would like to thank the Cornell University Library and the Chemical Heritage Foundation for the research opportunities, and Evan Hepler-Smith and the reviewers for detailed feedback on the articulation and readability.

References

1. Metanomski, V. M. *50 Years of Chemical Information in the American Chemical Society 1943–1993*; American Chemical Society Division of Chemical Information: 2003. <http://acscinf.org/CINF50/> (accessed May 29, 2014).
2. Willet, P. Chemoinformatics: a history. *WIREs Comput. Mol. Sci.* **2011**, *1*, 46–56; DOI: 10.1002/wcms.1 (accessed May 29, 2014).
3. Warr, W. A. Some Trends in Chem(o)informatics. Chemoinformatics and computational chemical biology. *Methods Mol. Biol.* **2011**, 672, 1–37; DOI: 10.1007/978-1-60761-839-3_1 (accessed May 29, 2104).
4. Chen, W. L. Chemoinformatics: past, present and future. *J. Chem. Inf. Model.* **2006**, *46*, 2230–2255; DOI: 10.1021/ci060016u (accessed May 29, 2014).
5. Kozlowski, A. W. Evolution of chemical information instruction. Chemical Information Bulletin, Technical Program, 244th National Meeting of the American Chemical Society, Philadelphia, PA, Aug. 19–23, 2012; American Chemical Society Division of Chemical Information, 2012; CINF 62. <http://bulletin.acscinf.org/node/347#S62> (accessed May 29, 2014).
6. Mellon, M. G. *Chemical Publications: Their Nature and Use*; 4th ed.; McGraw-Hill Book Co. Inc.: New York, 1965.
7. Warr, W. A. Representation of chemical structures. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 557–579; DOI: 10.1002/wcms.36 (accessed May 29, 2104).
8. Rayward, W. B. Scientific and Technological Information Systems in Their Many Contexts. In *The History and Heritage of Scientific and Technological Information Systems: proceedings of the 2002 conference*; Chemical Heritage Foundation, Philadelphia, PA; Rayward, W. B., Bowden, M. E., Eds.; Information Today, Inc.: Medford, NJ, 2004; pp 1–11.
9. Grethe, G. The History of Chemical Reactions Information, Past, Present and Future. In *The Future of the History of Chemical Information*; McEwen, L., Buntrock, R. E., Eds.; ACS Symposium Series 1164; American Chemical Society: Washington, DC, 2014; Chapter 6.
10. *Otlet Fellow, Beckman Center for the History of Chemistry*; Chemical Heritage Foundation: Philadelphia, PA, September–November, 2013.

11. International Committee on Machine Documentation in the Chemical Field (1968-1978). Box #91. International Union of Pure and Allied Chemistry – Addenda, Series VII. Standing Committee. Donald F. and Mildred Topp Othmer Library of Chemical History, Chemical Heritage Foundation, Philadelphia, PA, November 2013.
12. Riegal, B. Report to the IUPAC Bureau, July 1969, Cortina, Italy. Box #91, Folder: B91FF1. International Committee on Machine Documentation in the Chemical Field (1968-1978), International Union of Pure and Applied Chemistry – Addenda, Series VII. Standing Committee. Donald F. and Mildred Topp Othmer Library of Chemical History, Chemical Heritage Foundation, Philadelphia, PA, November 2013.
13. Herman Skolnik Award, “to recognize outstanding contributions to and achievements in the theory and practice of chemical information science.” American Chemical Society Division of Chemical Information. <http://acscinf.org/content/herman-skolnik-award> (accessed May 29, 2104).
14. Brecher, J. Graphical representation standards for chemical structure diagrams (IUPAC Recommendations 2008). *Pure Appl. Chem.* **2008**, *80*, 277–410; DOI: 10.1351/pac200880020277 (accessed May 29, 2014).
15. IUPAC International Chemical Identifier Project. International Union of Pure and Applied Chemistry, Chemical Nomenclature and Structure Representation Division; McNaught, A., Project Chair. Project No.: 2000-025-1-800, 2001-2005. <http://www.iupac.org/home/publications/e-resources/inchi.html> (accessed May 29, 2014).
16. Scandone, M. Spectra and Searching from Punch Cards to Digital Data. In *The Future of the History of Chemical Information*; McEwen, L., Buntrock, R. E., Eds.; ACS Symposium Series 1164; American Chemical Society: Washington, DC, 2014; Chapter 10.
17. Hoffmann, R. *The Same and Not the Same*; Columbia University Press: New York, 1985; pp 53–84.
18. Schenck, R. J.; Zapiecki, K. R. Back to the Future: CAS and the Shape of Chemical Information to Come. In *The Future of the History of Chemical Information*; McEwen, L., Buntrock, R. E., Eds.; ACS Symposium Series 1164; American Chemical Society: Washington, DC, 2014; Chapter 9.
19. Simmons, E. S. Patents and Patent Citation Searching. In *The Future of the History of Chemical Information*; McEwen, L., Buntrock, R. E., Eds.; ACS Symposium Series 1164; American Chemical Society: Washington, DC, 2014; Chapter 5.
20. Dubois, J.-E. Chemical Complexity and Molecular Topology: The DARC Concepts and Applications. In *The History and Heritage of Scientific and Technological Information Systems: proceedings of the 2002 conference*, Chemical Heritage Foundation, Philadelphia, PA; Rayward, W. B., Bowden, M. E., Eds.; Information Today, Inc.: Medford, NJ, 2004; pp 149–167.
21. Currano, J. N. Searching by Structure and Substructure. In *Chemical Information for Chemists: A Primer*; Currano, J. N., Roth, D. L., Eds.; Royal Society of Chemistry: Cambridge, U.K., 2014; pp 109–145.
22. *Principles of Chemical Nomenclature: A Guide to IUPAC Recommendations*; Royal Society of Chemistry: Cambridge, U.K., 2011.

23. Hepler-Smith, E. *Herdegen Fellow, Beckman Center for the History of Chemistry*; Chemical Heritage Foundation: Philadelphia, PA, 2013/14. Personal communication, 2013.
24. Zass, E. Looking Back, but Not in Anger. In *The Future of the History of Chemical Information*; McEwen, L., Buntrock, R. E., Eds.; ACS Symposium Series 1164; American Chemical Society: Washington, DC, 2014; Chapter 4.
25. Currano, J. N. Teaching Chemical Information for the Future: The More Things Change, the More They Stay the Same. In *The Future of the History of Chemical Information*; McEwen, L., Buntrock, R. E., Eds.; ACS Symposium Series 1164; American Chemical Society: Washington, DC, 2014; Chapter 11.
26. Lesk, M. The Future Value of Digital Information in Digital Libraries. In *Development of Digital Libraries: An American Perspective*; Marcum, D. B., Ed.; Greenwood Press: Westport, CT, 2001; pp 63–82.
27. Buntrock, R. E. Chemical Information: From Print to the Internet. In *The Future of the History of Chemical Information*; McEwen, L., Buntrock, R. E., Eds.; ACS Symposium Series 1164; American Chemical Society: Washington, DC, 2014; Chapter 2.
28. Rusch, P. F. Computer-Based Chemical Information: The Transition Years. In *The Future of the History of Chemical Information*; McEwen, L., Buntrock, R. E., Eds.; ACS Symposium Series 1164; American Chemical Society: Washington, DC, 2014; Chapter 3.
29. Database Counter. Chemical Abstracts Service: Columbus, OH, 2014. <http://www.cas.org/content/counter> (accessed May 29, 2014).
30. *CrossFire Gmelin: A Tutorial*, Manual Version 2.3; Beilstein Informationssysteme GmbH: Frankfurt, 1997.
31. Maizell, R. E. *How To Find Chemical Information: A Guide for Practicing Chemists, Educators, and Students*, 3rd ed.; Wiley: New York, 1998.
32. DataCite. <https://www.datacite.org/> (accessed May 29, 2014).
33. Chapman, A. Is This Data Fit for My Use? The Challenges and Opportunities Data Provenance Presents. Presented in Dealing with the Data Deluge: Successful Techniques for Scientific Data Management, NISO Virtual Conference, April 23, 2014. http://www.niso.org/news/events/2014/virtual/data_deluge/ (accessed May 29, 2014).
34. Adams, S. E.; Goodman, J. M.; Kidd, R. J.; McNaught, A. D.; Murray-Rust, P.; Norton, F. R.; Townsend, J. A.; Waudby, C. A. Experimental data checker: better information for organic chemists. *Org. Biomol. Chem.* **2004**, 2, 3067; DOI: 10.1039/B411699M (accessed May 29, 2014).
35. *NIST Interactive Data Evaluation Assessment Tool*; National Institute of Standards and Technology: Washington, DC, 2005. <http://www.ceramics.nist.gov/IDELA/IDELA.htm> (accessed May 29, 2014).
36. *The Anatomy of Data*. Center for Information and Numerical Data Analysis and Synthesis (CINDAS), Purdue University: Lafayette, IN, 1970. <https://cindasdata.com/about/history> (accessed May 29, 2014).

37. *Crystallographic Information Framework, a standard for information interchange in crystallography*. International Union of Crystallography website. <http://www.iucr.org/resources/cif> (accessed May 29, 2014).
38. *JCAMP-DX* (fka: Joint Committee on Atomic and Molecular Physical data and the group Data eXchange). International Union of Pure and Applied Chemistry, Committee on Publications and Cheminformatics Data Standards, Subcommittee on Spectroscopic Data Standards; Davies, A., Chair. http://www.iupac.org/nc/home/about/members-and-committees/db/division-committee.html?tx_wfqbe_pi1%5Btitle%5D=Subcommittee%20on%20Spectroscopic%20Data%20Standards&tx_wfqbe_pi1%5Bpublicid%5D=034 (accessed May 29, 2014).
39. Bachrach, S. M.; Nitsche, C. I. Tying It All Together: Information Management for Practicing Chemists. In *The Future of the History of Chemical Information*; McEwen, L., Buntrock, R. E., Eds.; ACS Symposium Series 1164; American Chemical Society: Washington, DC, 2014; Chapter 15.
40. Walker, M. A. Public Chemical Databases and the Semantic Web. In *The Future of the History of Chemical Information*; McEwen, L., Buntrock, R. E., Eds.; ACS Symposium Series 1164; American Chemical Society: Washington, DC, 2014; Chapter 12.
41. Clark, A. M. Cheminformatics: Mobile Workflows and Data Sources. In *The Future of the History of Chemical Information*; McEwen, L., Buntrock, R. E., Eds.; ACS Symposium Series 1164; American Chemical Society: Washington, DC, 2014; Chapter 14.
42. Batchelor, C. Chemistry Ontologies. In *The Future of the History of Chemical Information*; McEwen, L., Buntrock, R. E., Eds.; ACS Symposium Series 1164; American Chemical Society: Washington, DC, 2014; Chapter 13.
43. Gordin, M. D. Beilstein Unbound: The Pedagogical Unraveling of a Man and His Handbuch. In *Pedagogy and the Practice of Science: Historical and Contemporary Perspectives*; Kaiser, D., Ed.; MIT Press: Cambridge, MA, 2005; pp 11–39.
44. Mourning Jean-Claude Bradley, PhD, Department of Chemistry. News Archive, Department of Chemistry, Drexel University. Posted May 14, 2014. <http://www.drexel.edu/chemistry/news/archive/mourning-jean-claude-bradley-department-of-chemistry> (accessed May 29, 2014).

Chapter 2

Chemical Information: From Print to the Internet

Robert E. Buntrock*

Buntrock Associates, 16 Willow Drive, Orono, Maine 04473

*E-mail: buntrock16@roadrunner.com

As presented in the cited symposium (*1*), five decades of progress in chemical information, including publication, media, and retrieval are described in an expanded version. The evolution of the chemical information industry, from both the user and vendor standpoint, is illustrated by the personal career history of the author. The emphasis is on the chemical, petroleum, and petrochemical industries along with interaction with the publishing and academic sectors. Knowledge of the past is helpful and even necessary to analyze the present and attempt to predict the future.

... “What we owe the future
is not a new start, for we can only begin
with what has happened. We owe the future
the past, the long knowledge
that is the potency of time to come.”...

Wendell Berry, At a Country Funeral,
from The Country of Marriage, Wendell Berry, 1973

Introduction

It was both an honor and a pleasure to be part of the symposium that is the basis for this ACS Symposium Series publication on the Future of the History of Chemical Information (*1*). Although the title seems anachronistic, the future does

depend on the past and how we use it and adapt to it. In that vein, as a veteran of the information wars, I'll describe, in an expanded version, the evolution of chemical information from the days of print and embryonic computerization to the onset of the Internet. The time span will be about 50 years from the mid-50s to the mid-2000s, corresponding to most of my active experience in the field. In addition to the introductory poem, per this quote from Santayana (2), "Those who cannot remember the past are condemned to repeat it", the past can be used to guide the future.

Several aspects of chemical information are being covered in more detail by other symposium speakers and authors but I've attempted to be supplemental rather than redundant. Pioneers in the chemical information field will also be highlighted. The emphasis will be on the non-academic, industrial sphere but that arena has obviously interacted with the academic throughout history, cantilevering or bridge building if you will. In addition, for the industrial scene, I will concentrate on the petroleum and petrochemical industries. The pharmaceutical and specialty chemical industries had their own requirements for information, often emphasizing other databases. Due their heightened interest and increased funding by organizations within these industries, the prodding and support of these organizations drove many of the developments in chemical information, especially for chemical structures and for patents.

"Classical" Searching

As noted in chemical information books (3, 4), as with information sources in other disciplines, chemical information can be categorized into primary—original articles and documents; secondary—abstracting and indexing services, databases, encyclopedias, monographs; and tertiary—directories, guides. Searching of primary sources is done by reading, scanning, or use of tables of contents and annual source indexes. Searching secondary sources usually involves subject searching using the indexing provided. Chemical information has two additional aspects as compared with other information: chemical structures and chemical reactions. The latter not only involve starting materials, reagents, and products, but have a vector aspect (the direction of the reaction arrow).

Using my experiences throughout my educational and jobs will hopefully be exemplary. The narrative begins with my childhood experiences in chemistry. I had a basement lab but had only a mediocre high school chemistry teacher and textbook. I was overjoyed to be able to pursue further an excellent education in chemistry at the nearby University of Minnesota. I had the good fortune to work with my Organic Chemistry teacher after my sophomore year. Wayland Noland was also my first of several mentors in chemical information. The Chemistry Department at the University of Minnesota had a departmental library, a resource becoming extinct in too many schools. If one had a free hour (a rarity given our intensive schedule with more classes than the ACS Certification requirements) one could study in the reading room before the next chemistry class. The resources were primarily Chemical Abstracts (CA)—in print of course, journals, and several reference books and monographs. Since Dr. Noland's group specialized in indole

chemistry, the Weissberger indole volume (5) was a required reference. With some hints from the boss and graduate student group members, we taught ourselves how to use CA. Another professor gave a biennial course on chemical information. Unfortunately, I never took it due to a scheduling conflict, but fellow students said I didn't miss much since the course was largely confined to a laborious process to search Beilstein, involving System Numbers.

After graduation, I had a summer job at the Veterans Hospital Research Lab in Minneapolis along with a group of other U of M chemistry students. My boss was Herbert Nagasawa, Research Professor of Pharmaceutical Chemistry at the University of Minnesota. We worked on preparation of novel amino acids for incorporation into novel peptides. Once again, Dr. Nagasawa and his assistant, Jim Elberling, were not only our lab mentors but information mentors as well. A discarded but obsolete set of Chemical Abstracts (CA) was shelved in the lab, but another student and I, who lived near campus, were often asked to stop by the U of M chemistry library on the way to work to search deeper in CA with the decennial and volume indexes for specific compounds.

In graduate school at Princeton, I worked for E. C. Taylor, a heterocyclic guru, on several projects. Once again, my lab mentor was also my information mentor and the departmental library was again excellent. Still in organic synthesis, my information resources expanded to monographs including Organic Syntheses, Organic Reactions, Houben Weyl (*Methoden der Organischen Chemie*), Theilheimer (*Neuer Methoden*; later editions were in English Translation, *Newer Methods in Organic Synthesis*), the various volumes in the Heterocyclic Compounds series, and the emerging Patai series, *The Chemistry of Functional Groups*. Between my research and passing language requirement exams, I began to use and scan the German and French literature as well as that in English (although "fudging" by reading the International Edition of *Angewandte Chemie* was easier and faster). Dr. Taylor also gave a course on heterocyclic chemistry which broadened our horizons in this important area even further.

We were strongly encouraged by our professors to read current journals since most of the Cumulative Exam questions came from these sources. Reading journals was the primary mode of current awareness for our research. Dr. Taylor consulted with various pharmaceutical companies, at least one of which reimbursed him with novel information products and services, examples of what the companies were using. One was a subscription to ASCA, the ISI (Institute for Scientific Information) current awareness product based on the updates to the Science Citation Index (SCI). The subscription was for 10 key references to be monitored for citation updates. It did prove to be a good current awareness tool. Taylor also recommended that his students conduct a literature search and write it up in publishable form as the introduction to the thesis. He also encouraged me to recycle my bibliography through the SCI. I did and uncovered a few more references. After I received my degree, we published the literature review in *Chemical Reviews* (6).

With this educational background, I was well prepared to cross the bridge into the world of industrial research. My first job was with Air Products and Chemicals, in a fairly new agricultural chemicals group. They already had a potential winner but I was commissioned to explore new areas. Although

primarily an engineering company, the library was fairly well stocked with chemistry journals plus encyclopedias and other resources necessary for support of industrial research. Not knowing where my next idea was coming from, I read or at least scanned a number of journals with organic syntheses topics. Requiring broader coverage, I began scanning a number of various “trade” journals for new disciplines (agriculture, engineering, etc.) with which I needed to become familiar. After a brief time on the library shelves, the journals were circulated to those who wished to see them at their desks. In fact, our next series of “winners” was inspired by an article in a then current journal issue.

What to do with this increasing flood of journal articles and CA abstracts? I briefly attempted to index my filing system with a set of edge punch cards; three tiered for additional sub category indexing. However, I abandoned this as too time consuming to be cost effective. Computerized personal citation and reference systems were decades in the future.

Computerization

The decade of the '60s began the onset of publication preparation by computer systems which expanded current awareness capabilities. Chemical Abstracts now appeared weekly with a keyword index in the back section. The bio and organic sections appeared every other week, but I scanned some subsections and used the keyword index in each. By this time, our research group had some key ring systems we were developing which were adequately indexed by Chemical Abstracts Service (CAS) with keywords. In addition, KWIC indexes (Keyword in Context) appeared. Text phrases of about 5-6 words were extracted from digital documents with keywords in the middle of the phrase. One of the first publications was Chemical Titles, which appeared weekly, beginning in print in 1961 and in computer readable format in 1965, consisting of rotated title words from journal articles abstracted in current issues of Chemical Abstracts.

About the same time, ISI began publishing Current Contents. Tables of contents from the journals used in preparation of the SCI were reproduced along with a KWIC index. Like the SCI, the coverage was more than three thousand journals covering many fields of science. The title page of each issue listed the journal issues covered and was followed by an editorial by Gene Garfield, founder of the whole enterprise and pioneer in scientific citation searching, which covered a variety of topics not necessarily just on scientific information. Subjects covered the making ice cream (and why Breyer's was the best) to citation ratings of articles and journals. Garfield also championed the writing and value of reviews. Some waggishly referred to these editorials as the “Thoughts of Chairman Gene”. Observing the rise of English as the predominant language of science, most of the journals covered in the services of ISI were in English and Garfield further championed the use of English in the communication of scientific research results. For more background on ISI and citation searching see the chapter in this volume by Bonnie Lawlor.

In addition, ISI began Current Abstracts of Chemistry/Index Chemicus (CAC/IC) and Current Chemical Reactions (CCR) which reported “new”

chemistry and “new” chemical reactions. A fair amount of “not-so-new” chemistry was also included in context. Along with the bibliographic citations, graphic reproductions of the new structures and new reaction schemes were shown. Compounds were indexed with WLN (Wiswesser Line Notation), which was a linear notation for chemical structures. Scanning those indexes also provided good current awareness. Some companies, mostly chemical and pharmaceutical, acquired the tapes of all of these products and ran current awareness profiles in-house.

To facilitate current awareness, at Air Products we had subscriptions to the weekly issues of Chemical Abstracts plus print subscriptions to Chemical Titles, Current Contents Chemistry Section, CAC/IC, and CCR.

In 1968, Chemical Abstracts Service (CAS) began demonstrating CA Condensates, which was the computer tape version of the bibliographic citations for the abstracts appearing in current issues of Chemical Abstracts along with the keyword indexing that appeared in the back pages of the printed issue. First used for current awareness, an online accessible version of archived collections of CA Condensates was demonstrated at the Spring 1969 ACS Meeting in Minneapolis (my home town) and several of us saw the potential value. Tapes were available but the real future lay in access to the backfile as it grew from its 1967 startup.

Need for Subject Expertise

Over the last several decades, those most interested in accessing and retrieving chemical information were chemists themselves. However, the explosive growth of technical information after World War II led to overloading of both the publishing and indexing of the information as well as the burdening of the users: research chemists. I came to realize that chemists were simultaneously the most blessed yet the most cursed of users of scientific information. Most blessed because the chemical information resources were so much better than for other disciplines, but cursed because the sheer size of and often cryptic access methods for the resources was depriving many researchers from access to needed information. Observations such as “two hours in the library can save two weeks in the lab”, although often true, all too often fell on deaf ears. This situation gave rise to the emerging careers of chemical information specialists who were chemists. This alternative career switch came for many after a beginning career in research. It’s noteworthy that currently an increasing number of chemistry graduates are making this switch immediately upon receipt of their degrees. Many technical organizations hire chemists and engineers for service positions in technical information. One of my maxims was “It’s usually easier to train a chemist to be an information specialist than it is to train a librarian or information specialist to be at least comfortable with chemistry”.

After losing my first job at Air Products, I secured a lab position with Amoco Oil at the Research Center in Whiting Indiana, a few blocks from the refinery, also in pesticide synthesis. The group was much more marketing oriented and I was the only chemist doing synthesis. Once again, I signed up for circulation of several dozen journals. This brought me to the attention of the Director of

Information Services who personally gave me an orientation to the services of the Division. At the end of the session, he said that he hired chemists to do searching services because they could “speak the language” and that if I ever considered a new position, to contact him.

Later that year, the Library moved out to the new Research Center in Naperville Illinois, a Chicago suburb about 50 miles away, along with Amoco Chemicals and Corporate Research. Although a small library remained in Whiting, several key resources went to Naperville. Funding of the Ag Chemical Group was not from research funds but from Amoco Oil Marketing. When that funding decreased the next year, my position could no longer be supported. After a few weeks of interviewing for lab positions in the various research groups of the Amoco companies, a searching position opened up in the Research Information Division at the Naperville labs and I started full time in my second love: information work.

I was hired to provide information services to the research staff. Since I was a chemist, I spoke the language of chemists and engineers and could better interpret their questions. My motto became “I am a chemist, I work in a library, and I’m not a librarian”. The Amoco Research Library was even better stocked than was Air Products. In addition to a complete set of Chemical Abstracts, the library had additional reference sources to those already cited such as Beilstein, Engineering Index, the Kirk-Othmer Encyclopedia of Chemical Technology, the Encyclopedia of Polymer Science and Technology (Mark), and additional specialized monographs on topics like oxidation reactions and catalysis. Several data resources, especially thermo- and physicochemical data were also available. Depending on the search request, examination of this wealth of reference material was often done first, either by the requesting scientist or the information services personnel, before diving into Chemical Abstracts.

Unlike the experiences in several other companies, the Information Services director and the head librarian had been able to plan the new research library, as one wing on the main office building. The research campus was expanding and the library had to serve a growing number of users scattered over 180 acres in several building complexes. An MIT study had shown that a typical engineer would walk 75 yards to a library but only 50 yards if stairs were involved. So, the library was made attractive as possible. We joked that it was a good suburban library: trilevel. The middle level featured shelved reference materials and current journal racks. Interspersed were reading areas which were well used by researchers, especially during the lunch hour (the cafeteria was in the same building just down the hall). The upper (and entry) level had the front desk, offices, and books. The lowest level had back runs of journals, additional reference series, and microfilm. As soon as the microfilm was received, the Chemical Abstracts weekly issues were retired and motor driven film reels with printers became the archival resource for CA.

Since the various research groups of the operating companies were dispersed among several buildings, branch libraries with limited but relevant resources were established. The Amoco Chemical Library had some duplicate journal holdings, a few reference materials, and a duplicate but less complete set of Chemical Abstracts. We later innovated by recruiting laboratory technicians to be trained to

search in addition to running the branch libraries. Since lab technicians typically had a bachelors level degree in chemistry, they too “spoke the language”.

Support of the operation varied. Library services were “free” in that they were supported, for subscription and staff costs, by overhead assessed to the research groups on a per capita basis. However, “non-traditional” services, including literature searching and the supporting computer services were charged to the requester’s project account. Initially, the searching services were largely manual and the expenditures were for the billable hours incurred by the searcher.

Current Awareness

This two tiered support structure became important in the early 70s as the number and availability of computerized services and databases increased. Although I began searching with printed CA as my primary resource, things changed rapidly. The first innovation was current awareness or SDI (Selective Dissemination of Information). CAS instituted their ISS (Individual Search Service) service. Customers submitted a profile (a custom designed search strategy) on a coding sheet and the profiles were batch run against the weekly CA file updates in Columbus. Output was returned on computer paper, tearable into 4×5-1/2 inch sheets for filing. My predecessor offered a “teaser” to the research staff: a free profile for a year. After that (and for any profile over the trial offer of 40 subscriptions), the customers project number was charged. When my colleague transferred back to the lab, I was thrilled to take over the project and administered it for the remainder of my career at Amoco (I maintained a profile myself). I coded the profile sheets, maintained the profiles, and distributed the results.

My usual “pitch” to new customers was, “Nothing beats a good reading program to keep up with developments in your area of interest, but even good reading programs need to be supplemented”. We usually recommended that when a client requested a background search for a startup project that they institute an SDI profile to keep up with developments. The service improved over the years and full subject indexing was added. Our SDI services expanded over the years to include “automatic”—saved search strategies periodically updated by the vendor system-- or “hand executed” online updates—a saved strategy executed periodically by the user. The service remained patronized into at least the ‘90s. My colleague Tom Wolff and I published an article on our SDI services (7), which presents details on profile construction.

I’ve already mentioned online services. In the ‘60s, six regional NASA information centers (RDCs) were set up on university campuses to spin weekly tapes of various databases for current awareness including NTIS (National Technical Information Service) federally funded research reports, and CA Condensates. The services were originally for current awareness only but virtually no one could provide retrospective services. NERAC (New England Research Applications Center), ultimately the only survivor, eventually did provide retrospective services in batch mode runs of the archived tapes.

Advent of Online Services

The existence of data in digital form is necessary but insufficient to create a usable and successful information service. This has been true for the entire existence of digital data up to and including the present. Early in 1972, an information salesman began making the rounds, pitching an online information service with a few databases of interest. I don't remember his name or the name of his company (the information services were a spinoff of the primary business), but we and several other librarians and information specialists went to a demo in the Chicago area. Even though Chemical Condensates was mounted on the rudimentary system, we weren't too impressed. For one thing, due to limited memory, all three letter words were "stop" words (stop words are those words not placed in the inverted file and therefore not searchable). As my boss said at a meeting later, "Oil is a three letter word". The librarian from American Can reminded us that "can" was also. I don't think that entrepreneur made very many sales, certainly not to us, and his enterprise soon disappeared.

Soon after, the Lockheed Business Group, spearheaded by Roger Summit, and SDC (System Development Corp), led by Carlos Cuadra, introduced their new online information programs, DIALOG and ORBIT respectively. They began making presentations around the country demonstrating their new services. The demos were typically held in hotel conference rooms, invariably with problems getting an outside telephone line via the hotel switchboard. Both seemed far superior to the previously described service with a far more limited list of stop words. Both mounted databases like ERIC (educational), NTIS, and COMPENDEX (Engineering Index). Since SDC had also designed ELHILL, the program that ran MEDLINE at NLM (National Library of Medicine), they also had MEDLINE. Carlos Cuadra himself often led the demos.

Many stories developed around these two prominent online pioneers. Evidently Summit acquired a large number of used IBM "data cells" or memory units from the parent Lockheed Corp. As a result, DIALOG seemed to always have more memory capacity but they were slow. In contrast, Cuadra had to pay "retail" for computer systems and memory from RAND, SCD's original parent company (8). An apocryphal story circulated around the industry that whenever customers complained about slower response times, Summit had the equivalent of a system rheostat and would turn up the dial a bit. Cuadra was quite active in information circles, especially ASIS (American Society for Information Science, now ASIST). He created and edited several editions of the Annual Review of Information Science and Technology. A former lounge pianist, Cuadra would often have a piano brought to his suite at ASIS meetings and entertain the guests at a social hour. (For more on these pioneering information sources, see the chapter in this volume by Peter Rusch.)

At first, access to ORBIT required a subscription fee over and above connect time fees but SDC soon dropped that for ORBIT usage. Ironically, although the Amoco library was comfortable with library subscriptions due to support by corporate overhead, the Information Services group, funded by pay as you go, was not. The suite of databases available was not yet of prime interest to us at Amoco. However, when CA Condensates, the file of primary interest to us, was

mounted on ORBIT, we signed up. We were confident that our clients would pay for enhanced services and would be, along with us, guinea pigs. Later, when DIALOG mounted CA Condensates, we also signed up with DIALOG. First learned is probably best learned and often leads to more comfortable usage so these and further developments led us to favor ORBIT over DIALOG even in later years. Corporate frugality usually prevented me from going to training sessions, usually at sites far remote from Chicago, so I was largely self-taught. However, as we added searchers to our staff, we sent them to online training classes.

Other broad-based searching services appeared as well as systems dedicated to a narrow set of databases. The BRS system was an outgrowth of online services provided through BCN, the SUNY Biomedical Communication Network. It used the STAIRS system from IBM. STAIRS was among the first of systems capable of handling semantic full text searching but was widely regarded as a memory hog. BRS was able to tweak the STAIRS program to produce an effective search system. Their innovative marketing featured “unbundled” pricing by separating out the royalty due the producer of the database and adding a flat connect time fee for their portion of the service (9).

Although it never was adopted as a commercial search service, the SMART system, designed by Gerard Salton (10), was a powerful program. Based on Salton’s concept of searching for concepts rather than by coincidence, by means of vector cosine correlations, it was designed for effective contextual searching of full text. Output was ranked for relevance in descending order per the vector overlap between query and retrieval. Salton had a demo system and his associate, Michael McGill, continued the endeavor with a version called SIRE (Syracuse Information Retrieval Experiment) (11, 12). A commercial version, MASQUERADE, was adopted by a few companies for their internal corporate files (13) but never caught on to any great extent. I was able to participate in a multi-user comparison test of searching a subset of the API abstract file using the indexed file on ORBIT vs. a loading of a corporate version of MASQUERADE. For the publically available database used in the study, in general, searches of the database on ORBIT were better both in recall and relevance than the version on MASQUERADE. Unfortunately, the complete results were never published.

Compared to searching resources in print, the advantages of online access to digital databases rapidly became apparent. In addition to indexing, all bibliographic details were searchable, some of which possibly for the first time including title words, all authors, corporate/institutional authors, publication source titles, dates, language, and patent numbers.

The “Business” of Information Services and Searching

I should point out that funding of information services by hours/dollars per request got us into the online game more rapidly than the typical college or university. Connect hour fees are more difficult to absorb into typical library budgets, more accustomed as they are to subscription fees. Building bridges back to academia, whenever we described our experiences formally or informally, the academics were quite envious until they found ways to finance their own services,

often years later. However, charge back accounting is a two edged sword. When times get tough, as they did three times in the '90s, our clients and customers cut back on allegedly non-essential services and decreased their support of fee-based services as well as cutting their overhead reimbursements to libraries.

Because of the size of the Amoco Research Center—1500 employees scattered over 180 acres in seven building complexes—the Research Information group was less dependent on walk-in business than many information centers. Requests came in by company mail, telephone, and later via company e-mail. I usually insisted on some sort of pre-search interview, quizzing the requester on what was known or had been searched, and what was trying to be accomplished. Customer supplied keywords were often necessary but insufficient. Reporting of results was done via a cover letter, an official company document, indexed and archived. These documents were searchable internally, often accessed to avoid duplication or as foundations for more current retrieval of information. As well as a descriptive title, the nature of the request was described as well as what was searched, what wasn't searched, and often some evaluation of the results. The results were categorized as those "of interest", "possibly of interest", and non-relevant hits were discarded. At first, references were cited bibliographically and abstracts, copied from microfilm reels, were attached. Later, the bibliography was generated from the online printout and even later digital abstracts were included. After I left and went out on my own, I based my search reports on these models.

Physical distance between offices and search rooms (where the terminals were located) as well as cantankerous systems usually precluded us from searching with the client present. Therefore my colleagues and I stressed good pre-search interviews. Whenever we reported on our searching practices at meetings, a vigorous discussion ensued on the value of having the requester present. Most of those searchers who stressed the presence of the customer present were in academic libraries or hospitals and were not necessarily, especially in the latter case, subject experts. For searching medical information, having the requester present probably was the better policy. In our case, if we did encounter a problem, relevancy, too many hits, or zero hits, we'd contact the requester before proceeding. Often, a successful search requires more than one searching session, a "quick and dirty" to determine the extent and quality of the retrieval, and subsequent sessions to refine the output. Often, we looked for reviews first, analyzed them and their cited references, and then either updated or supplemented the review.

Many research information specialists have found that superior communications between clients and searchers occurs when information specialists are adjunct members of research groups and attend research group meetings. We in the Research Information group promoted such liaisons but we rarely achieved that goal. I was able to interact with one group for a couple years and the experience was mutually beneficial. Lacking group meeting attendance, we searchers usually attended in-house presentations by the research groups and encouraged our input before and after the presentations.

Patents

Until recently patents have been a form of literature used much more by industry than academia. Because of their complexity, patents are a form of literature unfamiliar to many since, unlike non-patent publications, more than one document is often associated with a given patent number. From the late '60s on, funded by Amoco Patents and Licensing, Amoco Computer Services in Naperville processed biweekly tapes of the IFI (Information for Industry) Comprehensive US Chemical Patent File and merged the tapes into a searchable backfile. Requests were submitted by code sheets and a card deck was keypunched. This deck was run against the database overnight to incur the lowest charges. The search program involved a numerically ranked hierarchical system. Numerical scores of documents retrieved indicated the presence and hierarchical relation of terms present so the system was quite good for both relevance and comprehension. Broader terms provided higher recall and more specific terms provided enhanced precision. This multi-level retrieval was especially valuable when using a non-interactive, overnight batch system. If use of very specific terms produced no hits, the more general results could be analyzed for possibly relevant content. Later, three versions of the IFI files (CLAIMS) were mounted on both DIALOG and ORBIT which used more standard search protocols including numerical codes for chemical compounds, subjects, and corporate authors.

Even for a chemical company, mere knowledge of chemical patents is necessary but insufficient. In the 1960s, Derwent, founded by Monty Hyams, began abstracting and indexing patents of interest to the chemical and pharmaceutical industries as well as supplying tapes of the files to be processed by subscribers. Customers, now appreciating the advantages of online vs. batch tape access, wanted online searchable files and the Derwent WPI files (World Patent Index) were mounted first on ORBIT and later on DIALOG. Coverage was expanded to subjects other than chemical and pharmaceutical as well as covering an increasing number of countries and patent granting organizations. Chemical compositions could be searched by an alphanumeric coding system for chemical structure fragments developed by Peter Norton (14).

In addition to the Derwent and IFI CLAIMS patent files, searching services later mounted full text patent files from various countries and regional patenting consortia. One was JAPIO, an English language full text version of Japanese Kokai, or unexamined patents. Since a double translation was involved for patents submitted from other countries, from English to Japanese for the submission and back to English for the database, some interesting quirks often developed in the text of the resulting Kokai documents, especially with non-Japanese names.

For additional background on patents and patent searching, see the chapter in this volume by Edlyn Simmons and references cited therein. In addition, a recent book, *Chemical Information for Chemists* (15), has, *inter alia*, an excellent chapter by Michael White.

Physicochemical Data

Petroleum and petrochemical companies have extensive needs for physical and chemical data. At Amoco, we had several sources, especially for thermochemical, engineering, and physical properties. Chemists and engineers as well as we information specialists consulted these sources first before proceeding to the secondary literature. In *Chemical Information for Chemists (15)* A. Ben Wagner has an excellent and comprehensive chapter on resources and searching of physical properties and spectra, both printed and digital.

Collegial Interaction with Vendors

Since the '60s, Amoco had been an active subscriber to the products of CAIS (Central Abstracting and Indexing Services) of API (American Petroleum Inst.). Using the world's second best thesaurus, 2nd only to NLM's MeSH (Medical Subject Headings), an expert staff abstracted and indexed (based on the article abstract) a select list of journals of interest to the petroleum and petrochemical industries (16). The thesaurus also covered petrochemicals and the hierarchical structure allowed for both generic and specific indexing and retrieval of chemical compounds and related materials. The literature file was also a good source of information on engineering, environmental, and transportation topics. In addition, they re-indexed an appropriate subset of Derwent patents. Bulletins were published for subscribers for both files and tapes were provided for in-house access.

API/CAIS had several subscriber advisory committees and committee members requested online access. After negotiations, the API files were mounted on ORBIT. Features like SENSEARCH and STRINGSEARCH allowed precision and proximity searching of indexing and text. They were later mounted (minus these latter, often important features) on DIALOG and STN (Scientific and Technical Information Network; an online service run by CAS and others). The API files are now known as ENCOMPLIT and ENCOMPAT. ORBIT was later merged into Questel.

Online Searching; Nuts and Bolts

So, with the advent of online searching, how was searching accomplished by the user? At first, 10 cps Teletypes were used but by the time we began in 1972, 300 baud phone modems with attached terminals were used. Connections were tricky and Murphy's Law said that the user was often dropped at crucial times. Since the cost of usage was based on connect time, especially for slower typists like myself (world's fastest four fingered, most error-prone), it was advisable to prepare a search strategy ahead of time but also to be flexible enough to take advantage of the interactive nature of the process (a boon over batch submission and running). Like users everywhere, we online pioneers immediately began asking for higher speed connections. Network connection capabilities rose to 1200 baud in the later '70s. The connections seemed better and higher rates of output printing were welcomed. We found that although it was possible to read 300 baud display output

“live”, one was limited to scanning 1200 baud and we had to depend on reading the printout. Some database producers, like Monty Hyams of Derwent, were leery of the higher speeds, worried that their files would be stripped and duplicated. In 1978, I was commissioned to go to a Derwent subscriber meeting in Stratford England to publicize 1200 baud searching and to reassure Hyams that his database was already too large to enable wholesale copying even at that speed.

Interactive communication was via packet networks, especially Telenet and TYMNET. Along with the advantages of ORBIT and DIALOG, the relative merits of the use of the networks were hotly debated among information professionals every time they got together. One’s location and local phone company often determined which service was better.

Online searching activity continued to grow rapidly in the ‘70s, at first in industry and eventually in academia. Not only for subject searching on a wide variety of topics in a wide variety of fields, both scientific and non-scientific, but the searchable fields outlined previously allowed for verification and identification of specific references, a boon to any reference desk in any library.

A typical search performed at Amoco for a client, usually a research scientist or engineer, involved a discussion of the problem to be solved and the questions to be asked, investigation of applicable standard resources including encyclopedias and other reference works, followed by a search for previous reviews. After analysis and evaluation of these results, if any, an online search would proceed unless the information retrieved was sufficient.

In my opinion and probably that of several other information specialists, there are advantages of offline search strategy preparation to optimize the effective use of interactive searching. This holds true even if connect time fees are no longer an issue since these pricing systems may not be available to all.

Chemical Compound Searching

With printed CA indexes, compound searching was done via chemical nomenclature and/or molecular formulas. CAS indexing of chemical substances is based on IUPAC nomenclature rules, with a somewhat modified CAS “dialect”. The advent of systematic nomenclature, especially the Ninth Collective indexing (9CI), allowed at least some extent of structure and substructure searching. For example, most aromatic amines were named with the “heading parent” benzeneamine. Ring systems, named by Ring Index policies, were particularly well suited for searching by index name (e.g., 1,3,4-thiadiazole). In addition, the creation of the CAS Registry System, using CAS Registry Numbers (CASRN) in 1965, greatly facilitated compound identification and retrieval. For more on the language of chemistry see the chapter by Bill Town. For more on ontologies of chemistry see the chapter by Colin Batchelor.

Printed lists of CASRN became available, including the list of Common Chemicals and later complete lists of CASRN from CAS as well as the growing TSCA list (Toxic Substances Control List). However, we chemical information searchers wanted online chemical dictionaries even if they were only searchable by text or CASRN. The first such file I knew of was CHEMLINE from NLM.

At first, it contained about 130,000 compounds from references indexed in TOXLINE, the toxicity information subset of MEDLINE. However, NLM, in a dispute with SDC, had pulled the MEDLINE file in-house and restricted access to all three files only to those who were trained on the MEDLINE and MeSH systems. Of course, further access was lost to improvements in ELHILL, the NLM online search system version of ORBIT, even as ORBIT continued to be upgraded. Prior to that time, MEDLINE/MeSH training consisted of two-week classes typically given at NLM headquarters or at regional NLM libraries. After the takeover, NLM offered 2 day classes in Bethesda. A side benefit was access to CHEMLINE and TOXLINE.

Since we had no responsibility to the Medical and Industrial Health groups in our company, we had no great use for MEDLINE but the brief training became attractive. I attended one of the sessions and our CHEMLINE/TOXLINE training was provided by Bruno Vasta, the “father” of the two files. I used CHEMLINE until something better came along. The NLM files eventually were mounted on DIALOG along with other related biomed files as well as eventually on the STN system from CAS. Later, we seldom did toxicity searches but when we did, we took advantage of efficient “one-stop searching” of the expanded suite of toxicity databases (in addition to the CA File) on STN.

The venerable Beilstein Handbook of Organic Chemistry also underwent extensive evolution in the online transition period. For much of its existence, Beilstein has been the premier source of reviewed data on organic chemical compounds. Before the database existed online, the best way to search for existence of organic compounds (absence implied the possibility of novelty) was to search the CA file and/or indexes in reverse chronological order and supplement with a search of Beilstein.

Although the excellent Beilstein System Number process was the best way to index the database, searching exclusively by this method (often the subject of entire chemical literature courses) was not necessarily the most facile or effective. Over the period of several successive ACS national meetings, the Division of Chemical Information (CINF) sponsored vendor symposia, organized by several vendors and producers of information online. At the Beilstein symposium, Reiner Luckenbach detailed a facile and accurate method for searching the database in print. The formula index for the 2nd supplement (based on literature through 1929) should be searched and if the compound was not found the postings for relevantly named compounds should be examined. Using the Beilstein System Number (BSN), determined from the postings, subsequent indexes should be searched. If the compound could not be found in the Basic or first two Supplementary Series, the System Number could be determined by searching similar compounds and the search proceeded from there.

The origin and development of the Beilstein Handbook is described in the first two chapters of an ACS Symposium series volume (17). The fifth edition in English appeared in 1984 and the online version was first mounted on STN in 1988 and on DIALOG in 1989. The DIALOG loading used the S4 structure searching program produced by Beilstein/Softtron, regarded by many to be superior to other structure searching programs. (A subsequent comparison of S4 with other substructure search systems found S4 to produce the fastest searches.)

(18) The CrossFire in-house version of the online file appeared in 1995 and the complementary inorganic Gmelin file was added the next year. For more background on Beilstein online see the two ACS Symposium Series volumes (17, 19). Beilstein and Gmelin are now part of the Reaxys database system from Elsevier (20). (Also see the chapter in this volume by Swienty-Busch, *et al.*)

In the '80s CAS, first via CAS ONLINE (21), then via the newly formed STN Network, began vending their files online. Detailed indexing had already been added to the CA online files and structure searching and CA abstracts were also added. The structures were offered to DIALOG but were subsequently withdrawn. This produced a few years of acrimony between the two groups leading to lawsuits. Fortunately, these were eventually settled out of court but only after depositions were taken from some users (including me). Concurrently, the DARC system for chemical substructures, developed in France, was established and was quite comparable to the STN system (22).

Prior to the addition of the CA abstracts to the online CA file, users had to convince even some of the staff at CAS that searching abstracts in addition to using the CA indexing would lead to more comprehensive search results. After the addition of the abstracts, I wrote articles demonstrating the value added (23–25).

For some time, the CA file online was limited to 1967 onward. Prior to that, even with the advent of pre-67 material online, searching was better accomplished by searching the printed indexes. For some time, I performed searches for presumably novel compounds in reverse, doing the online search first followed by searches in the collective indexes in reverse order as well as searching the original Beilstein Handbuch and the first two supplements.

As mentioned previously, chemical information is unique among all other kinds of information in at least two aspects: chemical structures and chemical reactions (the latter involve chemical structures, associated data, and vector aspects). The need for computerized storage and retrieval preceded the availability of graphical capabilities. William J. Wiswesser invented WLN—Wiswesser Line Notation—in 1949 (26). Chemical Structures were represented and searched by means of coded text strings made up of characters on a standard typewriter or terminal keyboard. Pharmaceutical companies used it, along with several “dialects”, to index and search their chemical libraries. Eugene Garfield and ISI used it to index their CAC/IC and CCR databases and bulletins. It is still used in some information systems (26).

SMILES, the Simplified Molecular-Input Line-Entry System, was first developed in 1980 (27). It has come to be more standardized than WLN and is possibly more readable by humans than InChI (see below). However, several valid SMILES can be written for the same molecule. As a result, algorithms have been developed that generate canonical SMILES strings that are unique for each molecule. Stereochemistry and chirality can be specified.

Another text-based chemical structure identifier is InChI (IUPAC International Chemical Identifier) (28). Developed by IUPAC and NIST in 2000–2005, InChI was designed to enable searching chemical structures on the Internet and is non-proprietary. Early versions were available under an open-source license but version 1.04 (2011) is available under a custom license. More information can be encoded than can be with SMILES. InChI is being used

by many databases, with varying success, including ChemSpider and PubChem (28).

Other authors in both the meeting symposium and in chapters in this symposium volume have previously described the immense field of chemical structure representation and searching. In his chapter, Roger Schenk as well as others (4) have described historical and current developments at CAS and STN. In *Chemical Information for Chemists* (15), Judith Currano has an excellent chapter on searching for chemical structures.

Chemical Reaction Information

As described previously, the other unique aspect of chemical information concerns chemical reactions. Various books and references series (29–32) were valuable print resources. When computerization of reaction information data began, there were additional issues to consider. Not only must structural data and representations be entered and searchable, but also the identity and nature of reactants/starting materials, products, reagents, catalysts, and conditions must be explicitly searchable. Positioning of functional group addition or deletion from the reactant—the reacting group—must be documented. On STN, chemical reaction information originating within CAS is consolidated in the CASREACT File. In addition, reaction databases from other sources are also vended by STN and provide supplementary and complementary information. For more on these databases and services, see the chapter in this volume by Schenck. For a detailed description of Reaxys, the successor and expansion of the Beilstein and Gmelin services, see the chapter in this volume by Swienty-Busch, *et al.*

MDL Information Systems, founded as Molecular Design Limited, Inc., was founded by Stuart Marson and W. Todd Wipke, in 1978 (33). Their mission was to automate chemical syntheses. Wipke and Corey had designed a computerized “retro” synthesis program where one started with the target molecules and used the system to work backwards toward possible starting materials via feasible reaction paths (34). MDL provided the commercial version of the program and, spurred on by support from the pharmaceutical industry, extended their novel chemical structure and reactions programs to database management systems. The MACCS (Molecular Access System) program was used to represent and archive chemical structures, along with linked chemical and biological property data, which pharmaceutical and chemical companies used to maintain their proprietary chemical “libraries” and the associated testing data. The REACCS program allowed storing and retrieval of chemical reaction data including reagents, reactants, catalysts, and conditions. REACCS used commercially available reaction databases including Theilheimer, the print version of which had a coding system for groups of reactions categorized by bonds broken, bonds formed. However, searching REACCS provided an easier method for more comprehensive searching. As far as I know, MACCS and REACCS were never used for databases with public online availability.

Amoco, more of a process and commodity company than many chemical and pharmaceutical companies without an extensive library of chemicals, was never able to justify complete in-house loadings of MACCS and REACCS. A far-sighted Amoco Chemicals Senior Research Associate had Amoco Research Computer Services load a truncated version of REACCS for his own group. The program saw only limited use and the subscription was terminated in a corporate downsizing.

For more on chemical reaction searching, see the chapter in this volume by Guenter Grethe. In addition, in *Chemical Information for Chemists (15)*, Judith Currano has an excellent chapter on searching for chemical reactions.

Information Resources Complementary to Chemistry

The number of online databases and breadth of topics continued to grow and evolve. COMPENDEX/Engineering Index was supplemented by the addition of INSPEC/Physics Abstracts. Beilstein and Gmelin became available. BIOSIS/Biological Abstracts, EMBASE, the Merck Index, RTECS (Registry of Toxic Effects of Chemical Substances), and HSDB (Hazardous Substances Databank) were added to the biomed armamentarium. Although not one of my favorite topics, “business” information is essential for industry including the research centers so the availability of Chemical Industry Notes (CIN), ABI INFORM, and Predicasts PROMT files were welcomed. STN added reaction databases, their own and others. As STN acquired more and more of these databases, one-stop searching caused many of us to depend more and more on STN, while DIALOG remained the favorite of many academic libraries and users because of much broader breadth of offerings and extensive marketing and training. ORBIT and successors continued to be the service for Derwent Patents until those files, along with IFI patents also became available on DIALOG and STN. Descriptions of the various databases available in digital form have always been available from the vendors or industry wide in the Gale Directory of Databases from Gale Research, currently updated and available in print form or by online access through vendors including Data-Star and ORBIT/Questel. Descriptions of chemical information databases can also be found in the resource texts including Maizell (4), Wiggins (3), and the Wikibook update to the latter, Chemical Information Sources (35).

Amoco Information Services was never able to justify a subscription to Science Citation Index. We never considered it to be a primary resource of information, secondary and complementary to the indexed resources we used regularly. If needed to supplement a search, we travelled to other libraries in Greater Chicago to do a manual search. Therefore, we welcomed the loading of SCI on DIALOG and STN. Even later, citation searching capability was added to the CA files. The results of the searches of both files citation supplemented each other.

In *Chemical Information for Chemists (15)*, Dana Roth has an excellent and comprehensive chapter on resources and searching, both printed and digital, of commercial availability, safety, and hazards associated with chemicals.

Polymer Information

Polymers have been a category of chemicals typically more of interest to the chemical industry than to academics. Although my company, Amoco Corp., was primarily a petroleum company and secondarily a petrochemical company, in addition to monomers Amoco Chemicals was also active in polymers. Their worldwide predominance in terephthalic acid production also led to activity in research on PET (polyethyleneterephthalate) and other polyesters of aromatic acids. In addition, they were large producers of polyethylene and polypropylene. The former has several forms, high and low density, which are process and catalyst specific. The latter has at least three secondary structures, syndiotactic, isotactic, and atactic, with differing properties. We were especially concerned with these nuances for both indexing and retrieval of these compositions along with details on catalysis and processes. Also, Amoco was preeminent in polybutenes, covering a wide range of butene liquid oligomers of varying compositions and viscosities. In searching these compositions over the years, we at Amoco determined that essentially all butene polymers needed to be searched using about forty CAS registry numbers plus the various names, and the polybutene oligomers parsed out by inspection or other aspects of the search request (36). This reference also describes difficulties in searching the various forms of polyethylene. Searching condensation polymers also produces problems (37).

Confusing and even inaccurate indexing of butene polymers is not limited to CA files. Searching the PROMT, API, and Derwent WPI files also produces problems although in fairness it should be pointed out that vague or incorrect description of the compositions in the original literature, especially in patents, can drive both indexers and searchers somewhat crazy. Although not extensively used by companies with interests in polymers, the MDL (Molecular Design Limited) indexing system with nested bracket functions was also applicable to polymers since they are often multicomponent compositions.

Description of polymers, and therefore searching for them is also complicated by their structures. Polymers can be indexed by their monomer components (CAS indexes those as CRN—Component Registry Numbers) or as SRUs (Structural Repeating Units) for regular, single component polymers. Secondary and tertiary structural aspects must be dealt with like stereochemistry, blocks (shorter polymeric strings linked further polymerically and regularly), grafts (addition of terminal components), mode of formation, catalysts, properties (average molecular weight), etc. For a primer on searching polymers, see the chapter by Donna Wrublewski in *Chemical Information for Chemists* (15).

Further Collegial Interactions

Cantilevers may or may not be one way, but most bridges are two way. We at Amoco were fortunate to be able to interact with a number of database producers and vendors. We served on advisory committees and otherwise interacted both formally and informally, especially at ACS and other meetings, including user groups and training sessions. Of course, the discussion typically went along the lines of, “That’s a nice development, but how about this improvement?”

I personally served on committees, with API, Derwent, CAS (including ACS CCAS, the Council Committee on Chemical Abstracts Service), and STN, the latter two even after I “retired” and went independent. One effort was a focus group on polymer nomenclature that I was asked by CAS to organize. Polymers are of particular interest to industry and categorization is confounded by variable compositions and at least three levels of structure. I also consulted on the development of STN Easy.

The information needs of the chemical and pharmaceutical industry obviously drove developments in chemical structure representation and searching by CAS, MDL, and other groups. In addition, similar needs drove developments in that other unique aspect of chemical information, chemical reactions. Representatives from these companies worked extensively with vendors like MDL to advance the capabilities for storage and retrieval of chemical structures and reactions, especially for the massive internal compound libraries that these companies produced. Incorporation of biomed data was also a feature of these files.

I have been saying that I’m the most delinquent founding member of PIUG, the Patent Information Users Group. Although still a member, I’ve only been to two meetings since. I was also on the advisory committee for the Journal of Chemical Information and Computer Sciences (JCICS) and helped select two editors and defend another.

Education and Training

The online service providers and database producers often gave training sessions and workshops. Presenters included Ken Ostrum (aka Dr. O) for CAS and STN databases and services and Peter Rusch, Mary Ann Palma, and others for DIALOG, often given at professional society meetings or dedicated workshops. We knew that many users could not regularly attend meetings (including several of us at Amoco). So, we often hosted regional training sessions for database producers and vendors in a conference room nearby to the Information Center. One advantage for us was that we were able to conveniently have more of our staff attend these sessions. Presenters and attendees from throughout Metro Chicago found these sessions valuable. They also seemed to like the donuts and lunches available from our nearby cafeteria. However, Draconian budgetary and cost recovery measures caused our management to charge room rent for the conference room even if there were Amoco attendees. Since there was no admission charge for attendees, our hosting of these sessions unfortunately ceased. Amoco was not alone in corporate hosting of training sessions and workshops as several other companies, especially pharmaceutical, also hosted.

We were also able to publish extensively on developments on searching especially online. For several years, I wrote “ChemCorner” columns for ONLINE and DATABASE magazines and also published papers in JCICS, usually based on presentations at ACS meetings.

We also crossed bridges back to academia. Several of us encouraged colleges and universities to provide instruction in information resources especially online searching. One method was to make road trips to Midwestern colleges and

universities and give demos featuring searches of local interest including author searches of prominent faculty members. For several years we also presented online searching sessions in at least two chemical literature classes at Chicago area colleges.

Soon after Arleen Sommerville, Adrienne Koslowski, and Bartow Culp formed the CINF Education Committee, I joined and I was often its only industrial member. One of our first efforts led to the preparation and circulation of searching modules, based in part on sample searches I prepared and also ran for demo sessions for academia. Judith Currano has described other education activities in detail in her chapter in this volume.

Inspired by ancient proverbs (38) and developments in online searching, a few of us in the industry also got into the education business ourselves (39, 40). We realized that improvements in chemistry databases could make them more attractive to end-users. In the 80s, we embarked on a program of our own design. Amoco research staff, with the consent of their supervisors, could sign up for three training sessions. Firmly believing that online databases are valuable outgrowths of databases previously available only in print, the first session introduced concepts of information in general and the use of printed CA in particular. The text booklet, "How to Search Printed CA", was obtained from CAS. In the second session, sample searches were run live along with printed output. In the third session, the attendees were strongly encouraged to bring their own sample searches. After completing the three sessions, attendees could opt to get personalized, one-to-one training in doing their own searching. The majority of the attendees decided not to take further training but left with a much better appreciation of technical information and were better clients of our services as a result. Several months later after training, we had a fairly good retention rate of active users. They tended to do the "quick and dirty" searches and the alumni still came to us for more complex questions. Two alumni of the program later joined the search staff. For whatever reasons, we apparently had the most successful pre-SciFinder program of end-user training.

Unfortunately, just as I was leaving Amoco, SciFinder was just becoming available. I've only had limited experience ever since because it was of no use to me as an independent information consultant. I did collaborate with Carmen Nitsche at nearby Nalco Chemical to publish a paper on the necessity for training for SciFinder use and management of the program (41). Just recently, ACS made limited use of SciFinder available to all ACS members so I was finally able to search it directly. It is indeed an excellent search system for end-users. However, since I'm more familiar with searching on STN, for future searching for customers I'll continue to use STN. ACS members also now have limited access, at no charge, for e-copies of articles from ACS journals even if one does not have a current subscription.

I agree with Engelbert Zass (see his chapter in this volume) that it is much easier for chemist end-users to search SciFinder than STN. However, we also agree that even with the marketing hype that training for SciFinder use is necessary for effectively obtaining search results. Even then, Zass points out that some searches are not as comprehensive as they could be. I believe that STN is still the premier system for comprehensive searching of chemistry and related topics.

Quality Control

In addition to providing advice and corrections to database producers, the Information group was always concerned with the quality of our own services. In the early '90s, after our merger into the computer services group, our new boss encouraged us to develop a quality control process. Dissatisfied with our brief exposure to Crosby Quality programs, we developed our own Quality process, which was quite successful. One of the concepts we developed was differentiating clients (the requester of the search) from customers (those who paid the bills or furnished the project number, usually a supervisor or manager). In the last year of my tenure, faced with departures of supervisors from Information and Computer Services and another round of "reengineering", we formed a self-managed work team and managed to learn how to run such a group. We were prepared to contribute the results of our success to the rest of the company, but the next downsizing eliminated that possibility.

Eye to the Future

I've taken this tale into this century. I hope I've been able to illustrate the incredible transition from print resources up until the equally incredible impact of Internet resources. I've not attempted to list all of the mergers, acquisitions, departures, and other changes in the industry. For a number of reasons, including mergers, purchases and the economy, the business of Buntrock Associates has pretty much wound down. The number of advances in information access, documented in Gary Wiggins' Chemical Information Sources Wiki (35) and elsewhere in just the last decade continues to amaze me. Although I've been a career long advocate of current awareness, I do not and probably will not have access to RSS feeds and the like but I can see where active researchers would find them valuable. I find recent discussions on the role and place of the physical academic library very interesting. I'm not convinced that the "classic" methods of keeping up with the literature are all that obsolete even in the face of the trend to electronic journals and electronic books. Searching success still hinges not only access but on methodology involving concepts, relevance, and recall. Unfortunately, with more emphasis on unedited source material, veracity is an increasing problem. There's still a need for indexing, Boolean logic, and Venn diagrams. Many researchers still find these methods superior to Google-style searching. Even though Google Scholar uses citations for evaluation of results, at least one study (42) showed that use of the Web of Science (WOS) is superior to the use of Google Scholar (admittedly, Google Scholar is free and use of WOS is by subscription).

However, I hope that I've illustrated that the previous evolution of chemical information has led to current developments and further evolution in storage, access, and usage. I see no reason why these developments will not continue, still grounded in the unique fundamentals of chemical information. *Plus ça change, plus ces la meme chose.*

In conclusion, I'd like to thank my mentors, both academic and non, my colleagues, too numerous to list, both at work and in the profession, the organizers

and other presenters at this symposium, the authors of the other chapters in this symposium volume, and my family, especially my wife Gloria, for putting up with this puttering library nerd and chemist for so many decades. The presentation that this chapter was based on might be my last presentation at one of these meetings. I'd like to say that it's been a great ride and in spite of occasional pitfalls and disappointments, I wouldn't trade my experiences for anything. I'm looking forward to monitoring and writing about the advances the rest of you will be making in creating the continuing future of chemical information. I hope that our shared past will aid in the development of our shared future. Join the revolution!

References

1. Twiss-Brooks, A.; Solla, L.; Organizers, Presiding. Future of the History of Chemistry, Symposium, Div. of Chemical Information (CINF), 244th American Chemical Society National Meeting, Philadelphia, PA, Aug. 20, 2012.
2. Santayana, G. *Reason In Common Sense*; The Life of Reason; 1905, Vol. 1.
3. Wiggins, G. *Chemical Information Sources*; McGraw-Hill: 1991.
4. Maizell, R. E. *How to Find Chemical Information*, 3rd ed.; John Wiley: 1997.
5. Sumpter, W. C.; Miller, F. M. *Heterocyclic Compounds with Indole and Carbazole Systems*; Interscience: New York, 1954.
6. Buntrock, R. E.; Taylor, E. C. Cyclization Reactions of 2,2'-Disubstituted Biphenyls. *Chem. Rev.* **1968**, 68 (2), 209–227.
7. Buntrock, R. E.; Wolff, T. E. Information to Order. *CHEMTECH* **1994**, 24 (4), 8–12.
8. <http://www.infotoday.com/searcher/oct03/CuadraWeb.shtml> (accessed May 31, 2014).
9. http://www.infotoday.com/searcher/nov04/ardito_bjorner.shtml (accessed May 31, 2014).
10. Salton, G.; *Automatic Information Retrieval*; McGraw-Hill: New York, 1968.
11. McGill, M. J. Knowledge and Information Spaces: Implications for Retrieval Systems. *J. Am. Soc. Inform. Sci.* **1976**, 27 (4), 205–210.
12. Noreault, T.; Koll, M.; McGill, M. J. Automatic Ranked Output from Boolean Searches in SIRE. *J. Am. Soc. Inform. Sci.* **1977**, 28 (6), 333–339.
13. Boyle, S. O.; Miller, A. P. Feature Comparison of an In-House Information Retrieval System with a Commercial Search Service. *J. Am. Soc. Inform. Sci.* **1980**, 31 (5), 309–317.
14. Kaback, S. M. Chemical Structure Searching in Derwent's World Patents Index. *J. Chem. Inform. Comput. Sci.* **1980**, 20 (1), 1–6.
15. *Chemical Information for Chemists*; Currano, J. N., Roth, D. L., Eds.; Royal Society of Chemistry: Cambridge, U.K., 2014.
16. Brenner, E. H.; Zarembler, I. Petroleum Information Services: API's CAIS. *Bull. Am. Soc. Inform. Sci.* **1979**, 6 (2), 18–19.

17. *The Beilstein Online Database: Implementation, Content, and Retrieval*; Heller, S. R., Ed.; ACS Symposium Series 436; American Chemical Society: Washington, DC, 1990.
18. Hicks, M. G.; Jochum, C. Substructure Search Systems. 1. Performance of the MACCS, DARC, HTSS, CAS Registry MVSSS, and S4 Substructure Search Systems. *J. Chem. Inf. Comput. Sci.* **1990**, 30 (2), 191–199.
19. *The Beilstein System: Strategies for Effective Searching*; Heller, S. R., Ed.; American Chemical Society: Washington, DC, 1998.
20. <https://www.reaxys.com/info/> (accessed May 31, 2014).
21. Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. The CAS ONLINE Search System. 1. General System Design and Selection, Generation, and Use of Search Screens. *J. Chem. Inf. Comput. Sci.* **1983**, 23 (3), 93–102.
22. Attias, R. DARC Substructure Search System: a New Approach to Chemical Information. *J. Chem. Inf. Comput. Sci.* **1983**, 23 (3), 102–108.
23. Buntrock, R. E. Dual Theme: Corporate Intelligence; Searching Abstract Text. *DATABASE* **1986**, 9 (3), 106–107.
24. Buntrock, R. E. Abstract Searching Revisited. *DATABASE* **1987**, 10 (2), 114–115.
25. Buntrock, R. E. In the Abstract. *DATABASE* **1992**, 15 (2), 94–95.
26. http://en.wikipedia.org/wiki/Wiswesser_line_notation (accessed May 31, 2014).
27. <http://en.wikipedia.org/wiki/SMILES> (accessed May 31, 2014).
28. <http://en.wikipedia.org/wiki/InChI> (accessed May 31, 2014).
29. *Organic Reactions*; Wiley & Sons: New York, 1942.
30. *Organic Syntheses*; Wiley & Sons: New York, 1932.
31. Fieser, L. F.; Fieser, M. *Reagents for Organic Synthesis*; Wiley & Sons: New York, 1967.
32. *Newer Methods of Preparative Organic Chemistry*; Interscience: New York, 1948.
33. http://en.wikipedia.org/wiki/MDL_Information_Systems (accessed May 31, 2014).
34. Corey, E. J.; Wipke, W. T. Computer-Assisted Design of Complex Organic Syntheses. *Science* **1969**, 166, 178–192.
35. http://en.wikibooks.org/wiki/Chemical_Information_Sources (accessed May 31, 2014).
36. Buntrock, R. E. Documentation and Indexing of C4 Compounds: Pathways and Pitfalls. *J. Chem. Inform. Comput. Sci.* **1989**, 29 (2), 72–78.
37. Wilke, R. N.; Buntrock, R. E. Condensation Information: Problems and Opportunities. *J. Chem. Inform. Comput. Sci.* **1991**, 31 (4), 463–468.
38. “If you give a man a fish,
He will have a single meal
If you teach him how to fish,
He will eat all his life.”
Kuan-tzu

39. Buntrock, R. E.; Valicenti, A. K. End-Users and Chemical Information. *J. Chem. Inform. Comput. Sci.* **1985**, 25 (3), 203–207.
40. Buntrock, R. E.; Valicenti, A. K. End-User Searching: the Amoco Experience. *J. Chem. Inform. Comput. Sci.* **1985**, 25 (4), 415–419.
41. Nitsche, C. I.; Buntrock, R. E. SciFinder 2.0: Preserving the Partnership Between Chemist and Information Professional. *DATABASE* **1996**, 19 (6), 51–54, 56–58.
42. Buntrock, R. E. The Better Mousetrap. Searching Wars: Google Scholar Versus Web of Knowledge Versus ... ? *Searcher* **2012**, 20 (1), 34–37; <http://pqasb.pqarchiver.com/infotoday/doc/920321091.html?FMT=ABS&FMTS=ABS:FT:PAGE&type=current&date=Jan/Feb%202012&author=Robert%20E%20Buntrock&pub=Searcher&edition=&startpage=34&desc=Searching%20Wars> (accessed May 31, 2014).

Chapter 3

Computer-Based Chemical Information: The Transition Years

Peter F. Rusch*

Rusch Consulting Group
***E-mail: PFRusch@aol.com**

Use of computer-based online searching of chemical information is now the preferred method of searching the chemical literature. It has fully supplanted many printed publications that no longer exist such as the printed Chemical Abstracts Collective Indexes. The technological advances in hardware, software and computer-readable information sources are reviewed showing how they contributed to the transition to online searching. Generalized search software was applied to content derived directly from printed sources that was often insufficient for direct computer-based usage. The advent of these services set the stage for the modern offerings in chemical information. Many of the early principles may be “rediscovered” as current popular (or simplified) search methods interact more with growing amounts of chemical information that may require more precise searching methods.

Introduction

For centuries chemical observations have been recorded and exchanged with other practicing chemists. This long history of chemical information is critical to the advancement of chemistry and is an ever-increasing body of documentation that is as important to that advancement as laboratory experiments. This review is about a period in the development of chemical information that formed the basis for the evolution that brought us from exclusively print products to the current state of chemical information access. Selection of examples is solely at the discretion of the author and is meant to be illustrative not exhaustive.

Large Chemical Information Databases

For there to be computer-based access to chemical information there needed to be significant sources to support it. In the late 1960's, Chemical Abstracts Service (CAS) recognized that its manual methods of preparing chemical information were not sustainable. The huge growth of the worldwide chemical literature was staggering. The traditional methods of preparing the annual and collective indexes relied upon thousands of index cards each with a hand-written, single index entry. This was true for both chemical subjects and chemical substances. What followed was months of tedious hand sorting of the index cards to produce the indexes. Although the traditional plural of index is indices, CAS always favored the form indexes.

A sensible alternative was emerging in the form of computer processing to collect and sort this vast amount of information. The savings, particularly in view of the exponential growth of the chemical literature, were enormous. That growth combined with the traditional methods posed an existential threat and a solution was needed.

Computing power and mass storage were expensive in the 1960's. To utilize these labor saving devices required a huge investment. Largely due to the efforts of Fred Tate, then Associate Director of CAS, a solution was found. The National Science Foundation (NSF) was solicited for a grant to help CAS invest in the equipment and manpower required to survive. Thus, an NFS Grant funded the transformation for many years.

By the end of the 1960's, CAS had operational computer systems producing both the General Subject and Chemical Substance Indexes. To fulfill the distribution requirement of the NSF grant, CAS produced and made available several products from the growing store of computer based chemical information.

Among the earliest was a printed product called "Chemical Titles" that was a Keyword in Context (KWIC) index of article titles covered by CAS (1). A KWIC index of titles is produced by rotating the words in the title so that each word appears at the beginning of the new KWIC entry. Each entry is completed with the other words in the title. For example, the title of this chapter would give rise to five new entries as follows. "Based Chemical Information The Transition Years # Computer;" "Chemical Information The Transition Years # Computer Based;" etc. This could only be produced by computer as the amount of manual effort was prohibitive.

Other products were available for license and distribution on magnetic tape. CAS developed its Standard Distribution Format (SDF) for all of its distribution tapes. The broadest coverage SDF database was CA Condensates, a database that contained the bibliographic data and keyword phrases for every article covered by CAS and that appeared in *Chemical Abstracts*. Keyword phrases contained three or four words that indicated the content of the abstracts printed in the issue. They were replete with abbreviations and their vocabulary was uncontrolled. They were often permuted so that other words in the keyword phrase would appear as the first word in the alphabetically sorted list of phrases. Normally, these keyword phrases were ephemeral as they appeared only in the weekly printed issue and were not repeated in other printed indexes.

The Institute of Scientific Information (ISI) also entered the chemical information database market with magnetic tapes of bibliographic and chemical substance information derived from its processing system. Their greatest contribution was in the form of the Science Citation Index. Recognizing that scientific ideas propagate through the literature, this unique index was designed to provide access to both cited and citing publications using standardized bibliographic references.

Patent Information

When the chemical enterprise accounted for approximately 20% of US GDP in the mid twentieth century, chemical patents comprised a significant fraction of patents issued worldwide. In general, chemical patents could be issued for uses of chemical substances; chemical processes and contents of matter.

Chemical patents had high value and major chemical companies had specialized departments to manage information about their own and related chemical patents. With such an important market, databases of chemical patent information followed. Among those of broadest importance were: Chemical Abstracts, IFI, Derwent WPI and INPADOC. Due to the vast number of chemical patents worldwide these became large chemical databases.

Chemical Abstracts Service had broad coverage of all of the chemical literature including patents. Patent coverage was for a group of major patent-issuing countries and database content for patents was similar for the journal and patent literature with additional items found only in patent bibliographic information. Still, the other databases found significant markets and were a part of most careful, deep searches.

IFI produced its Comprehensive Database (CDB) with a proprietary, deep indexing vocabulary for both chemical subjects and substances. This indexing vocabulary was created by DuPont and sold to IFI that expanded it and made it a commercial service. A portion of the complete indexing was offered to the general search community at reduced pricing while the full indexing was reserved for subscribers that paid IFI for the right to access and the online search service for the search and output process. Throughout the long life of this patent database it was led by the affable Harry M. Allcock who hosted legendary social events for the chemical information community.

Derwent World Patents Index was the inspired creation of its founder, Monty P. Hyams, who was a patent agent for a British fire extinguisher company, Pyrene. In the 1950's he observed that most chemical patents from European countries issued first in Belgium often several weeks before issuing in any other country. This observation coupled with his knowledge of French caused him to fly to Brussels periodically to read through newly-issued Belgian patents. He disciplined himself to write 150-word abstracts in English for each of his selected patents. Upon returning to England he and his wife transcribed his hand-written notes onto carefully arranged type-written sheets that were reproduced and mailed to subscribers. The fledgling service proved so popular that the redoubtable Mr.

Hyams created his own company named after the building where they lived, Derwent House.

One of the unique features of the Derwent WPI was its chemical fragmentation coding developed by Peter Norton who received the Skolnick Award for this work. This was a collection of alphanumeric codes that described major structural features of a chemical substance. The codes were applied by indexers who selected all of the appropriate codes for a chemical substance. As almost all chemical patents contained Markush structures (named after the inventor in whose patent the structures were legally recognized). The codes were designed to be highly descriptive but lacking in connectivity. Thus, the code for two or more carboxyl groups was useful but it did not describe where in the structure these groups appeared.

INPADOC (the International Patent Documentation Center) was created by a treaty between the Austrian government and WIPO (World Intellectual Property Office). Headquartered in Vienna, Austria, INPADOC set out to create a computer-based master file of patent “equivalents.” From the earliest days of patents, intellectual property rights were valid only in the jurisdiction that issued them. As the chemical enterprise became more global, chemical patents covering the same invention issued in many different countries usually selected by the size of market for the invention in a country. Such patents for the same invention are known as “equivalents.” Using magnetic tapes of information from dozens of different patent offices, the talented Wolfgang Pilch established computer programs to bring these disparate files together to form families of equivalent patents.

INPADOC covered all kinds of patents, not just chemical patents, and it lacked any chemical structure information other than chemical substance names in the patent titles that appeared in a variety of languages including transliterated titles.

In the beginning, INPADOC was represented in the US by IFI. As the INPADOC database grew in the number of countries it covered to create families of equivalent patents, Chemical Abstracts Service ceased production of its similar patent family collection known as the CA Patent Concordance and used the information from INPADOC.

Early Uses of Computer-Based Chemical Information

To promote use of CA Condensates, CAS offered a mainframe, batch search software known as “System 360” named after the then-current top of the line IBM mainframe computer. Only a handful of installations were made. Although not terribly successful and later withdrawn, these installations provided a much-needed testing ground. Searches of CA Condensates showed the value and speed of computer-based searches. They also highlighted the differences between manual searching of printed indexes and computer-based searching.

Manual searching of printed indexes encouraged “browsing” as the user viewed many items, indeed pages and pages of index entries in the search for some particular topic. Computer-based searching offered speed but introduced spurious results that were not expected or relevant because the searcher had to

predict through the use of “keywords” or other terminology, precisely how his topic would be recorded so that it could be found.

Still, computer-based searching of CA Condensates was a huge benefit. Since CAS no longer offered any software to search its distribution tapes, several organizations began independent development of search software. Once search software was available, searches of CA Condensates became a commercial business.

Among the more successful of the early commercial searching systems were those developed by Illinois Institute of Technology Research Institute (IITRI) in Chicago and the United Kingdom Chemical Information Service (UKCIS) in Nottingham, a project of the Royal Society of Chemistry. Through an internship program, CAS hosted chemical information researchers from several countries. Many of them returned to their national chemical societies to open computer-based chemical information centers. Eventually there were centers in the UK (UKCIS); Belgium, France, the Netherlands, Germany and Finland.

Apart from the national chemical society centers, an industrial cooperation was started in Basel, Switzerland, then home of three of the largest pharmaceutical companies: Ciba-Geigy, Sandoz and Hoffman-LaRoche. The Basel Center for Chemical Information (BASIC) was both an information center running searches and a research center developing new ways to search computer-based chemical information, particularly chemical structure information of interest to the pharmaceutical industry.

The Move to Online

To understand the impact and advantages of online searching, it is useful to review the batch (or offline) searching described above.

Typically, the person desiring the search (the “end-user”) engaged the help of a searcher who understood how to translate the concepts of the user’s search query into the terminology and commands required by the search software. All of the systems were different and quite idiosyncratic. Once a sufficient number of such queries were ready, they were run as a “batch.” The software passed each of queries against the database that was on magnetic tape and, therefore, processed linearly as a sequence of records processed one record at a time. Records on the database that responded to a given query were printed and sent to the user for evaluation.

Surprising and irrelevant responses were a common problem causing the user the request another translation of the query terminology to be re-submitted to another pass of the database in hopes of reducing the irrelevant responses. The process was time-consuming, often taking days to complete as further iterations were tried to produce a better set of responses.

Clearly, speeding-up the process was not only desirable but also valuable to the user.

By the early 1970’s the price of mainframe computers and mass storage devices had declined while the performance was markedly advanced. The real driving force for online searching was the appearance of packet-switched telecommunications networks. Communication with computers was developed

but it required dedicated telecommunications lines that were exclusive to the single link between one user and one computer. Packet-switching obviated all of that by “packaging” telecommunications so that the resulting packets of information could be sequenced and exchanged using any of myriads of telecommunications lines between user and computer.

Mass-storage devices are, by their nature, random-access devices. The need to process records sequentially was no longer a requirement. That coupled with faster and cheaper telecommunication made online searching a reality permitting users to enter their own searches and obtain results directly. Not only was this faster but it opened the opportunity for rapid refinement of searches to obtain a more relevant results. Instead of days to refine and complete a search, only minutes were necessary.

The only remaining problem for the user or information consumer was to learn the method to translate a query into some form meaningful to the computer running the online search software.

Early Online Systems

It is important to remember that the barriers to online information offerings were primarily costs of processing and storage equipment. Accordingly, some early efforts were funded by US federal government agencies. A particularly important example of this is the work of the National Library of Medicine (NLM) resulting in Medline and Chemline. This is not the only example but it is a good one as it set the stage for other developments and the expectations of the commercial customer community.

Both Medline and Chemline were derived from databases produced by CAS. As part of the long-running NSF grant (*vide supra*) used to fund the development of computer processing at CAS, it was agreed that “products” of that system would be created and made available to third parties. Among the early adopters was NLM that created search software that could be used in an online mode. The Medline file was derived in part from the CAS product CBAC (Chemical Biological Abstracts) that contained bibliographic information and abstracts in addition to detailed chemical substance information.

Because of the constraints on just how much processing and storage equipment could be afforded, NLM took the step of producing Medline’s companion Chemline. This was a separate database composed of just the chemical substance information found in CBAC. The most-commonly cited chemical substance in CBAC at that time was *D*-glucose. Rather than repeat the chemical substance identifiers (such as name, molecular formula, etc.) in each Medline record, Chemline held that information only once. It was linked to each Medline record using the CAS Registry Number. This separation of chemical subject and chemical substance information mirrored the way CAS produced the information. The CAS Registry Number is a numerical identifier unique to a specific chemical substance without conveying any structural information. For internal processing purposes, it was economical to process and store chemical substance information separately. For products released to third parties, it was easy to add the complete

chemical substance information for each occurrence in the product through linkage using the CAS Registry Number.

Although initiated as a convenience, the separation of chemical substance information from chemical subject (or text) information led to development of separate search techniques for the two distinct types of information.

The Transition

Now the stage was set for the entrance of the large commercial online services offering chemical information. Computing power and mass storage devices were getting less expensive and their capabilities were constantly and dramatically increasing. Packet-switched networks were a proven technology and several providers offered public access albeit dial-up but with increasing speeds as modem technology improved.

As this covers only the transition to such services, two examples will be discussed: System Development Corporation (SDC) and Lockheed Information Systems. These were widely available and shared some similar characteristics as well as exhibiting significant differences for competitive advantage. Each was developed as a unique computer program to be applied to general problems of online information retrieval. They shared a basic architecture of “accession number”, “inverted file” and “linear file”. The accession number is a carefully chosen, unique number for each complete record in the database. The inverted file is a list of accession numbers that contain a given search term. It can be compared to back-of-the-book index where index terms are presented with a list of page numbers (think “accession numbers”) on which the index term appears. Search terms in a query were compared to index terms in the inverted file to produce sets of accession numbers. Index terms were identified by the field from which they were taken (e.g., document title, author names, keyword phrases, etc). The real value in these systems was in the choice of index terms to be listed in the inverted file. Finally, the linear file has a complete displayable record for each accession number. Major contributors to this transition were Carlos Cuadra of SDC and Roger Summit of Lockheed Information Systems.

They were command-driven systems where functions were initiated through a command line where the user provided the command usually followed by one or more operands. They were Boolean search systems where search terms (“keywords”) were linked by the Boolean AND or OR with notable extensions to be described later. The Boolean OR was always the inclusive OR. The Boolean NOT was also available but its use was advised only with great caution as it could lead to the unintended elimination of relevant items.

The Boolean logic operators used in online searching have a property known as commutation. That is to say that the order of operands is not significant as in arithmetic addition or multiplication (e.g. $1+2 = 2+1$ or $3 \times 4 = 4 \times 3$). Thus, keywordA **AND** keywordB gives the same result as keywordB **AND** keywordA. The same is true for the Boolean OR. The Boolean NOT does not commute.

Operations resulted in “sets” that were uniquely identified and were conformable with other sets in certain operations. Sets were simply lists of

the accession numbers of items responding to a query. As such they could be combined with Boolean operators giving rise to a subsequent unique set or they could be output in whole or in part. This provided a great advantage as partial results of a larger search could be tried, retained, re-used or ignored without starting over with a refined query against the entire database as was true with batch searching.

Searchers were able to have online access to multiple databases selected by the user. It is fair to say that extensive training was important to use them effectively and both services had training staff and programs. Initially, users paid by the “connect hour” (literally wall-clock time of the online connection to the database calculated to some fraction of an hour) and for output printed and sent to the user by the postal service.

Each of these systems had a means to view the alphabetical listing of search terms thereby giving the user insight into possible search terms.

System Development Corporation

The Pharmaceutical Manufacturers Association (or PMA as it was then known) engaged in an exclusive contract whereby SDC would provide online search service using CA Condensates that was at the time quite large among databases and was certainly the largest chemical information database. The exclusive agreement meant that the commercial risk was reduced for SDC as PMA paid certain costs to make the CA Condensates database available. This arrangement did not preclude others from offering online access to CA Condensates as CAS had a non-exclusive licensing policy. That such an agreement was reached is testament to the high desirability of accessing chemical information online. Due to the demands on the available storage resources, the CA Condensates service was online only part-time. The SDC Search Service with its software known as “Orbit” was available online for most of the day with different databases available during different blocks of time. Over time this changed so that all of the databases were available simultaneously and continuously.

For some searches this non-discriminating property could result in excessive and unwieldy sets of results. SDC devised a means of making a search more precise using its STRINGSEARCH command that could be applied to any set formed with Boolean operators. The operand for the STRINGSEARCH command was a literal string of alphanumeric characters placed in quotation marks. The result was another set containing only those records that had the exact string of characters requested. Typically, this was a significantly smaller set as the records found were more constrained than the starting set formed by Boolean operations.

This particular command was useful in chemical information searching because it was possible to find responses to a query consisting of embedded characters. Accordingly, it was possible to find “chloro” in “dichloro” or “trichloro.” Indeed, the customer community used it in exactly that way and increased relevant answers to queries.

There were some problems with this approach. First, was how to form the set against which the STRINGSEARCH command could be used. In the simple

example above it was not possible to form a completely inclusive set using an initial search term of “chloro” because that alone missed both “dichloro” and “trichloro.” Second, was the potential ambiguity of the string being searched. For example, “ethyl” invariably found “methyl.” Lastly, processing this command was slow and, therefore, costly. Still, the command was quite popular and customers always asked other search services to implement something like it.

Lockheed Information Systems

Bearing in mind the basic architecture described above, this system, with its software known as DIALOG[®], was well-adapted to offering chemical information. The first database was CA Condensates that appeared after it was available from SDC. Eventually, almost all of the databases that were licensed by CAS were made available. Because of the size and scope of that amount of chemical information, several unique strategies were used to accommodate the breadth of information.

By 1980, CAS offered for license its CASIA (Chemical Abstracts Subject Index Alert) database. The records on this database complimented those on CA Condensates as they provided both the General Subject and Chemical Substance index entries found in the printed indexes to *Chemical Abstracts*. Although these records appeared from six to ten or more weeks after they appeared on CA Condensates, they had exactly the same CA abstract numbers and could be successfully matched to the corresponding records on CA Condensates that contained the bibliographic information and keywords.

The in-depth indexing of CAS was in high demand by the customer community. For General Subjects, the index terms were from a controlled vocabulary that was applied in accordance with indexing rules. Rules for the use of General Subject headings were described in the CA Index Guide. The vast majority of index entries had an uncontrolled vocabulary modifying phrase to describe further the use of the heading for the document being indexed. Using all of this information more than doubled the number of index terms in the inverted file and increased the lists of accession numbers for any index term.

This system also had a feature of connecting related terms to any index term in the inverted file. To assist customers the CA Index Guide was used as a source to provide alternatives and preferred terms to the online user.

The non-discriminating property of commutation of Boolean operators was always present. To provide compensation for this property, adjacency operators were used. By far the best known was the operator (W) that was placed between operands. This meant that keywordA (W) keywordB was not the same as keywordB (W) keywordA. The (W) operator was a Boolean AND that did *not* commute. Order mattered. Using the (W) operator provided more precise search statements. Additionally, this had great advantage that the search was run against the entire inverted file, not against a previously determined set; the resulting sets were created based upon the ordered occurrence of the search terms in the entire database. To implement such a feature required the ability to process and store enormous sets as each search term had to have attached not only the accession

number of the record from which it came but also the position of the term within the record.

There was a range of adjacency operators most of which were less precise and were effectively Boolean AND with commutation but with restrictions on the location of terms in the same field (e.g., document title, CA index entry, etc.). The (T) adjacency operator was as precise as (W) but with special characteristics brought about by chemical substance name searching. It provided a means to locate search terms within the same chemical name preventing retrieval where the search terms were in different chemical names.

One of the ancillary projects at CAS in the production of "Chemical Titles" was a chemical name segmentation algorithm. Dissection of long chemical substance names produced chemically significant parts. For example, a term such as "dichloroethylmethyl" could be reliably reduced to di#chloro#ethyl#methyl (where # is used to indicate a segmentation point). Each of the chemically significant segments was placed in the inverted file for direct searching. The (T) adjacency operator required that the search term appear in the same original term. Some ambiguity arose because the example term would respond to di(T)methyl because both segments were in the same original term. Still, the direct access to each of the chemically significant segments proved advantageous as one could also search for ethyl(T)methyl. Such searching was useful for both uncontrolled chemical substance names where the order of the chemically significant segments was unpredictable and in controlled chemical substances names where the order was pre-determined.

Inventions

In spite of technological advances, in the late 1970's it became clear that file sizes were growing quickly and maintaining tractable search times and costs were important in the commercial online business. In searching for ways to reduce search effort and costs for users, an interesting fact appeared from the use of the CASIA database. Some number, less than 23%, of all chemical substances covered by CAS were referenced more than once. This was true over decades and millions of publications covered and millions of chemical substances. Some core group of chemical substances was often referenced and the size of the group grew rather slowly. With this knowledge it was easy to separate the large body of chemical substance information into one group that grew rather slowly while the number of singly-indexed chemical substances grew explosively.

Separating the chemical substances from textual information offered numerous advantages to both the online service provider and to the user. The textual information consisting of general subject indexes, their modifying phrases, titles, authors, keywords and bibliographic information had well-developed online text searching methods. Indeed, most databases accessible online were purely textual in nature. It was the chemical substances that offered the challenges due to sheer numbers of them and the precise descriptors used. Using the CAS and NLM models of separating the search for chemical substances from the search for text

was an acceptable and beneficial compromise. The link was the well-established CAS Registry Number.

In the online text databases derived from CAS information (primarily the CASIA database of General Subject and Chemical Substance indexes entries), there was an unusual problem that arose with the indexing of huge numbers of chemical substances. In the CAS printed indexes it was customary to report preparation of chemical substances with an entry consisting *only* of the CAS systematic chemical name and the CAS Registry Number. Occasionally, but not often, additional words such as “prepn.” or “synthesis” were used but were not required.

For years it was understood by users of these printed indexes that such minimal Chemical Substance Index entries signaled the preparation of the named chemical substance. In the online search environment it was impossible to convey to the searcher that the absence of words was meaningful. For this reason and on advice and consent of CAS staff, the letter “P” for preparation was added to CAS Registry Numbers in the conversion process when no other words were present. The letter “S” for synthesis was considered but it resembled the numeral “5” and could be confusing. This convention was later adopted by other online services.

Challenges with chemical substances were large as the CAS Registry Nomenclature File (RNF) contained every chemical substance known to CAS and the number of them grew rapidly. At the time of the transition to online searching, the full scope of CAS coverage and indexing policies were little understood by users. Converting this information to online searching focused more attention on coverage and policies because search results were full of apparently irrelevant results. In fact, the seemingly irrelevant results were valid responses to well-constructed searches and were useful. Not all chemical substances appearing in CAS products are well-defined. There are addition compounds, alloys, coordination compounds, mixtures and polymers. The methods of naming and, therefore, searching such chemical substances is quite challenging.

Some straightforward searches yielded little, no or incorrect results due to policies not well understood. For example, the molecular formula for table salt (sodium chloride) does not exist as NaCl in Chemical Abstracts Indexes. The “Hill Order” for molecular formulae has always been used by CAS. For carbon-containing compounds, carbon is cited first followed by hydrogen (as H) and all other elements cited alphabetically by element symbol (including D and T for the isotopes of hydrogen); without carbon, element symbols are placed alphabetically. Thus, the molecular formula for sodium chloride is properly given as ClNa.

As illustrated above, chemical substance nomenclature is characterized by the repeated use of a relatively small number of unique terms. Once it was possible to license the entire CAS RNF for the millions of chemical substances registered by CAS, this became even more evident. Something more than just transforming words from a database to search terms was needed.

Some guiding principles became evident. Even though there were millions of systematic chemical substance names, a relatively small number of systematic nomenclature terms were used to correctly and completely name them using the rules of CAS systematic nomenclature. Combining search terms with huge

numbers of valid responses permitted a large collection to be precisely reduced to a tractable number of chemical substances. Therefore, one strategy was to generate useful search terms without regard to the large number of chemical substances to which they applied. Another important strategy was to generate search terms that worked to eliminate some chemical substances.

Much of the information about a chemical substance, particularly one that has a known structure, is inferred by our chemical knowledge. For example, if a heterocycle is seen in a chemical structure or explicit in a systematic name, this is obvious to a chemist. In the online search environment using systematic chemical names one could simply enter all of the possible heterocycle names and obtain a useful answer. Although possible, such a strategy is impractical. Identifiers in the RNF permitted the generation of “higher-level” terms that collected all heterocycles under a single search term “HETEROCYCLE.” This single term, not part of the actual chemical substance nomenclature, permitted the collection of all identified heterocycles irrespective of the hetero-atom present. It had the additional property of eliminating from a search all heterocycles by using it as the object of a Boolean NOT. With hundreds of thousands of heterocycle chemical substances, this search term in combination with other search terms proved useful.

At the risk of being pedantic, the generation of heterocycle terms was expanded to include the hetero-atoms from the small set of Nitrogen, Oxygen, Phosphorous, and Sulfur. “HETEROCYCLE-N” was used to describe all chemical substances with heterocyclic nitrogen irrespective of the ring nomenclature. Similar, O, P, and S terms were generated. Again, there were from tens to hundreds of thousands of chemical substances for each of these terms but in combination with other terms they were useful. The ability to include or exclude certain heterocycles was also useful. Rather than relying on Boolean NOT logic that could be misleading, other terms were “pre-coordinated.” Thus, “HETEROCYCLE-NS” was used to describe heterocycles containing nitrogen and sulfur but neither oxygen nor phosphorous.

Another useful generation of search terms came from molecular formulae. Element counts were generated with an element symbol such as “C” followed by a four digit number with leading zeroes. Thus, C0012 collected all chemical substances with twelve carbons in the molecular formula. This was done purposefully so that the counts would appear sequentially with C0012 followed by C0013, etc. By doing so, ranges of element counts were directly searchable such as C0012 to C0016. Because of the importance of isotopes of hydrogen, element counts for deuterium (D) and tritium (T) were also generated.

Because carbon is present in well over 94% of all chemical substances with molecular formulae, it was easy to produce the search term C0000 meaning no carbon in the molecular formula. This could be an identifier for “inorganic” substances if one ignores carbonates, for example. Another use of this search term was with a Boolean NOT to mean all chemical substances containing carbon in their molecular formula.

Many other examples abound. Most of them are not directly available as chemical substance searching has moved more toward structure searching. One of the tenets of structure (or sub-structure) searching is to screen a large collection of chemical substances to eliminate those that cannot possibly be answers to a query.

Many of the inventions described above serve this purpose while others specifically include desired characteristics.

Still, all chemical substances have names and, if systematic, the names are fully descriptive of structure. By understanding systematic chemical substance nomenclature, searches can successfully locate relevant answers that transcend some of the finer points of nomenclature such as isotopic substitution.

Summary

The transition in chemical information searching from printed sources to online searching was due to the adaptability of the hardware, telecommunications and software that were decreasing in cost while increasing significantly in capability. Large databases of chemical information including bibliographic general subject and chemical substances were generated to support the growing chemical industry including pharmaceuticals and agrochemicals. Online chemical information search satisfied a growing need for cost-effective searches of high economic value. Adaptations and inventions of better search terms and search techniques enabled increasing amounts of chemical information to be handled at reasonable cost.

The economic factors of this transition have been mainly about hardware. There were, of course, issues about royalties, licensing fees, pricing and competition that influenced the transition but the market was so big and expanding worldwide that it absorbed many of these changes as the benefits of online chemical information were more widely spread.

Acknowledgments

The author wishes to thank all of those mentioned in this review for their enthusiasm, assistance and integrity throughout his chemical information career. Those mentioned specifically are but a fraction of the many colleagues around the world who made his career meaningful.

References

1. Heym, D. R.; Siegel, H.; Steensland, M.; Vo, H. V. *J. Chem. Inf. Comput. Sci.* **1976**, *16* (3), 171–6.

Chapter 4

Looking Back, But Not in Anger

My View of the History and Future of Chemical Information

Engelbert Zass*

ETH Zürich, HCI H 309, CH-8093 Zürich, Switzerland

*E-mail: zass@chem.ethz.ch

The history of chemical information retrieval started a long time ago with printed sources, soon differentiated by their function into primary, secondary, and tertiary literature. With the advent of appropriate technology, they were converted into electronic databases, starting with secondary sources. These impressive developments are illustrated by landmarks and examples. Despite tremendous progress in the last four decades, improvements are still necessary, as traditional sources have lost most of their “must use” reputation in the face of Google and Wikipedia.

The mission for chemical information retrieval was aptly defined shortly before chemistry started as a science in a modern sense by the famous English writer *Samuel Johnson* in 1775: “Knowledge is of two kinds. We know a subject ourselves, or we know where we can find information about it”. In contrast, in times like ours where information overload is much more of a threat than scarcity of information, it may be useful to head the recommendation phrased at about the same time by *Georg Christoph Lichtenberg*, first professor of experimental physics in Germany, in one of his famous aphorisms: “Leute die sehr viel gelesen haben, machen selten grosse Entdeckungen. Ich sage dies nicht zur Entschuldigung der Faulheit, den Erfinden setzt eine weitläufige Selbstbetrachtung der Dinge voraus. Man muss mehr sehen als sich sagen lassen” – in its essence, it states that reading too much may actually inhibit scientific discovery, as this is fostered more by one’s own observation than by being told about interpretations of observations by others. These two quotations set the difficult goal for appropriate

information to support research, to navigate between the *Scylla* of information overload (wasting time and running the danger of becoming too prejudiced) and the *Charybdis* of a too cursory examination of the state of the art (leading to a waste of resources by simple repetition of earlier work).

The Chemical Literature

Chemical information has always been communicated directly by personal discussions, personal letters, public lectures and presentations at conferences, or regular sessions of scientific academies. This was the major route chemical information was exchanged before scientific journals (2) became widely available and dominant in the process. Remarkably, by using modern electronic tools like e-mail, mailing lists (above all the indispensable CHMINF-L inaugurated by G. Wiggins; (3)), or blogs, this direct communication, both in a one-to-one and a one-to-many mode, is increasing in importance.

Important as these informal (i.e. not normally documented) ways are, only formal means of disseminating chemical information will be discussed here. The term “chemical literature” will be used referring as to how information is *organized*, i.e., not restricted to its traditional meaning for the medium print on paper which dominated it until quite recently. For the development of the chemical literature in the print era, important landmarks are shown in Table 1.

Table 1. Landmarks in the History of the Chemical Literature

Primary Literature

1474	first patent law: Venice (Italy)
1665	first scientific journals (2): <i>Journal des Sçavans</i> (Paris) <i>Philosophical Transactions of the Royal Society</i> (London)
1790	first modern US patent: making of pot ash and pearl ash
1778	first chemistry journal: <i>Crells Chemische Annalen</i> (2)
1789	first chemistry journal still published: <i>Annales de chimie, ou recueil de mémoires concernant la chimie et les arts qui en dépendent, (et spécialement la pharmacie)</i>

Secondary Literature

1817	first handbook: Leopold Gmelin, <i>Handbuch der theoretischen Chemie</i> (finally <i>Gmelin Handbook of Inorganic and Organometallic Chemistry</i> (4))
1830	first A & I publication: <i>Pharmaceutisches Central-Blatt</i> (later <i>Chemisches Zentralblatt</i> (5))
1881	Friedrich K. Beilstein, <i>Handbuch der Organischen Chemie</i> (6)
1907	<i>Chemical Abstracts</i> (7)

Chemical information is published (i.e. formally communicated and documented) in the *primary* literature: journal articles, patents, conference proceedings, research reports, theses. When numbers and volume of these publications became too big to be overviewed and read individually by chemists, *secondary* literature was created as a tool to locate the required information in the primary literature, first handbooks, then dedicated abstracting and indexing (A & I) publications (see Table 1). The following discussion is about such secondary sources used for searching.

The *tertiary* literature is less well defined and somewhat difficult to differentiate from secondary literature: while many scientists group handbooks like Beilstein or Gmelin into this category, we restrict tertiary literature to monographs, encyclopedias (e.g., Ullmann, Kirk-Othmer) and handbooks like Houben-Weyl, Patai etc., all dominated by what might be called “prose” in contrast to the highly structured text in the secondary literature. In the secondary literature, there exists a distinct relation between an individual primary publication and the corresponding entry/record (print or database) in the secondary literature. This relation is either one-to-one, as in A & I publications like *Chemical Abstracts* (CA (7, 8)), or in a handbook like Theilheimer’s *Synthetische Methoden der Organischen Chemie* (together with the *Journal of Synthetic Methods* basis for the first reaction database (9, 10)), or a many-to-one relation in the Beilstein Handbook: for a given compound, many chemical and physical properties are collected in a well-structured way, each with the corresponding references to the primary literature. The Gmelin Handbook with its “prose” structure is a borderline case, but counted here as a secondary source. Tertiary literature is characterized by a higher degree of transformation and processing of the primary information than the secondary literature (see Figure 1).

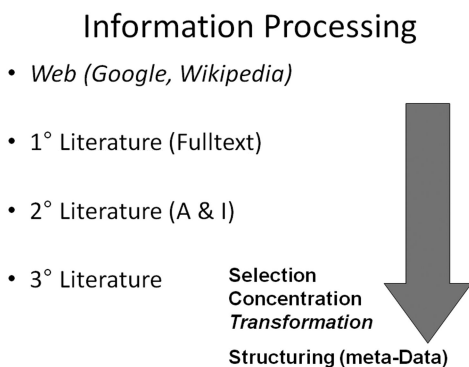


Figure 1. Chemical Literature Categories

Tertiary sources became available much later in electronic form than secondary sources; this may reflect their lesser importance, being more specialized, but it certainly reflects the fact that their content was less amenable to conversion into databases than A & I publications or handbooks like Beilstein or Theilheimer with a highly structured content already in print. Many tertiary sources are e-books and not databases in a stricter sense.

Searching the Printed Chemical Literature

For searching, only secondary or tertiary sources were available in the print era. Among these, A & I sources like CA (7) or the Science Citation Index (SCI (11)) abstracted the primary literature continuously in chronological order, with no content structure or only a minimum (e.g., CA Sections). Within a defined coverage of types of primary sources, there was no differentiation by the kind of content: CA indexed authors, topics (including reactions), compounds (by systematic name and molecular formula), while SCI provided citation links in addition to bibliographic data as an alternative to bringing together publications of similar content by indexing (12). On the other hand, handbooks focus on specific kinds of information: compounds with properties (Beilstein, Gmelin), reactions/synthetic methods (e.g., Theilheimer, Houben-Weyl), or physical properties (Landolt-Börnstein). This information is covered for an entire time range, and presented in a highly structured way easy to perceive. Whatever kind of chemical information handbooks covered, it was arranged in a very systematic way based on chemical principles, highly formalized, but readily understandable by chemists. Such systematics usually covered not only the arrangement of the chemical entities reported, like compounds or reactions, within the handbook, but pertained also to the information about them (13, 14). The price to pay for this structured information, however, was a significant lack of actuality compared to A & I sources.

These different printed secondary sources had therefore distinct missions and uses well known to chemists from their established brand names, and this implied that for many searches, more than one source had to be used.

Figure 2 shows notes taken during a search for the isomeric tetramers of HCN about 1978, just before online database searching became available at the ETH Chemistry Department: from the Beilstein Handbook, only two pages were needed which, when photocopied, contained data about preparation and physical properties with the appropriate references to the primary literature in an ordered, easily perceivable fashion – but this relatively fast result was only achieved with some knowledge about the information structure within the Beilstein Handbook (cf. Figure 3); support for how to search was readily available in form of brochures. For searching printed CA, some knowledge about the appropriate index (12) was necessary, provided in its introduction, or in support publications like CAS Printed Access Tools. The lack of content structure, however, implied a tedious repetitive search over time in Decennial/Collective Indexes, then in Volume Indexes, and finally - restricted to authors and keywords without the very important controlled CA indexing (12) - in issue indexes of the current volume, doing the same search all

over again. One "synergy effect" should be noted here: given the time coverage by Beilstein and the first mentioning in the literature of these compounds, the tedious search in CA started with Vol. 41, not with Vol. 1!

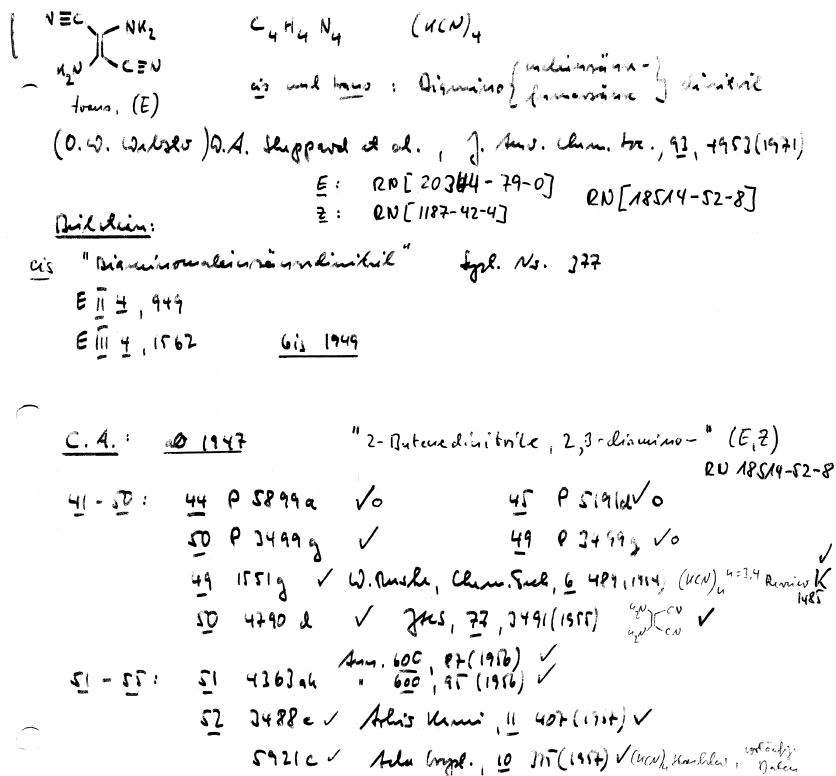


Figure 2. Protocol (part) of a compound search in printed Beilstein and CA

The problem accessing information in the Beilstein Handbook at a given time is illustrated in Figure 3. It should suffice to mention that a multi-volume handbook could provide indexes across all volumes (General-Register, later Centennial Index) only when the entire series had been completed, but the systematic arrangement of entities came here to the rescue: provided one located the volume a given compound had to be reported in according to the Beilstein System (13), the indexes for this individual volume (Gesamt-Register) or subvolumes (Teilband) could be used.

As an aside: this handout was hand-drawn with the help of an IBM Selectric ball typewriter, using the large Orator font – in 1984, the author had not yet access to word-processing or drawing software for producing teaching material.

and author indexes. In 1973, with CASIA (Chemical Abstracts Subject Index Alert (28)), the more thorough indexing information (12) going into the CA Volume Indexes became searchable, first as a separate database, then combined with CA Condensates. The abstract text (29) was available only much later in 1983.

Looking for compounds in printed sources involved systematic names, being plagued by the vagaries and complexities of the index nomenclature one had to use, while the only alternative, searching in molecular formula indexes, suffered from their non-uniqueness, particularly for organic compounds. The transition from printed to electronic secondary sources was therefore kind of crowned with what can be considered the biggest improvement over print, the facility to search for fully or partially defined structures (substructures) of compounds. The two competing (30) systems CAS Online (18) and DARC (19, 20) both used then and now the largest compound inventory, CAS Registry (16).

Compound searches in CAS Registry were limited by the fact that (with few exceptions), any compound appearing only in the primary literature before 1965 was not registered and therefore not searchable. In 1983 CAS tried to raise funds from companies for a Pre-1965 Registration project. This led to a database with limited data (CAS Abstract Number, CAS Registry Numbers of compounds, no other bibliographic data, no indexing) available at STN International as CAOLD, but only back to 1962 (7th Collective Index period). A second attempt was started in 1998 and finally led back all the way to 1907 with virtually the complete information available in 2003 via SciFinder (Scholar) or STN International (31).

CAS Online was operated by CAS themselves, and later transferred to the host STN International (31) where CAS is the major partner. In contrast, in the early phase of electronic information, producers like CAS or the U.S. National Institute of Health for Medline were dependent on companies like Lockheed (DIALOG (32)) or System Development Corporation (ORBIT (33)) to host their already then large and complex databases with appropriate command-driven retrieval languages (17, 34). These and other hosts played a major role in propagating the online use of databases (35) which should not be forgotten in a time when most producers directly offer their databases to users.

With chemical databases running first on terminal-server, then on client-server systems, it is interesting to look at the location of the servers: in the beginning, they were exclusively running at hosts (32, 35) which were usually not the producers of the databases. That changed significantly with the reaction database systems REACCS, ORAC, and SYNLIB (cf. Table 2 (21, 22)). They were the first end-user systems, due to their graphic user interfaces (instead of the so far dominating command-language driven retrieval systems) and their subscription pricing. Characteristic for these and later database systems for end-users like CrossFire (26, 27) Beilstein and Gmelin (cf. Table 2) was also the fact that their servers were operated decentralized and *in-house* by the user's organization. The SCI (11) database was also available for local servers.

For all major chemical databases, this later changed again to servers operated by the producers themselves when dedicated client software was replaced by Web browsers as clients. SciFinder and SciFinder Scholar were always operated by CAS which seems more strongly motivated to completely control its information than other producers.

Table 2. Landmarks in the History of Electronic Chemical Information^a

1960	Chemical Titles (CAS (15)): first computerized information service
1965	CAS Registry System for compounds introduced (16)
1968	CA Condensates (CAS (17)): first major chemical database
1980/81	CAS Online (18), DARC (19, 20): first substructure search systems (based on CAS Registry (16))
1982	REACCS: first graphic user interface for chemists, first reaction database for end-user searching (21, 22)
1986	CJACS (ACS): first e-journal (23)
1991	World Wide Web
1993/94	CrossFire (Beilstein (24, 25)): first major chemical database for end-user searching (26, 27)

^a These landmarks refer to *publicly available* sources; several of these were preceded by similar systems developed for internal use in companies.

Later Enhancements

The field of electronic A & I sources for chemistry was well covered by CAS literature and structure databases, particularly after their integration into the other database offerings of the host STN International (31) founded in 1983, which in due course also made the Science Citation Index (11), BIOSIS, Medline, Derwent World Patent Index and other scientific and patent database available under the common retrieval language STN Messenger, since 1988 supported by the front-end software STN Express (36). Missing in the electronic arsenal up to the late eighties, however, were electronic versions of the handbooks which had played such an important role in the print era, providing users with well-organized compound information, in particular property data, all the way back to the beginnings of chemistry as a science. This was changed starting with an extraordinary meeting of the Board of Trustees of the Beilstein Institute on March 16th, 1982 which decided to examine the feasibility to create a Beilstein database. This became partially available at STN International in 1988 and complete in 1991, followed by the Gmelin database in the same year (25).

Coverage of patents in Beilstein was terminated in 1980; when later the lack of recent patents in the Beilstein database (Gmelin never had them covered) was seen as a disadvantage, MDL (now Elsevier) newly created a Patent Chemistry Database which is now also part of Reaxys (37). This is one of the rather few examples of a major chemistry database not based on a printed predecessor; CASREACT (38) being another prominent example.

While several reaction databases with intellectually selected reactions examples already existed in 1985 (e.g., REACCS, ORAC, SYNLIB; (21, 22); or CRDS (9)), reaction information in CA and Beilstein, by at least an order of magnitude larger, was only present as what may be called “half reactions”: it existed at individual *compound* database entries for the respective reaction partners (starting materials, products, to a much lesser extent reagents and

solvents), but was not linked, and thus not directly searchable as reactions. This was changed in two quite different ways: CAS started in 1985 to build the new, specific reaction database CASREACT (38), made available at STN International in 1988. The Beilstein database, which in 1993 (cf. Table 2) was also offered under the in-house database system CrossFire (26, 27) as the first major general chemistry database with a graphical user interface, was in 1995 extended to reaction searching (39) by linking the “half reactions”. This was made feasible by the fact that preparation information in the handbook in a compound record (i.e., with the reaction *product* thus unambiguously identified by its recorded structure) was in the form of systematic entries “prepared from (systematic name of the starting material)”. This could be transformed in most, but not all cases by automated conversion of chemical names to structures into a structure-searchable full reaction. In this way, also reactions like kinetics, treatment of a compound with a reagent to look at the product distribution with no specific preparative intention, were made available, a specialty of Beilstein in contrast to other reaction sources.

Secondary sources truly dominated electronic information until the mid-nineties of the last century. The only landmark in Table 2 not in this domain is CJACS (Chemical Journals of ACS) as part of CJO (Chemical Journals Online (23)). These (too) early version of an e-journal lacked some important features which about a decade later made e-journals the tremendous success they are: CJO missed tables and graphics (i.e., no complete full text), a graphic user interface (command-driven search by STN Messenger), a suitable price model (pay per view instead of subscription as for print journals), and it was restricted to a single host and thus lacked the universal platform later provided by the Web, an indivisible part of the success of electronic primary literature.

Searching Online

While many younger chemists do not know any more from personal experience how time consuming and tedious chemical information had to be retrieved from printed sources (see Figure 2), their content and many of their traditional forms of organization, indexing rules (12), and data structures are still with us. This concerns not only the legacy information produced before the introduction of electronic data processing, but is also mostly true at present.

Although the primary literature is now to a very large extent available electronically and thus in principle searchable, it is at present too dispersed and not structured enough for direct searching. So, even after the transformation from print to electronic, the search process consists still of two distinct steps: searching secondary (or tertiary) sources for topics, authors, compounds, reactions, physical properties, then identifying relevant references and acquiring the corresponding primary publications.

Up to about 1970, only printed sources could be used for chemical information retrieval (cf. Table 2). When electronic databases became available in the early seventies of the last century, they were at first tools for information specialists only. This restriction was due to the complexity of searching via

command-driven textual user interfaces, the requirement of knowledge about content and data structure of databases, and the price model pay-per-use where one incurred usually charges for connect time to the database, search terms, and for items displayed. Added to this were communication charges which in the beginnings from Europe to the U.S. might run up to almost one-third of the total cost of a search.

About 1985, databases for end-user searching started to become available, with easier to use graphical interfaces and subscription cost models which made the total cost incurred independent of the way a search was executed, and also of the number of searches – both factors are indispensable if a database is intended to be a routine tool for chemists. These databases were then isolated, stand-alone sources. Integration of sources, particularly linking publication records in databases belonging to the domain of the secondary or tertiary literature to the electronic full text of the primary literature (journal articles, patents), which is now taken for granted by users, became only available around the turn of the century. This feature is not only dependent on linking technologies (OpenURL, SFX (40)), gateways like CAS ChemPort (41), or standards like the DOI (Digital Object Identifier (42)), but above all on a sizeable body of electronically available journal articles (43, 44), patents etc. This had to include not only the recent publications already produced electronically for distribution as e-journals on the Web after about 1995, but also backfiles generated later by scanning.

These development phases had a direct influence on searchers: in printed sources, chemists used to search themselves, in the library, assisted either by librarians or by more experienced colleagues; formal education and training were rarely available. Nowadays, with information dominated by the WWW/Internet, with information either directly in the Web (Google (45), Wikipedia), or accessible via the Web (all major databases, e-journals, patents, e-books), chemists search almost entirely themselves, and no longer in the library, but directly at the bench (46).

The intermediate time, from about 1975 to 2000, were the heydays of information specialists and librarians. Chemists at that time not only often had to turn to the library to use databases at all, but due to the cost and complexity of database searches, they had to rely on specialists to do the actual search and also some of the interpretation of the results, setting them in the context of their at that time rather limited capabilities. In addition, librarian assistance was also needed to acquire the necessary primary literature in printed form – until around 1995, electronic information was almost entirely limited to the domain of secondary literature.

Figure 4 shows access routes at the ETH Zurich Chemistry Information Center, the former chemistry library, near the end of this period. A user had to come to the library not only for printed sources which included at that time most of the primary literature, but also for a range of special databases offered locally on CD-ROM. He had to do likewise for a search in CAS databases at STN International, because for reasons of cost and accessibility, we licensed SciFinder Scholar ((47) started in 1997, preceded by SciFinder in 1995) only in 2002. CrossFire Beilstein and Gmelin as well as ISIS (successor to REACCS (48)) with several selected reaction databases, and SpecInfo (49) with C13 NMR

and IR spectra were available already at the chemist's workplace, all running on in-house servers.

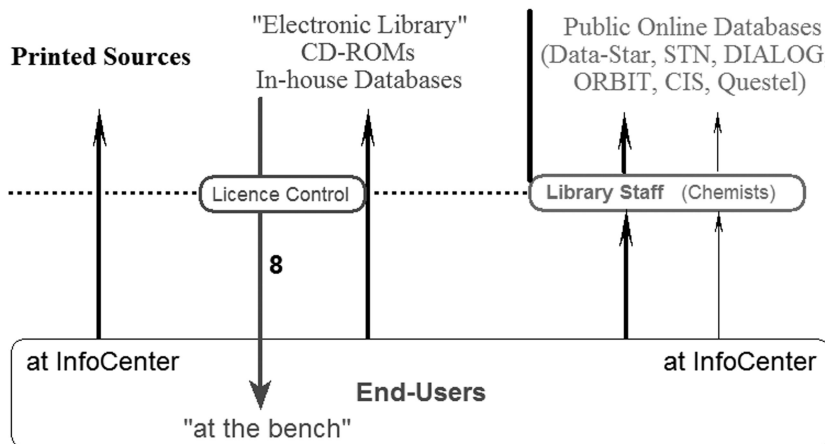


Figure 4. Access to Chemical Information at ETH Zurich around 1998

The access routes shown in Figure 4 changed dramatically quite soon thereafter, with the further propagation of graphic user interfaces and subscription-based pricing, and particularly with the widespread availability of e-journals and their backfiles.

Support, Training, and Education

Printed sources were characterized by distinctive brands focusing on a well-defined mission known to users. This to a certain extent disappeared with the replacement of the printed sources by more or less equivalent databases: nowadays, the CA literature database contains also cited references, formerly an exclusive domain of the Science Citation Index (11). The SCI is still the citation source going farthest back in time, but now it has also Scopus and Google Scholar as competitors (50). CAS Registry contains now measured data and many more calculated data for compounds, and Reaxys (37) with Beilstein and Gmelin offers also titles and abstracts for references, not included in the respective handbooks.

From a marketing point of view, this may look like a good strategy, providing a user with what looks like a one-stop-shop, fitting every need, trying to emulate the success of hosts like STN International (31) in a single system from a single producer. Alas, chemical information is not as simple as that, and from long-time experience, we know that one source alone, even be it SciFinder (47), Reaxys (37), or Web of Knowledge (51), will in many cases not be sufficient (46, 52). Particularly dangerous is the fact that the time coverage of these “add-ons” to traditional sources is not obvious to a user – he probably will not notice that abstracts and titles in Reaxys are practically nonexistent before 1980, and that measured data in CAS Registry became available only in 2002, and although

extended back to 1975 with third-party data, Registry misses a lot of measured data to be found in Reaxys or Springer Materials (Landolt-Börnstein): e.g., of 275'372 compounds with a steroid skeleton retrieved in CAS Registry, only 19 % had any measured data or spectra at all, compared to 80 % of the 225'550 compounds retrieved with the same substructure in Reaxys; for such compounds with melting points, the difference is even more striking: 48 in SciFinder Registry vs. 127'709 in Reaxys (all searches Dec 15th, 2011).

An Example: Searching for Substructures with Properties

The example in Figure 5 searched in March 2011 will further illustrate this: looking for compounds with this substructure, limited to only those which do have NMR data published is quite straightforward in Reaxys, and retrieved 328 such compounds. Using a similar approach in SciFinder with *Refine: Property Availability – Any selected experimental property – NMR spectrum*, the same search found only 36 compounds. While SciFinder permits such property searches only for 14 of the most common spectra/data, this may be done for any of the many more data types present in Reaxys; for an NMR, one might narrow down further by specifying the nucleus or the solvent.

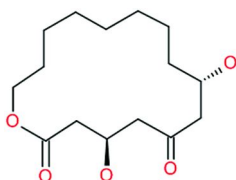


Figure 5. Chiral Substructure

While Beilstein has always covered routine spectra and data, CAS indexed those only if emphasized in the primary publication – but this property *indexing* going much further back in time than the data in Registry is missed if one does not do another, quite different search in SciFinder, this time in the CA literature database. After the same substructure search in Registry for all compounds with the structure in Figure 5, narrowing down to only those references where NMR data may be present for our compounds can be done in no less than three different ways: *Get References – Limit results to: Spectral Properties* gave seven references, all relevant, although NMR had not even been mentioned in our query. Using *Get References – Categorize – Physical Chemistry – Spectra & Spectroscopy – NMR* (four entries, including Overhauser): 9 references, 8 relevant. In a third approach, the total references were narrowed down by *Refine - Research topic: NMR* to no less than 31 refs., where only 21 were relevant, but nine of those had been missed in the earlier searches described here. Comparing the results from CAS Registry (36 substances with NMR) with those of the last search in the CA literature database (25 compounds with NMR), one finds only eight compounds in both searches, i.e., the majority of results in both approaches are exclusive. Not only searching in more than one database system may be

necessary (52), a given database system may also have to be searched in more than one way.

All this should be known to a user doing such combined substructure/property searches, but is it really (46), and where does one acquire this information? There is indeed more training and support available now than ever before, a lot of it by publishers/producers, but one does not get the impression that it is used enough. In addition, as the help messages and other support material provided by producers are of necessity product-oriented and limited to their own products, users are much better served by an appropriate problem-oriented instruction in chemical information and local support. Why does one observe then that in institutions that do offer such education, training and support, often only a minority of users attend?

There are obvious reasons for that: just looking at the awe-inspiring shelves full with printed Chemical Abstracts or Beilstein volumes, printed sources were complex in a very obvious way and perceived as such before their indexes were even consulted. In stark contrast, modern information sources with their graphical user interfaces (GUIs) look very simple in comparison, and the marketing efforts of producers enhance this perception for obvious reasons. In order to further discuss this problem, we have to differentiate between the *use* (operation) and the *utilization* (application) of an information source; this difference is in the author's experience too often ignored.

While present electronic sources are indeed simple to *use*, the assumption that searching in these databases may not really require training and support is definitely false. The core problem lies in the fact that below the easy-looking GUIs lurks the entire complexity of large chemistry databases, with their still important legacy of information going back many decades, their changes in content and indexing policies reflecting both the complexities and also the paradigm changes in the long history of chemistry. In order to *utilize* these sources, the desired information must be harvested from this complex content, and this involves definitely more than just a few mouse clicks.

This problem is aggravated by the fact that the array of essential printed sources to be utilized by chemists in earlier times was much smaller than nowadays plethora of electronic sources, bewildering even to an information specialist.

Excepting some simple, straightforward searches or those only intended to give a first orientation, many chemists thus have problems they may not even properly recognize as such. Therefore, they often do not ask for support, let alone training, when in reality it would be very helpful, if not essential to solve their problem at hand (46). In our experience, in particular academic institutions are challenged by this situation, and must answer it appropriately by readily available support and obligatory instruction (53).

Experiences at ETH Zurich

At ETH Zurich, rather soon after we had started to use databases in 1979, we found that the training offered by producers or hosts did not meet a lot of our requirements: instruction was focused on a single product or a group of products

which producers could not really put into the context of other (competing!) sources, and examples used were often too far away from the area of experience and need of our users. So we started our own courses already in 1981, turning to a regular, two-semester course (one hour per week) covering the entire field of chemical information in 1984. This was offered as an elective, and worked quite well in the beginning, addressing both printed and electronic sources. With increasing end-user searching, attendance went down. In attempts to reach users needing less than such a full course or at least feeling that way, we started in 1995 to offer eight different one-hour courses (two per week, repeated monthly in the library) addressing important sources, e.g., CA on CD-ROM, CrossFire, Current Contents, printed Landolt-Börnstein, and Houben-Weyl, or problems like searching for reactions, data and spectra. These were not as successful as hoped for.

After a lot of experimenting and discussing, we came to the conclusion that only integrating appropriate parts of chemical information instruction directly into lecture and lab courses would be a good solution. We are convinced chemical information instruction will work best when it can answer an immediate need, and when the utility of learning how to search is obvious right on the spot, and not perhaps only weeks later. Being part of an obligatory course, chemical information instruction is then also implicitly obligatory. We found it tough, however, to get into other courses with our programs. Only with the reorganization of the curriculum due to changeover to the Bologna system with Bachelor and Master after 1999, we had a chance to fully realize our designs at ETH. This implied a kind of patchwork instructions which we were not very pleased with in the beginning, doing it out of sheer necessity. We found, however, that by detailed planning, good contacts to the lecturers and teaching assistants overseeing lab courses, and above all the students themselves, needs could be well met, and an entire array of important sources and search types could be covered within the undergraduate curriculum.

With the master part where students are much more “delocalized” in specialized courses, the situation is yet less satisfactory. For those and for Ph.D. students, we used to offer a stand-alone *advanced* course in chemical information at ETH as an elective (only in summer term, 1 hour per week, one credit point). Another means to foster utilization of databases are individual courses for research groups, specifically tailored to their needs. Such courses need a lot of effort to prepare, but that is rewarded by personal contacts and efficiency (54). The needs of students who did their undergraduate work at other universities without the instruction provided at ETH are not yet addressed.

A different approach is being used by the author at the University of Berne and at Innsbruck University: a separate course as the only formal instruction in chemical information is offered as one of the electives students have to take as a master requirement. This implies not only integrated practical searches, but also an examination at the end of the course where students have to do actual searches; Figure 6 shows a typical example.

Students prefer this significantly over being asked written question testing just their knowledge about sources or search strategies. A prerequisite for this were reliable IT infrastructures, and practically unlimited access to Reaxys, SciFinder

etc. for groups of about 20 students for the entire duration of the course and during the one-hour practical examination. The key to success is to make such a course as practice-oriented as possible, and to bravely resist the natural temptation of an information specialist to get as much of his background information to the students as possible.

19.12.2012

7. You are asked to search for **ZrSiO₄**. Write down the exact variant of the **molecular formula** you must use for searching in

a) **Reaxys** [5]

b) **SciFinder** [5]

c) Give a reference for the **Enthalpy of Formation** of this compounds: name of first author, journal title, volume, first page (publication year)
In which **database** did you find this reference ? [10]

Is this article available **electronically** at the Univ. Bern ?

d) Give a reference for a **²⁹Si NMR spectrum** of this compound and the **database** you used to find it [10]

Figure 6. Example of an examination question in a master course

State of the Art

Whoever observed the field of chemical information in the last four or five decades has seen enormous progress and a tremendous *evolution* due to the influence of information technology (55). Outside this domain, we have even seen a *revolution* with the Web, and some revolutionary changes in user behavior, but not yet a revolution in chemical information itself: it still shows the literature categories and types, and the major players as producers, intermediaries and customers of more than a century ago. The majority of databases evolved from print products, and they still carry on the legacy of their content and data structure (52).

Despite all progress made in chemical information, problems significantly impeding access to chemical information are found in every one of the important secondary sources, i.e., the databases used for searching (52). The following discussion will concentrate on CAS databases under SciFinder (47); this is marketed as “the world’s most comprehensive and authoritative source of references, substances and reactions in chemistry and related sciences” on the present starting screen of SciFinder which indeed it is - but this makes solutions to the following problems seem all the more mandatory.

The natural language interface for topic searches (56) in SciFinder (47) is, in the author's experience, not yet really up to the complexity and size of the underlying CA database: e.g., for the search terms "ketone", "oximes", or "olefin", singular and plural are automatically searched for; for olefin, even the synonymous controlled indexing term "alkenes" is automatically included. This is not the case, however, for "porphin" and "porphyrin" – they are not treated as synonyms, and not even their singular and plural are taken care of. One can force the interface by entering "porphin (porphins, porphyrin, porphyrins)" to search all of these (56), but this is limited to a maximum of three terms in parentheses, and a bigger question arises: do SciFinder users know that, or how can they find out? Searching for "total synthesis of estrone", "total synthesis of estron", "total synthesis of oestrone", or "total synthesis of oestron" gives quite different results, in the last phrasing even dramatically different – should this really be so?

The way CAS treats molecular formulas of salts is a very unfortunate legacy from printed formula indexes: normalized dot-disconnected formulas which make a simple compound like manganese(II)sulfate into $\text{H}_2\text{O}_4\text{S.Mn}$, $\text{Ca}_3(\text{PO}_4)_2 \cdot \text{H}_2\text{O}$ becomes $\text{Ca}_{.2/3} \text{H}_3\text{O}_4\text{P.H}_2\text{O}$, and the iron(II) salt of formic acid has to be searched as $\text{CH}_2\text{O}_2.1/2 \text{Fe}$, a real disaster with students. This has been brought to the attention of CAS for more than a decade, unfortunately yet without results. Structure searches are not a good alternative in such cases, as the definitions of charges, valencies and coordination numbers in all major structure databases make their use in searching more of a danger than a tool.

TIS (tabular inorganic substances) which perhaps few outside CAS really understand (57), are another problem area as they describe compounds not only important to inorganic chemistry, but also the material sciences. Although according to the rules set by CAS (known to specialist, but most probably not to SciFinder users), for salts of non-chalcogenic acids like NaCl, a dot-disconnected formula is *not* correct, Cl.Na finds all (?) the TIS of sodium and chlorine because they are registered with such a formula regardless of their actual stoichiometry. It is hard to understand in this context that SciFinder still does not have a search for compounds by their elemental composition, i.e., a given list of elements and the total number of them to be present, regardless of stoichiometry. Both Reaxys and Springer Materials do have that search facility, as does STN Registry.

For nucleotides and peptides, SciFinder does at present not permit sequence or subsequence searches – they can again easily be done in STN Registry (58) with assistance from STN Express (36), but how many academic institutions (in contrast to industry) do continue to have access to the STN/CAS academic program, and still have librarians/specialists knowledgeable in using this alternative, more powerful but also decidedly more complex interface to CAS databases? This is rather unfortunate as we could show to students in courses that the many sources free on the Web they use may lack a significant number of sequences found in CAS Registry (59).

Even with (sub)structure searching, something very central to chemistry, there are common problems where the established computer representation of structures is at severe odds with chemistry: π -complexes, particularly allyl complexes, or metallocenes. At long last, we need structure databases which are user-friendly, not computer-friendly!

As a legacy from the Gmelin database, Reaxys has even more problems here, as a coordination compound may show up with or without coordinating bonds, often with both representations present which must be both searched for separately. Unfortunately, the Gmelin database did originally not contain structures for important compound categories like NaCl composed of single-atom ions, or for solid compounds without discrete ions or molecules. When Gmelin was transferred into Reaxys, structures were formally added for these categories, obviously based on composition data. But this seems to be incomplete: searching for all compounds containing B, F, O, but no other elements with a text search (element symbol, number of elements; Oct 31st, 2013) retrieved 49 compounds in Reaxys, while in a *substructure* search with B F O and IDE.NE = 3, only 30 of those were found.

Preparation and Other Reactions

Once compounds are retrieved in a search, one may need literature about preparation and/or properties; too many users probably rely on one source only in such searches (cf. above). Taking the traditional anesthetic lidocaine (Figure 7) as example in a course (June 2009), and searching for any reaction with this compound as a product, we found appropriate reactions with 23 references in Reaxys in the time range 1946-2008, but only four (1984-2009) in the SciFinder reaction search using CASREACT (cf. ch. 7 in ref. (47)). These results made SciFinder look bad at first glance. However, this is not the only way to search there for preparations, but as the most precise and informative one, it suffers from the relative lack of time coverage of the CASREACT database (38); although this contains some old reaction as far back as 1840, real coverage does not extend back much beyond 1975 - CAS would be well advised to present more informative meta data about this database.

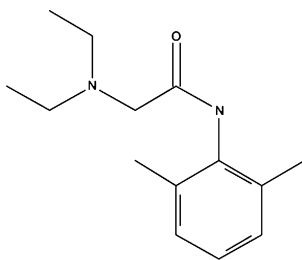


Figure 7. Lidocaine Structure

Given the obvious need to search for preparations via indexing in the CA literature database due to the demonstrated insufficient time coverage of reaction searches, this is neither as straightforward nor as precise as one might desire. A second, simple approach, available only since March 2012, uses *Additional Reactions* in the reaction search; this obviously goes into the CA literature database, but no explanation is unfortunately given for this search mode and its limitations. The probably most obvious and most comprehensive approach uses

Get References – Limit results to: Preparation, but for lidocaine in 2009, only 45 of the 109 references thus found were relevant. The favorite of the author's students was a publication about isolation of lidocaine from horse urine, definitely an analytical paper, but flagged out by CAS as preparation. This was one of many unfortunate incidences where this role is applied much too generously by CAS.

The best way in our experience is to display *Substance Detail* for the compound to be prepared, go to the CAS Role matrix, use the *preparation* line, (but do not click on *preparation*, this will include non-specific derivatives, and gave the same 109 results as above), click on the check marks under *Patents* and *Nonpatents*, respectively. Of course, this is not feasible in one run when one intends to do this for a group of compounds instead of a single one. Should looking for preparations in SciFinder not be easier and more precise?

On the other hand, reaction searching (10) in SciFinder can be more powerful than it actually looks: when trying to search the reaction substructure in Figure 8, one gets the message “stereo bonds will be ignored in reaction searching” – as indeed they are when one insists on searching directly with this query. A user should not wonder then how CAS could produce a large reaction database without the facility of stereosearch, but use the linking of the CAS databases to circumnavigate this problem:

1. Search for the chiral product (e.g., the cis-diol) or starting material – a substructure search for compounds in CAS Registry does have stereosearch
2. *Get Reactions – Limit results to reaction role: Product (or Reactant)*
3. *Refine by Reaction Structure*, using the original query (stereochemistry will still be ignored, but we got at it already by the backdoor)

If both sides of the reaction involve chiral compounds, save the result from step 2, repeat steps 1 and 2 for the other side of the reaction, and combine the result sets

If one retrieves too many irrelevant results (and only then!), use reaction center mapping in step 3

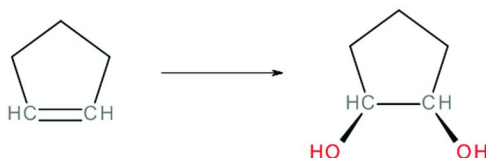


Figure 8. Reaction Substructure Search

In the spring term 2010, we thus retrieved 21 reactions with 14 different relevant references; compared to a Reaxys search (4 reactions, 16 refs.), eight references were only found in SciFinder, ten were exclusive to Reaxys. An investigation of the reagents used in all these cis-hydroxylations exemplified further that any chemist looking for a good result would be well advised to execute such a search in *both* SciFinder and Reaxys, use of additional sources

covering synthetic methodology in a more general sense like Science of Synthesis (60) notwithstanding.

Information specialists can probably live with the above problems in CA databases used as an example, and the many others in other sources; even end-users perhaps could if properly instructed – but most of them are probably not, and do we really want to spend precious instruction time on fixing even those problems that could and should be fixed at the sources?

Outlook

Chemists seem to be more conservative than other scientists in both their publishing as well as their searching/reading habits: While preprint servers (61) like arXiv.org (62) play an important role in physics and elsewhere, an attempt to establish a Chemistry Preprint Server (63) in 2000 was already given up only four years later, due to the reluctance of chemists to use this new communication channel, as well as to refusal of major publishers to accept manuscripts for publication already made available as preprints. While major secondary sources in the medical and pharmaceutical sciences (PubMed), molecular biology, or biochemistry are available on the Web for free (64), only PubChem comes to one's mind in chemistry (45, 65). Any outlook into future developments of chemical information must consider this. The stakeholders of the present system resisting major changes are not only the publishers, often attacked because of their profit interests, but in a publication system that is closely tied to a chemist's reputation and career, many chemists seem not so eager for changes as well – at least as long as the ever-increasing bill for access to information is continued to be paid for by their libraries.

A really novel system of chemical information will probably use as building blocks the primary literature (already available, but not yet readily accessible in machine-searchable form to a large extent (46)), combined with federated searching (46, 66), automatic entity recognition (67), meta data and mark-up of publications (68), machine-indexing (69), ontologies (70), and more open ways (71) to access information (46) as well as research data (72) and software. The feasibility of such a system has already been illustrated in an exemplary fashion by the seminal activities of Henry Rzepa (73), Peter Murray-Rust (74), and other pioneers. Of course, what is technically feasible may be politically problematic, given the resilient character of the established chemical information system, and in particular the present copyright practice where authors have to forsake their rights to publishers - but economic pressure as well as a differently minded coming generation of chemists may overcome this sooner than foreseeable at present.

Trying to look ahead further from the present functions of primary, secondary and tertiary sources in chemical information, it is very obvious that journals and patents will survive into the foreseeable future. The electronic medium will become ever more dominant or even exclusive (46). Although it is to be expected that direct exchanges of primary research data on new channels - blog-like, access to data archives (46, 73) - now in its infancy will become more

important, the existence and even the principle content structure of journal articles will be preserved, but significantly augmented with features mentioned above. Notwithstanding efforts of other publishers, the Royal Society of Chemistry seems to be taking a pioneering role here (75).

With the ever increasing amount of information, *tertiary sources* where somebody already pre-digested the masses of information also seem to have a good future. We all know that the best things to come across in a literature search are good review publications.

But the abovementioned technologies improving the primary literature have also the potential to make *secondary* sources as we have known them obsolete. The literature species thus endangered (52) encompasses SciFinder, Reaxys, Web of Knowledge etc., the big players – but can we really exclude a Titanic Syndrome for databases?

In the days of print, nobody questioned the value and necessity of libraries, or information products like Chemical Abstracts, Beilstein etc. They did not need a lot of marketing, their value, their indispensability and usage were clear to most chemists already as students, and propagated by word of mouth – that is how this author learned about chemical information when he started undergraduate work in 1968.

Propagation among present-day students, however, is dominated by Google and Wikipedia. This entitles serious consequences not only for libraries, but also for producers of traditional sources. The problem of conveying their unique selling points to potential users is aggravated by the fact that the former brand names which in the old days carried everything were given up: Chemical Abstracts replaced by SciFinder (47), Beilstein and Gmelin now part of Reaxys (37), Houben-Weyl became Science of Synthesis (60). Apart from problematic legacies of the print area like those mentioned above, from the point of view of a typical Google user, SciFinder, Reaxys, Web of Knowledge (51) *et al.* play all in the same league as Google (45), but they possess rather complex interfaces (46). Indeed, from a user's perspective, at present too many sources are needed, there are too many differences between sources serving the same purpose, and the cost & effort vs. utility ratio is not obvious enough. Wherever ease of use cannot be improved further along these lines (actually, it still can!), this must be made clear to a user, and he must be supported appropriately.

The best argument for traditional secondary sources is of course their *reliable* coverage of the primary literature and their *quality*. But quality must be perceived by users as such, and by the administrators who have to pay for it. Marketing hype, seen too often nowadays in a market with a lot of money to spend, will not solve this problem. The proof of the pudding is in the eating, the proof of database quality is in sufficient *meta data* readily available about their content, strengths, and particularly, limitations; only to a much lesser extent is quality recognizable in search results themselves, as these are often difficult to judge at first sight concerning their relative utility and comprehensiveness.

Unfortunately, none of the major chemistry databases is marketed yet in a way really conducive to critical, optimal use; for this, much more information about coverage (by time and sources of the primary literature), data structure, and the unavoidable changes in indexing policies (12) must be communicated to all

users in such a way that it cannot be overlooked. This information is definitely insufficient for *all* major databases at present. Given this black-box character, small surprise that students then take rather to the largest black boxes of them all, Google and Google Scholar!

A recipe for a future high-quality information supply in chemistry is naturally much easier to formulate than to realize: Get rid of legacies from print which impair access, use technology in a more innovative way, improve further both user interfaces (*46*) and content, and improve above all communication *to* (meta data) and feedback *from* users.

Notes

Based on a presentation at the 244th ACS National Meeting (Philadelphia, PA, Aug. 20th, 2012), and on *Chemical Literature* (A Celebration of the History of Chemical Information, RSC, London, Nov. 29th, 2010). For a related publication focusing on the influence of new technologies in libraries, see ref. (*1*).

References

1. Zass, E.; Brändle, M.: Information Services in Academia: The Impact of Changing Technology. In *Proc. Int. Chem. Inf. Conf. Nimes*; Collier, H., Ed.; Infonortics: Tetbury, U.K., 2003; pp 15–34.
2. Kronick, D. A. *A History of Scientific & Technical Periodicals. The Origins and Development of the Scientific and Technical Press 1665-1790*, 2nd ed.; Scarecrow Press Inc.: Metuchen, NJ, 1976.
3. Wiggins, G. J. *Am. Soc. Inf. Sci.* **1995**, *46*, 614–617.
4. Lippert, W. J. *Chem. Inf. Comput. Sci.* **1979**, *19*, 201–205.
5. Weiske, C. *Chem. Ber.* **1973**, *106* (4), I–XVI.
6. Luckenbach, R. J. *Chem. Inf. Comput. Sci.* **1981**, *21*, 82–83.
7. Powell, E. C. *Sci. Technol. Libr.* **2000**, *18*, 93–110.
8. Schenck, R. J.; Zapiecki, K. R. Back to the Future: CAS and the Shape of Chemical Information To Come. In *The Future of the History of Chemical Information*; McEwen, L. R., Buntrock, R. E., Eds.; ACS Symposium Series 1164; American Chemical Society: Washington, DC, 2014; Chapter 9.
9. Finch, A. F. J. *Chem. Inf. Comput. Sci.* **1986**, *26*, 17–22.
10. Grethe, G. The History of Chemical Reactions Information, Past, Present and Future. In *The Future of the History of Chemical Information*; McEwen, L. R., Buntrock, R. E., Eds.; ACS Symposium Series 1164; American Chemical Society: Washington, DC, 2014; Chapter 6.
11. Garfield, E. J. *Chem. Inf. Comput. Sci.* **1985**, *25*, 170–174.
12. Zaye, D. F.; Metanomski, W. V.; Beach, A. J. J. *Chem. Inf. Comput. Sci.* **1985**, *25*, 392–399.
13. Prager B.; Stern D.; Ilberg K. *System der Organischen Verbindungen. Ein Leitfaden für die Benutzung von Beilsteins Handbuch der Organischen Chemie*; Julius Springer: Berlin, Germany, 1929.

14. Gmelin-Institut für Anorganische Chemie und Grenzgebiete, Ed. *Gmelins Handbuch der Anorganischen Chemie. Systematik der Sachverhalte*, Verlag Chemie: Weinheim, Germany, 1957.
15. Anonymous. *Chem. Eng. News* **1960**, 38 (14), 27–28.
16. Weisgerber, D. W. *J. Am. Soc. Inf. Sci.* **1997**, 48, 349–360.
17. Prewitt, B. G. *J. Chem. Inf. Comput. Sci.* **1975**, 15, 177–183.
18. Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. J. *Chem. Inf. Comput. Sci.* **1983**, 23, 93–102.
19. Dubois, J. E. In *Computer representation and manipulation of chemical information*; Wipke, W. T., Heller, S. R., Feldman, R. J., Hyde, E., Eds.; John Wiley & Sons: New York, 1974; pp 239–264.
20. Attias, R. *J. Chem. Inf. Comput. Sci.* **1983**, 23, 102–108.
21. Zass, E.; Müller, S. *Chimia* **1986**, 40, 38–50.
22. Borkent, J. H.; Oukes, F.; Noordik, J. H. *J. Chem. Inf. Comput. Sci.* **1988**, 28, 148–150.
23. Love, R. A. *Online '86 Conf. Proc.* **1986**, 149–151.
24. Heller, S. R., Ed. *The Beilstein System. Strategies for Effective Searching*; American Chemical Society: Washington, DC, 1998.
25. Lawson, A. J.; Swienty-Busch, J.; Géoui, T.; Evans, D. The Making of Reaxys—Towards Unobstructed Access to Relevant Chemistry Information. In *The Future of the History of Chemical Information*; McEwen, L. R., Buntrock, R. E., Eds.; ACS Symposium Series 1164; American Chemical Society: Washington, DC, 2014; Chapter 8.
26. Wiggins, G. *J. Chem. Inf. Comput. Sci.* **1996**, 36, 764–769.
27. Meehan, P.; Schofield, H. *Online Inf. Rev.* **2001**, 25, 241–249.
28. Weisgerber, D. *Online* **1977**, 1 (1), 52–56.
29. Baker, D. B.; Horiszny, J.-W.; Metanowski, W. V. *J. Chem. Inf. Comput. Sci.* **1980**, 20, 193–201.
30. Meurling, A. *Database* **1990**, 13 (1), 54–63.
31. Baker, D. B. *J. Chem. Inf. Comput. Sci.* **1987**, 27, 55–59.
32. Rusch, P. F. *J. Chem. Inf. Comput. Sci.* **1985**, 25, 192–197.
33. Elchesen, D. R. *J. Am. Soc. Inf. Sci.* **1978**, 29, 55–66.
34. Ross, J. C. *J. Am. Soc. Inf. Sci.* **1979**, 30, 103–106.
35. Bourne, C. P. *J. Am. Soc. Inf. Sci.* **1980**, 31, 155–160.
36. Wolman, Y. *J. Chem. Inf. Comput. Sci.* **1989**, 29, 42–43.
37. Goodman, J. *J. Chem. Inf. Model.* **2009**, 49, 2897–2898.
38. Blake, J. E.; Dana, R. C. *J. Chem. Inf. Comput. Sci.* **1990**, 30, 394–399.
39. Zass, E.; Donner, W.; Sendelbach, J.; Zirz, C. Experience with the Use of Beilstein In-house (CrossFire) in an academic and industrial environment. In *Proc. Int. Chem. Inf. Conf. Nimes 1995*; Collier, H. R., Ed.; Infonortics: Calne, U.K., 1995; pp 134–144.
40. Lagace, N. *Ser. Libr.* **2003**, 44 (1–2), 77–89.
41. Shanbrom, E. Finding the full-text solution on the web. In *Proc. Int. Chem. Inf. Conf. Nimes*; Collier, H., Ed.; Infonortics: Tetbury, U.K., 1998; pp 16–25.
42. Paskin, N. *Interlending Doc. Supply* **1999**, 27, 13–16.
43. Peek, R. P.; Pomerantz, J. P. *Annu. Rev. Inf. Sci. Technol.* **1998**, 33, 321–356.

44. Kling, R.; Callahan, E. *Annu. Rev. Inf. Sci. Technol.* **2003**, 37, 127–177.
45. Marx, W.; Schier, H. *Nachr. Chem.* **2005**, 53, 1228–1232.
46. Wild, D. J.; Beckman, R. The Future of Searching for Chemical Information. In *Chemical Information Mining Facilitating Literature-Based Discovery*; Banville, D. L., Ed.; CRC Press: Boca Raton, FL, 2009; pp 171–184.
47. Ridley, D. D. *Information Retrieval: SciFinder*, 2nd ed.; J. Wiley & Sons: Hoboken, NJ, 2009.
48. Chen, L.; Nourse, J. G.; Christie, B. D.; Leland, B. A.; Grier, D. L. *J. Chem. Inf. Comput. Sci.* **2002**, 42, 1296–1310.
49. Neudert, R.; Penk, M. *J. Chem. Inf. Comput. Sci.* **1996**, 36, 244–248.
50. Jie, L.; Burnham, J. F.; Lemley, T.; Britton, R. M. *J. Electron. Resour. Med. Libr.* **2010**, 7, 196–217.
51. London, S.; Brahmi, F. A. *Med. Ref. Serv. Q.* **2007**, 24 (4), 59–66.
52. Zass, E. *Heterocycles* **2010**, 82, 63–86.
53. Currano, J. N. Teaching Chemical Information for the Future: The More Things Change, the More They Stay the Same. In *The Future of the History of Chemical Information*; McEwen, L. R., Buntrock, R. E., Eds.; ACS Symposium Series 1164; American Chemical Society: Washington, DC, 2014; Chapter 11.
54. Fong, B. L.; Hansen, D. B. *Issues Sci. Technol. Libr.* **2012** (fall issue); DOI: 10.5062/F4V122Q6.
55. Buntrock, R. E. Chemical Information: From Print to the Internet. In *The Future of the History of Chemical Information*; McEwen, L. R., Buntrock, R. E., Eds.; ACS Symposium Series 1164; American Chemical Society: Washington, DC, 2014; Chapter 2.
56. Wagner, A. B. *J. Chem. Inf. Model.* **2006**, 46, 767–774.
57. Wagner, A. B. *Issues Sci. Technol. Libr.* **2011** (winter issue); DOI: 10.5062/F4QJ7F79.
58. STN sequence searching aids. <https://www.cas.org/training/stn/substance> (last accessed June 5, 2014).
59. Andree, P. J.; Harper, M. F.; Nauche, S.; Poolman, R. A.; Shaw, J.; Swinkels, J. C.; Wycherley, S. *World Pat. Inf.* **2008**, 30, 300–308.
60. Mendelsohn, L. D. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 2198–2199.
61. McKiernan, G. *Sci. Technol. Libr.* **2001**, 20, 149–158.
62. Anonymous. *Nat. Photonics* **2012**, 6 (1), 1.
63. Weeks, J. R.; Kuras, J.; Town, W. G.; Vickery, B. A. *J. Chem. Inf. Comput. Sci.* **2002**, 42, 765–766.
64. Fernandez-Suarez, X. M.; Galperin, M. Y. *Nucleic Acids Res.* **2013**, 41 (D1), D1–D7.
65. Baykoucheva, S. *Online* **2007**, 31 (5), 16–20; cf. Morrissey, S. *Chem. Eng. News* **2005**, 83 (17), 5.
66. Eigner-Pitto, V.; Eiblmaier, J. *Abstr. Pap. – ACS Natl. Meet., 238th* **2009**, CINF-062.
67. Jessop, D. M.; Adams, S. E.; Willighagen, E. L.; Hawizy, L.; Murray-Rust, P. *J. Cheminf.* **2011**, 3, 41.
68. Murray-Rust, P.; Rzepa, H. S. *J. Chem. Inf. Comput. Sci.* **1999**, 39, 928–942.

69. Anderson, J. D.; Perez-Carballo J. *Inf. Process. Manage.* **2001**, 37, 231–254**2001**, 37, 255-277.
70. Hastings, J.; Magka, D.; Batchelor, C.; Duan, L.; Stevens, R.; Ennis, M.; Steinbeck, C. *J. Cheminf.* **2012**, 4, 8.
71. Guha, R.; Howard, M. T.; Hutchison, G. R.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J.; Willighagen, E. L. *J. Chem. Inf. Model.* **2006**, 46, 991–998.
72. Murray-Rust, P.; Rzepa, H. S. *J. Digital Inf.* **2004**, 5 (1), <http://journals.tdl.org/jodi/index.php/jodi/article/view/130/128> (last accessed June 5, 2014).
73. Rzepa, H. S. *J. Cheminf.* **2011**, 3, 46; cf. also http://en.wikipedia.org/wiki/Henry_Rzepa (last accessed June 5, 2014).
74. <http://www.ch.cam.ac.uk/person/pm286>; cf. also http://en.wikipedia.org/wiki/Peter_Murray-Rust (last accessed June 5, 2014).
75. Rzepa, H. S.; Casher, O.; Whitaker, B. J. A paradigm shift in chemistry electronic publishing. In *Proc. Int. Chem. Inf. Conf. Nimes*; Collier, H., Ed.; Infonortics: Tetbury, U.K., 1996; pp 141–148; cf. <http://www.rsc.org/Publishing/Journals/guidelines/AuthorGuidelines/AuthoringTools/> (last accessed June 5, 2014).

Chapter 5

Patents and Patent Citation Searching

Edlyn S. Simmons*

**Simmons Patent Information Service, LLC, 4021 Ambleside Dr., Fort Mill,
South Carolina 29707**

***E-mail: edlyns@earthlink.net**

As the second decade of the 21st century progresses, revolutionary changes in patent documentation, patent classification systems, and the availability of patent information are occurring. New patent information resources have resulted from cooperation among patent offices, which are sharing patent prosecution documentation as well as published documents with the public and providing access to patent translation tools and the new Cooperative Patent Classification system and Common Citation Documents.

Patents as Chemical Literature

Patents have been an important part of the chemical literature for centuries. The first patent granted by the United States government was granted to Samuel Hopkins for The Making of Pot Ash and Pearl Ashes on July 31, 1790 (Figure 1). When Chemical Abstracts began publication in 1907, its very first abstracts were for patents. Then, as now, much of the information published in patents never appeared in scientific journals, so access to the chemical information in patents was essential.



The United States.

To all to whom these Presents shall come, Greeting.

Whereas Samuel Hopkins of the City of Philadelphia and State of Pennsylvania hath discovered an Improvement, not known or used before, such Discovery, in the making of Pot ash and Pearl ash by a new Apparatus and Process, that is to say, in the making of Pearl ash 1st by burning the raw Ashes in a Furnace, 2^d by dissolving and boiling them when so burnt in Water, 3^d by drawing off and settling the ley, and 4th by boiling the ley into balls which then are the true Pearl ash; and also in the making of Pot ash by fluxing the Pearl ash so made as aforesaid; which Operation of burning the raw Ashes in a Furnace, preparatory to their Dissolution and boiling in Water, is new, leaves little Residuum; and produces a much greater Quantity of Salt: These are therefore in pursuance of the Act, entitled "An Act to promote the Progress of useful Arts", to grant to the said Samuel Hopkins, his Heirs, Administrators and Assigns, for the Term of fourteen Years, the sole and exclusive Right and Liberty of using, and vending to others the said Discovery, of burning the raw Ashes previous to their being dissolved and boiled in Water, according to the true Intent and Meaning of the Act aforesaid. In Testimony whereof I have caused these Letters to be made patent, and the Seal of the United States to be hereunto affixed Given under my Hand at the City of New York this thirty first Day of July in the Year of our Lord one thousand seven hundred & Ninety.

G. Washington

City of New York July 31st 1790. -

I do hereby certify that the foregoing Letters patent were delivered to me in pursuance of the Act, entitled "An Act to promote the Progress of useful Arts", that I have examined the same, and find them conformable to the said Act.

Edm: Randolph Attorney General for the United States.

Figure 1. The first U.S. chemical patent, granted to Samuel Hopkins, July 31, 1790.

Chemical patents have much in common with journal articles, allowing Chemical Abstracts and other chemical literature databases to index them with only minor modifications of their usual protocols.

- There is a title and an abstract.
- Individuals named as inventors are considered to be authors of the patent specification.
- Patents contain a general discussion of a problem and its solution, similar to the text of a journal article. In a patent document this section is known as the disclosure. There is usually a discussion of earlier work in the field, and there are examples of experimental procedures and their results.
- Many patent documents, but not all, have a list of cited references.

On the other hand, there are significant differences between patents and journal articles. Patents are a form of intellectual property. The inventors of a new, non-obvious and useful product or process agreed to permit a governmental patent office to disclose the details of their research to the public in exchange for a grant of the right to exclude others from practicing their invention for a limited time, most often for 20 years after the filing of a patent application. The patent offices perform prior art searches to ascertain whether the inventions claimed in patent applications describe an innovation that is patentable. In general, patentability requires that the invention be new to the world and sufficiently inventive that differences between the claimed invention and what was known before would not be obvious to a person of ordinary skill in the field of science or technology to which it belongs. As a result of the legal ramifications of the patenting process, patent documents have major differences from journal articles.

- The invention protected by a patent is defined by its Claims, a numbered list of sentences that define the metes and bounds of the invention for which exclusivity is to be granted. The claims submitted when a patent application is filed are examined by the patent office and are often amended before being published in the granted patent.
- A patent publication is not always a unique document. Because patents are granted by national governments or quasi-governmental organizations representing a group of countries, an applicant must apply for a patent in every jurisdiction where exclusivity is desired. This can be done within a year of the first filing date by claiming priority under the Paris Convention for the Protection of Industrial Property (the Paris Convention) or other treaties, conferring the same effective filing date in each of the countries. Most patent offices publish the patent specification, i.e., the text supporting the patent application, 18 months after its effective filing date, and each patent office publishes a granted patent at the time rights are granted. The result is a family of equivalent patent documents with nearly identical disclosures and with claims that may or may not vary significantly in scope.
- Bibliographic information for patent documents include dates and serial numbers relating to the filing and issuance of the patent and standardized

codes that identify the country or international patent issuing authority where the patent has effect.

- Patent rights are owned by individual or corporate patentees, typically the employer of the inventors. In the United States, where the law specifically states that the inventors are the owners of patent rights, inventors usually assign the rights to their employers. The names of assignees or corporate patentees are part of the bibliographic record of the patent. Patents can be reassigned, sold or used as collateral on a loan during their lifetime and they are part of the assets of a company that ceases to exist or the estate of a deceased individual patentee.
- The legal rights associated with a patent application - its legal status - change over time. The patent owner can license the patented technology or sue infringers during the term of a granted patent. The content of a pending patent application is kept in secret during the first 18 months after the priority filing date or until a patent is granted, and the published application serves as a warning that a patent may be granted in the future. In some jurisdictions, third parties can file a formal opposition to a granted patent and have the patent withdrawn or modified by demonstrating that the claims cover unpatentable subject matter, and accused infringers can demonstrate invalidity by showing the claims were described in the prior art. Maintenance fees must be paid to keep the patent rights from lapsing, and the patent expires at the end of its statutory term unless an extension of some kind has been granted. When a document is abandoned without grant or a granted patent lapses or expires, the claimed invention becomes part of the public domain.
- Patent claims use broad generic terminology. Having invented a method for treating headaches with aspirin, sodium acetylsalicylate, the patent could claim “a method for treating pain with esters of substituted benzoic acid or a salt thereof.” This allows the patentee to enforce its patent against competitors who attempt “design around” the claims.
- Chemical substances are often drawn as “Markush structures,” in which a substructure carries substituents selected from one or more groups of required substructures, substituted in fixed or variable positions by a closed set of permitted substructures. Markush structures can encompass millions of specific substances, many not exemplified in the patent specification or any other publication, and all considered to be disclosed in the prior art.

The first page of a typical 21st century patent specification, US 6,962,918 B2, assigned to Lilly Icos LLC, is shown in Figure 2. It contains its own bibliographical information, a drawing of a Markush structure, serial numbers and dates of priority applications and a corresponding PCT publication, and citations to patent and non-patent prior art found in the patent examiner’s search or provided by the applicant during prosecution of the patent. The PCT information points to an earlier publication of the patent specification under the procedures of the Patent Cooperation Treaty, which is discussed below. All of this information can be useful for patent searches of various types - patentability, freedom to

operate, invalidity, competitive intelligence, statistical analysis of various types, and more.



US006962918B2

(12) **United States Patent**
Orme et al.

(10) **Patent No.:** **US 6,962,918 B2**
(45) **Date of Patent:** **Nov. 8, 2005**

(54) **HEXAHYDROPYRAZINO[1'2':1,6]PYRIDO[3,4-B]INDOLE-1,4-DIONES FOR THE TREATMENT OF CARDIOVASCULAR DISORDERS AND ERECTILE DYSFUNCTION**

(75) **Inventors:** **Mark W. Orme**, Seattle, WA (US);
Lisa M. Schultze, Woodinville, WA (US); **Jason Scott Sawyer**, Indianapolis, IN (US); **Alain Claude-Marie Daugan**, Les Ulis (FR); **Raymond Brown**, Fishers, IN (US)

(73) **Assignee:** **Lilly Icos LLC**, Wilmington, DE (US)

(*) **Notice:** Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 127 days.

(21) **Appl. No.:** **10/363,569**

(22) **PCT Filed:** **Sep. 17, 2001**

(86) **PCT No.:** **PCT/US01/28972**

§ 371 (c)(1),
(2), (4) **Date:** **Feb. 27, 2003**

(87) **PCT Pub. No.:** **WO02/28858**

PCT Pub. Date: **Apr. 11, 2002**

(65) **Prior Publication Data**

US 2003/0236263 A1 Dec. 25, 2003

Related U.S. Application Data

(60) Provisional application No. 60/237,477, filed on Oct. 2, 2000.

(51) **Int. Cl.⁷** **C07D 471/14**; C07D 487/14;
A61K 31/4985; A61P 9/12; A61P 15/10

(52) **U.S. Cl.** **514/250**; 544/343

(58) **Field of Search** 544/343; 514/250

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,859,006 A * 1/1999 Daugan 514/249
5,981,527 A * 11/1999 Daugan et al. 514/250
6,140,329 A * 10/2000 Daugan 514/250

FOREIGN PATENT DOCUMENTS

WO	WO 95/19978	7/1995
WO	WO 97/03675	2/1997
WO	WO 9703675 A1 *	2/1997
WO	WO 97/03985	2/1997
WO	WO 02/11706	2/2002

OTHER PUBLICATIONS

Wang et al., *Organic Letters* vol. 1 (10) 1647–1649, 1999.*
Lucas et al. *Pharmacological Reviews* 52 (3), 375–413, 2000.*

West, Anthony R., *Solid State Chemistry and its Applications*, Wiley, New York, 1988, pp. 358 & 365.*

A. Madrigal et al., *Tetrahedron: Asymmetry* 11, 3515–3526 (2000).

H. He et al., *Med. Chem. Res.*, 9:6, 424–437 (1999).

H. Wang et al., *Org. Lett.*, vol. 1, No. 10, 1647–1649 (1999).

S. Edmondson et al., *J. Am. Chem. Soc.*, 121, 2147–2155 (1999).

A. van Loevezijn et al., *Tetrahedron Letters*, 39, 4737–4740 (1998).

A. Madrigal et al., *Tetrahedron: Asymmetry* 9, 3115–3123 (1998).

S.K. Pandey et al., *Tetrahedron*, 57, 4437–4442 (2001).

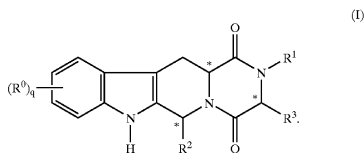
* cited by examiner

Primary Examiner—Venkataraman Balasubramanian

(74) *Attorney, Agent, or Firm*—Marshall, Gerstein & Borun LLP

(57) **ABSTRACT**

Compounds of the general structural formula (I), and use of the compounds and salts and solvates thereof, as therapeutic agents.



18 Claims, No Drawings

Figure 2. The first page of a typical 21st century chemical patent, US 6,962,918 B2, assigned to Lilly Icos LLC, granted November 8, 2005.

Searching for patent purposes presents significant problems that may not be encountered when searching for other purposes. Patentability searching is based on the assumption that all publications in all languages are accessible. Not only patents, but also other forms of technical literature need to be searched. Beyond the need for searching highly technical terminology in all of the world's languages

is the fact that new technologies require new terminology, allowing the authors of patent specifications to create a new vocabulary that may or may not be identical to the vocabulary used by other patent applicants. Simple text searches need to be supplemented by standardized indexing to keep from overlooking relevant references.

When the United Nations created the World Intellectual Property Office (WIPO), work began slowly on standardizing patent documentation (1). The Patent Cooperation Treaty now allows applicants to file a single application for patents in multiple countries, 148 patent offices by the end of 2013, entering the national phase of patent prosecution 2-1/2 years after their priority application date. The first PCT applications were published in 1978, with the country code WO. The European Patent Office also published its first patent applications in 1978, with the country code EP, and has grown to 38 member states by 2013. European patents do not automatically protect inventions in all EPO member states, but must be validated in each country after grant. Negotiations toward a new European patent that would have effect across the European Union came and went over many years, and an agreement was finally reached in 2012 to allow European Patent applicants to opt for a Unitary Patent to be enforced by a Unitary Patent Court. The European Patent Office will begin issuing European Patents with Unitary Effect after 13 member countries have ratified the agreement on the establishment of the Unitary Patent Court.

Individual countries also made progress in harmonizing patent laws and procedures in the last decades of the 20th century. The establishment of the World Trade Organization in 1995 reduced differences in patentable subject matter among countries and established a minimum patent term of 20 years from filing. The Patent Law Treaty brought standardized patenting procedures to signatory countries in 2005. In 1983 the United States, Japanese and European patent offices initiated trilateral cooperation, and in 2008 the trilateral offices were joined by the patent offices of China and Korea to form the IP5 group with the objective of "the elimination of unnecessary duplication of work among the IP5 Offices, the enhancement of patent examination efficiency and quality and guarantee of the stability of patent right" (2). Bilateral agreements among those five and other countries has resulted in the Patent Prosecution Highway, allowing pairs of countries to share the results of examination of equivalent patent applications, saving time and expense for both applicants and patent offices. Although standardized laws and procedures have made current patent documentation easier to understand, these changes and other changes in the patent laws and procedures of individual countries create confusion for anyone attempting comprehensive retrospective searches.

Retrieving Chemical Information from Patents

Classification Systems

Because only one patent can be granted on an invention in any country and a patent can be granted only if the invention has never been described in

a printed publication, patent examiner searches of the prior art are an essential feature of the patenting system. During the 19th century patent offices created classification systems to separate patents into manageable groups for review by patent examiners. Public search facilities were created so that members of the public could search the patent collections to find out whether an invention was patentable or find patents they might infringe before risking a lawsuit by entering a new line of business. The classified files needed to be searched manually. Libraries and law firms could subscribe to the Official Gazette of the United States Patent and Trademark Office or similar official journals from other countries that contained abstracts or exemplary claims from newly issued patents, but these were current awareness tools not designed for retrospective searching. For chemical patent searching, the chemical formula and subject indexes of Chemical Abstracts were invaluable. Demand for less time consuming patent search tools and for access to information from other parts of the world, particularly from chemical and pharmaceutical companies, led to the introduction of new patent indexing services in the middle of the 20th century. The IFI/Plenum Data Co. introduced a dual index of US chemical patents in the form of a book with duplicate pages of patents and corresponding index terms that allowed a user to search for two concepts appearing in a single patent, and Derwent Publications Ltd. began producing English language abstracts of international patents. Derwent's establishment of a classification system for files of abstracts facilitated searching for information from international patent collections; the codes were called Manual Codes to distinguish them from codes used for sorting by machines.

Even more important for international patent searching was WIPO's introduction of the International Patent Classification (IPC) code in 1968. IPC codes are assigned to patents in all technologies by patent offices around the world in addition to any national classification systems used in a particular country. The hierarchical IPC scheme was updated every 5 years, with new subdivisions added to cover new technologies. By contrast, the United States Patent Classification system was under constant review with reclassification of the manual search files at the Public Search Room as new classes were established. Noting that the IPC was insufficiently precise and its revision schedule was too slow, both the European Patent Office and the Japanese Patent Office created modified classification systems. The EPO applied its ECLA system to its internal patent family database, DOCDB, and updated its computerized family records whenever the ECLA system was revised. The Japanese File Index (FI) system was assigned only to Japanese patents.

An attempt was made to correct the deficiencies of the IPC system for its 8th Edition. The Reformed IPC, launched in 2006, had an Advanced Level of classifications, suitable for assignment by large patent offices, and a Core Level suitable for assignment by patent offices that processed few patents. When revisions of the Core Level were made, on a shorter 3-year cycle, the classifications assigned to all patents would be updated in electronic databases, requiring that the patent databases be redesigned to allow updating for efficient retrospective searching. It took only a few years to show that the revised system was not delivering the efficiencies that were hoped for. In 2011 a new IPC

design was introduced, with only one level of specificity, updates on a class by class basis rather than full published revisions, and reclassification of records for retrospective searching (3).

The United States Patent and Trademark Office remained the only major patent granting organization with a classification system that was structured differently from the IPC. In 2010, an agreement was reached between the USPTO and the European Patent Office to form a Cooperative Patent Classification system that would replace both the US classification system and ECLA (4). The CPC, launched in 2013, retains US classes for designs and plants, which are not part of the IPC scheme, and follows the hierarchy of the IPC and the ECLA scheme it replaces. The backlog of ECLA codes in the EPO's DOCDB file has been reclassified and distributed to online patent databases. Both Korea and China have agreed to adopt the CPC, and it may spread to many other patent offices in the future.

Chemical Structure Searching

When digital computers became available for quick sorting of indexed cards in the 1950s, chemical and pharmaceutical companies began developing systems for indexing and retrieving information about chemical substances (5). What we now call "value-added" indexing was the only tool for finding relevant patents. IFI merged its indexing system with a fragmentation code developed by DuPont and a separate fragment system for polymers developed by the Gulf Oil Co., giving access to its coded database exclusively to subscribing corporations (6). Derwent Publications Ltd. (now Thomson Reuters) also developed fragmentation codes for small molecules and polymers and offered access only to corporate subscription holders. Fragmentation codes describe chemical compounds by assigning search terms designating atoms, functional groups and ring systems from a closed or open-ended thesaurus, and they worked remarkably well for patent searching, particularly since Markush structures are built out of structural fragments. By the 1970s, large companies had in-house patent database systems and libraries of printed abstracts and microfilmed patent specifications.

With the introduction of online search services, particularly Questel, Orbit (now merged into Questel), Dialog and STN International, patent databases became widely available. Unrestricted subsets of data in the Derwent World Patents Index and IFI CLAIMS databases were opened to nonsubscribers, and new national patent databases containing first page data and the first claim and/or abstract of patents opened up new means of searching, particularly helpful for inventions without chemical structures as limiting features. Topological searching of Chemical Abstracts Registry file was made searchable through STN and Questel's DARC service, allowing searches for exemplified compounds, though not for Markush structures, by defining the atoms and bonds in a compound's skeleton. Additional research in computerized handling of chemical structures during the 1980s resulted in the introduction of Marpat, a new database of Markush structures from patents indexed for Chemical Abstracts and Markush

DARC, with indexing of Markush structures associated with Derwent records and patents in the French patent office's PharmSearch database (7). Derwent's fragmentation coding continued after the introduction of topological search systems, as it would have been impossible to generate Markush DARC indexing for the millions of patents in the back file.

Searching for chemical substances without costly structural indexing became a reality when full text patent data arrived on the scene. The full text of a patent typically mentions many substances that are of potential interest to searchers but are not important enough in the context of the patent to be indexed, and these substances could not be discovered using chemical structure search tools. Cheaper and more powerful computers and cooperation among patent offices led to an explosion of affordable patent documentation in the late 20th century. INPADOC, a joint venture of WIPO and the government of Austria formed in 1972, began collecting bibliographic information from patent offices around the world, creating a global patent family database. The European Patent Office released its internal bibliographic patent search files as the Espacenet database, allowing the public to search titles, abstracts and bibliographic information and display patent documents at no cost. The INPADOC database, which had been transferred to the European Patent Office, was integrated into Espacenet, where INPADOC's extended patent families can be viewed at the click of a mouse. Many national patent offices began providing full text databases through the Internet, and commercial services opened portals for searching United States, European and Patent Cooperation Treaty publications in combination with the JAPIO database of English language abstracts of Japanese patent applications.

By the second decade of the 21st century, full text patents could be searched on large and small full text databases with widely varying features and prices. Some, like Free Patents Online and Google Patents, are free; others require subscriptions to individual seats or offer subscriptions with access shared among the employees of a large corporation, costing hundreds of thousands of US dollars per year. Many of the databases now have translation functions to assist with searching or reading patent specifications. LexisNexis TotalPatent was the first to provide searchable English translations of patents. Espacenet, WIPO's PatentScope, Orbit.com, and MineSoft's PatBase and other databases allow users to obtain machine translations of displayed documents. The quality of translation varies widely from source to source and among language pairs on a single source. Chemical nomenclature is particularly problematic for translation engines without specialized vocabularies, but there have been great improvements in machine translation of patents in recent years as translation vocabularies can be built by comparing the text of equivalent patents with human translations provided for filing in patent offices that operate in different languages.

Searching full text is still not an efficient way to find chemical substance information. There is no standardization of chemical nomenclature in patents, and many compounds are disclosed as chemical structure drawings, embedded in Markush structures, or described in tables of partial structures. Most of the chemical structure search tools developed over the years for online databases are still available. Although Markush DARC and Marpat are searched by drawing chemical structures, they still require manual input of codes for elements and

functional groups. The cost of manually indexing of millions of patents had taken its toll. IFI had discontinued its high value indexing early in 2011 and opened access to its back file of premium indexing to all STN users in late 2013. Derwent fragmentation coding was being generated algorithmically from Markush DARC indexing, but is still searchable only with a corporate subscription. Markush DARC is available on Questel without restriction to Derwent subscribers, but with a significant search charge. Derwent's topological structure records for specific compounds are searchable on STN in the Derwent Chemical Resource module of the World Patents Index. There is reason for concern that the limited number of persons being trained to use the fragmentation code and the topological search systems threatens the continuing availability of these valuable resources (8).

Research on methods for extracting searchable chemical structures from patents has increased in recent years, and databases with chemical substance information extracted algorithmically from patents have been introduced to the market. Each of the systems use textual extraction of chemical names/or 2-dimensional drawings and converts them to structural records that can be searched with a variety of software tools. The Reaxys system incorporates the MDL Patent Chemistry Database, which has specific substance information from US, European and PCT patent publications, with links to displayed Markush structures. The SureChem system, created by Digital Science and acquired by EMBL to be remarketed as SureChEMBL in late 2013, extracts compounds from patents and deposits them into PubChem, as well as making them available directly (9). IBM's Strategic IP Insight Platform (SIIP, formerly called SIMPLE – Strategic Information Mining Platform for IP Excellence), developed through collaborative efforts with several major life sciences organizations, uses a computer-based algorithmic method for automatically extracting chemical entities from textual content as well as from images and symbols found in US, European and PCT patent publications (10, 11). A searchable database of more than 2.4 million chemical structures and pharmaceutical data extracted from the patents and scientific literature using the SIIP analytics methods was donated by IBM to PubChem and NIH CADD Group, which will support advanced drug discovery efforts in cancer research (12).

As Downs and Barnard (13) point out, "The extent to which automatic extraction of chemical information from patent documents can (either now or in the future) provide databases of equal quality to the traditional manually-curated, "value-added" ones, is arguable." There is no standard for naming substances or drawing chemical structures in patents, and patents often describe substances using a partial structure drawing completed by text terms. As difficult as it is for the mind of a trained chemist to resolve the ambiguities in substance descriptions in patents, it is far more difficult for a computer algorithm. Certainly, the indexing of specific compounds available as of 2013 is no more able to retrieve patents on the basis of chemical structures encompassed by Markush structures such as the one in Figure 2 than is the Chemical Abstracts Registry. IUPAC is currently working toward establishing requirements for extending the applicability of the IUPAC International Chemical Identifier (InChI) to Markush structures (14). It is difficult to imagine how InChI can be taught to retrieve the billions of compounds in real world patents.

Patent Citation Searching

Unlike references cited in the journal literature, patent citations are not always selected by the author/inventor. References cited in patents are provided by a patent examiner, a patent office technical expert whose role is to search the prior art, both patents and non-patent literature, to discover any publications that anticipate the subject matter of the claims or render it obvious. A comprehensive search is not required for this purpose; the end point of the examiner's search is one or more references that teach or suggest the invention defined by the claims. If such references are found, the claims are rejected as unpatentable, and the applicant has an opportunity to amend the claims to delete any unpatentable subject matter or argue for a different interpretation of the references. Along with publications used to reject claims, the examiner may cite references that show the general state of the art found during his or her search or submitted by the applicant or a third party.

Applicants for patents are never required to do a search before they submit their applications, though a thorough patentability search is invaluable as a tool for deciding whether a patent is likely to be granted and for drafting claims that are unlikely to be rejected during prosecution. The United States is unique in requiring that applicants provide material prior art to the Patent and Trademark Office if they are aware of it. Rule 56 (37 Code of Federal Regulations 1.56) imposes the duty of candor on all individuals involved in the filing or prosecution of the patent application, requiring that they provide information they have about prior art relevant to patentability and that any references cited by foreign patent offices be submitted to the USPTO during pendency. Many other patent offices require that applicants submit references cited in corresponding patent applications filed in other countries. After review, those references become part of the record of the patent application. By contrast, publications mentioned in the body of the patent disclosure as background of the invention may or may not be cited as references by an examiner.

Cited references are published by most countries as part of a granted patent. Under the Patent Cooperation Treaty and European Patent application procedures, patent office searches are done before publication of the application and a search report is published with the patent specification 18-months after the priority date or shortly afterward. Patent examiners' citations are not required to follow a standard format, and patent databases usually deal with the lack of standardized formats by omitting non-patent citations from their searchable data. There are significant differences among the references cited by multiple patent offices examining applications claiming the same invention, as each patent office has its own search protocols and somewhat different legal interpretations of novelty and inventiveness.

The idea of making examiners' citations available to the public in a searchable index was introduced by Eugene Garfield in an American Chemical Society talk in 1955 (15), but such an index was slow in coming. Only United States patent citations were indexed in the earliest patent citation database, IFI's CLAIMS Citation database. The United States was the first patent office to publish citations to the prior art on granted patents systematically when it began the practice in

1947, but many other patent offices took up the practice thereafter. Commercial patent databases began to consolidate citations from members of patent families when the Derwent Patent Citations Index was created. The cited reference information in commercial patent databases is extracted from the citation data from published search reports or granted patents. Additional references cited during prosecution of the patent or during opposition could be found in national patent registers, not in bibliographic databases. This has begun to change as patent offices have begun to share citation data with other patent offices.

Patent offices began sharing search results in 2006, when the US and Japan instituted the first Patent Prosecution Highway, a program through which patent applicants could request that the two patent offices share examination results to accelerate grant of their patent in the second country (16). The program has proved to be popular enough that many patent issuing authorities have agreements in place. Now that major patent offices have begun to work together, it has become possible to share this information systematically. In 2012 the IP5 offices began consolidating their search results in the Common Citation Document (CCD) and sharing them not only among the five patent offices but also with the public. These can be viewed at no cost through the EPO's Espacenet database (17).

To searchers outside of the examining corps, cited references are used to supplement subject-matter oriented search parameters. Citation searching is often the best way to identify publications similar to the topic of a search for subject matter that is not reliably indexed by patent databases or has an unreliable vocabulary. Web-based patent search platforms now have links to cited patents, and those that index patents in family records provide all of the citations available in the data feeds used to build the database. This can result in impressive sets of patents that may or may not relate to the aspects of the citing patents that are relevant to the current search. Although patent citations can provide relationships without the limitations of a classification or indexing system, they also can point to subject matter relevant to both citing and cited patents that is totally irrelevant to subject matter relevant to the searcher's interest in the citing patent - for example, a citation search might attempt to find references to methods for formulating printer ink by looking at citations for a known patent in which the ink is applied to labels, but may retrieve patents that describe labeling adhesives.

In addition to their use as sources of references to technical information, many patent analysts believe that citation metrics can be used to judge the value of a patent or a company's patent portfolio. The assumption behind this kind of study is that patent citations have the same significance as citations in scientific journal articles. In journals, citation metrics are assumed to measure the influence of a particular journal article or its author on subsequent research in the field. More important work would be cited more often than less important work, journals that publish more highly cited articles must be more important journals, and authors whose work is cited frequently must be more important scientists. Citation impact calculations generally discount self citations - after all, you can't measure someone's impact on the scientific community by considering the author's references to his own work. Quantitative methods intended to measure the quality or importance of research has been subject to discussion and criticism,

but they are widely used in managing library subscriptions and often in granting tenure.

Since patent citations are assigned by patent examiners and relate almost exclusively to claimed inventions, there is even more reason for skepticism. The real world value of a patent should be measured by the success of the claimed invention in the marketplace, not by its reputation as a publication. A patent with a well developed summary of a technology might be cited often as the technology develops, but the value of the disclosure for rejecting claims does not necessarily correlate with influence of that patent's claims on subsequent developments in the area. Citation of a reference in a patent application is not necessarily an indication that the applicant was aware of the cited reference when the claimed technology was invented and is even less an indication that the cited patent influenced the inventors' work. Also, when using patent citation metrics there is no justification for discounting the importance of "self citations," as the owner of the citing patent might be the only entity with the right to use the claimed invention for further research.

Conclusions

In the not too distant past there was a scarcity of patent information for use in chemical patent information research, and the chemical industry willingly paid for labor intensive fragmentation and topological indexing and provided training in the use of value-added indexing. The 21st century is a time of abundant patent information, and there is a heavy emphasis on technological approaches to information retrieval that allows searching by non-specialists with minimal training. Patent offices have developed incredibly valuable tools to help their examiners with patentability searching, and the public is allowed to benefit from free machine translation, classification and full text resources. Web based commercial search services offer all of the advantages provided by the patent offices at a wide variety of price points, competing on the basis of additional features such as corporate sharing and annotation features, graphing and analytical tools and hit term highlighting.

Indexing and abstracting services still provide the best access to chemical structures, and there is reason for concern that the cost of manual or machine-assisted indexing threatens their survival. Work will continue on development of automatic extraction of chemical structures from patent documents, but the loss of either value-added indexing or trained information scientists may endanger the quality of future patent information research.

For non-chemical searching, much remains to be done. Mechanical and electrical drawings are at the heart of patent disclosures for devices, machinery and electrical circuitry, and systems for indexing and searching drawings are not yet available. The past history of patent searching was focused on chemical structure retrieval, the present history of patent searching is focused on full text, and it is possible that the future history of patent searching will bring us algorithmically assisted drawing retrieval.

References

1. For a detailed description of the features of patent documents see Adams, S. R. *Information Sources in Patents*, 3rd ed.; K. G. Saur: Munich, 2011.
2. <http://www.fiveipoffices.org> (accessed March 7, 2014).
3. http://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide_ipc.pdf (accessed March 7, 2014).
4. <http://www.cooperativepatentclassification.org> (accessed March 7, 2014).
5. Hunsberger, I. M.; Frear, D. E. H.; Harmon, R. E.; Smith, E. G. *Survey of Chemical Notation Systems*; National Academy of Sciences - National Research Council: Washington, DC, 1964; Publication 1150.
6. Donovan, K. M.; Wilhide, B. B. *J. Chem. Inf. Comput. Sci.* **1977**, *17* (3), 139–143; DOI:10.1021/ci60011a008.
7. Simmons, E. S. *World Pat. Inf.* **2003**, *25* (3), 195–202.
8. Geyer, P. *World Pat. Inf.* **2013**, *35* (3), 178–182.
9. <http://chembl.blogspot.com/2013/12/surechembl-chemical-structure.html> (accessed March 7, 2014).
10. Spangler, S.; Ying, C.; Kreulen, J.; Boyer, S.; Griffin, T.; Alba, A.; Kato, L.; Lelescu, A.; Yan, S. Exploratory analytics on patent data sets using the SIMPLE platform. *World Pat. Inf.* **2011**, *33* (4), 328–339; DOI: 10.1016/j.wpi.2011.07.001.
11. Yan, S.; Chen, Y.; Spangler, S. Chemical Name Extraction based on Automatic Training Data Generation and Rich Feature Set. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2013**, *10* (5), 1218–1233.
12. <http://www-03.ibm.com/press/us/en/pressrelease/36180.wss#release> (accessed March 6, 2014).
13. Downs, G. M.; Barnard, J. M. *WIREs Comput. Mol. Sci.* **2011**, *1* (5), 727–741; DOI: 10.1002/wcms.
14. [http://www.iupac.org/nc/home/projects/project-db/project-details.html?tx_wfqbe_pil\[project_nr\]=2009-041-1-800](http://www.iupac.org/nc/home/projects/project-db/project-details.html?tx_wfqbe_pil[project_nr]=2009-041-1-800) (accessed December 8, 2013).
15. Later published as Garfield, E. *JPOS* **1957**, *39* (8), 583–585 and reprinted in *Essays of an Information Scientist*, *6*, 472–484. <http://garfield.library.upenn.edu/essays/v6p472y1983.pdf> (accessed March 7, 2014).
16. http://www.uspto.gov/web/offices/pac/dapp/opla/preognotice/pph_epo.pdf (accessed March 7, 2014).
17. <http://www.epo.org/searching/free/citation.html> (accessed March 7, 2014).

Chapter 6

The History of Chemical Reactions Information, Past, Present and Future

Guenter Grethe*

352 Channing Way, Alameda, California 94502-7409

*E-mail: ggrethe@att.net

The history of information about chemical reactions, their data and applications are broadly described in this chapter. The published information spans nearly four millennia, from the ancient alchemists in Egypt to modern researches in the age of computer technology. Because of the vast amount of information available on chemical reactions in general, most of the material in this chapter deals with organic chemical reactions. The scope of this chapter does not allow for a comprehensive discussion of each topic, but selected examples are mentioned to show the historic development in this field.

Introduction

This chapter attempts to broadly describe the history of information on chemical reactions, starting with the writings of alchemists to the modern applications of reaction information in synthesis. Because of the very large amount of information available, the chosen examples are only illustrative for a certain period and do not attempt comprehensive coverage. Because of the type of information available, the earlier parts of this chapter, particularly those involving alchemy, are not necessarily restricted to information about reactions, but the latter part exclusively uses examples from organic chemistry because of the tremendous growth of information about chemical reactions and its varied uses in all areas of chemistry.

The Past

Alchemy

Alchemy, whose early practitioners claimed profound powers, was practiced in Egypt as early as 2000 BCE and included Hermetic principles and practices related to mythology, magic and spirituality. Their objectives were varied but included the creation of the philosopher's stone, the capability of turning base metals into gold or silver and the creation of an elixir of life (1). A mysterious Egyptian called "Hermes Trismegistus" by the Greeks is generally considered the founder of alchemy. Unfortunately, his and many of the writings from antiquity were destroyed by emperor Diocletian who ordered the burning of alchemical books (2) to suppress a revolution in Alexandria in 292 CE. One of the few surviving documents was the *Emerald Tablet (Tabula Smaragdina)* (3), considered the primary document of alchemy. Although Hermes Trismegistus is the author named in the text, it first appeared as a book written in Arabic between the sixth and eighth century. Other noteworthy texts that survived included the *Papyrus Graecus Holmensis* or *Stockholm Papyrus* (ca. 300 CE) (4) which contains 154 recipes for dyeing, gemstone coloring, pearl cleaning and generating gold and silver imitations and the *Leiden Papyrus X* (5), a Greek papyrus codex containing alchemical texts, mostly about making dyes and alloys looking like gold. It is assumed that both of these texts were written in the 2nd or 3rd century by the same scribe. The core of these recipes and additional texts are found in the Medieval Latin text *MappaeClavicula* (6). Among the oldest known books on alchemy, of which quotations in Greek and Arabic are known, are the ones written by Zosimos of Panapolis (7), an Egyptian or Greek alchemist and mystic who lived between the end of the 3rd and the beginning of the 4th century. In summary, alchemists developed a framework of theory, terminology, experimental process and basic laboratory techniques that are still recognizable today.

The Beginnings of Modern Chemistry

A new approach to alchemy that relied more on scientific methodology and controlled experimentation in the laboratory instead of the most allegorically writings of the work of earlier alchemists was first introduced in the late 8th century by Jābir ibn Hayyān (721 – 815 CE) (8). Known in Europe as the "Geber", his processes and apparatus were clearly described and he used a methodical classification of substances. In his writings, Jābir proclaimed the importance of experimentation as follows: "*The first essential in chemistry is that thou shouldest perform practical work and conduct experiments, for he who performs not practical work nor makes experiments will never attain to the least degree of mastery*" (1). For these reasons, Jābir is considered by many the father of chemistry. In addition to Jābir, other Islamic chemists, for example Al-Kindi (801 – 873 CE) and Muhammad ibn Zakarīya Rāzi (865 – 925 CE), contributed key chemical discoveries, for example the synthesis of hydrochloric, sulfuric and nitric acids and the power of *aqua regia* to dissolve gold. As a philosopher, Jābir contributed greatly to alchemical hermeticism seeing as his ultimate goal the artificial creation of life in the laboratory. Many other influential alchemists

followed Jābir to practice and publish their work until the Middle ages. One of them is the “Pseudo-Geber” or “Paul of Taranto” (9), an anonymous European alchemist who published his *Summa Perfectionis* in the 13th century, clearly describing alchemical theory and practice.

Despite the many contributions of Jābir to chemistry, others reserve the title “father of chemistry” for Robert Boyle (1627 – 1631) (10), best known for Boyle’s Law. Though his research was rooted in alchemy, he employed modern experimental scientific methods in wide-ranging areas such as natural philosophy, inventions, physics, and chemistry. His book *The Skeptical Chymist*, written in 1661, is an essential book in chemistry that started the evolvement of alchemy into modern chemistry. He studied the composition of substances and differentiated between mixtures and compounds and described techniques to detect ingredients (analysis). He published the results of his studies, even negative ones, with detailed information about procedures, apparatus and observations.

A century later, Torbern Olaf Bergman (1735 – 1784) (11), a Swedish chemist and mineralogist, was the first one to use diagrams and symbols to explain chemical reactions. His most important contribution was an essay on electric attractions, in which he lists the elements in order of their affinity; at that time the largest listing of that type. He wrote a treatise on the manufacture of “alum”, double sulfate salts *e.g.* $\text{KAl}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$. He was the first to use the term “organic chemistry” and is considered one of the founders of analytical chemistry.

Another giant in the development of modern chemistry was Antoine Lavoisier (1743 – 1794) (12) who in 1785 disproved the phlogiston theory by correctly stating that combustion is a reaction with oxygen. Among other important achievements, he was instrumental in reforming literature and generating the first extensive list of elements.

Based on work on gases by Joseph Louis Gay-Lussac (1778 – 1850) and on the atomic theory of John Dalton (1766 – 1844), Joseph Proust (1754 – 1826) developed the law of definite proportions, the forerunner of the concepts of stoichiometry and chemical equations.

For a long time, it was believed that compounds obtained from living organisms were too complex to be synthesized. The concept of vitalism (13) stated that organic matter contained a vital force, distinguishable from inorganic materials. However, the synthesis of urea from inorganic precursors by Friedrich Woehler in 1828 (14) and the synthesis of acetic acid from carbon sulfite by Herman Kolbe in 1844 (15, 16) dealt the final blow to the vitalism theory.

Abstracting Services, Handbooks and Journals

Until the 18th century, almost all texts about or descriptions of chemical reactions were either written in Latin or translated into Latin from Greek or Arabic. In the 19th century, the situation changed with the advancement of modern chemistry, particularly in Europe, resulting in the publication of journals from different countries in various languages. One of the oldest and historically most important journal covering organic chemistry worldwide is the renowned *Justus Liebigs Annalen der Chemie* (*Liebigs Annalen*) which was founded in 1832 as *Annalen der Pharmacie*. After several title changes, it merged in 1998 with

Recueil des Travaux Chimiques des Pays-Bas (established in 1882), *Bulletin de la Société Chimique de Paris* (first published in 1858) and *Chemische Berichte* to form the *European Journal of Organic Chemistry*. Other important journals followed. The year 1849 saw the publication of the *Quarterly Journal of the Chemical Society*, which underwent many splits, mergers and name changes. The resulting journals are published by the Royal Society of Chemistry. *Berichte der Deutschen Chemischen Gesellschaft* was first published in 1868, it changed its name in 1947 to *Chemische Berichte* and later merged with *Liebigs Annalen*. The *Zeitschrift für die Chemische Industrie* was first published in 1887 by Ferdinand Fischer and, after several name changes, became *Angewandte Chemie* in 1947. The prestigious *Journal of the American Chemical Society* started its life in 1879.

With the increasing amount of data available in different languages, it was only natural that abstract services had to be established to provide researchers with the information they required. The first abstracting service for chemistry was founded in 1830 in Germany by Gustav Theodor Fechtner and published by Leopold Voss under the name *Pharmazeutisches Centralblatt*. In the first year, 400 abstracts were published covering important research in pharmaceutical chemistry including descriptions of the transformations of compounds (reactions). In 1850, the title changed to *Chemisch-Pharmazeutisches Zentralblatt*. In 1856 it became the *Chemisches Zentralblatt* (17, 18) to highlight the importance of chemistry. Because of the high costs of collecting and abstracting the primary chemical literature worldwide, the publication ceased in 1969. Over 140 years, *ca.* 2 million abstracts were produced. In order to conserve this important information, the complete work was digitized in 2011 and developed as a full text searchable database by FIZ CHEMIE Berlin (19). Subsequently, in 2012 InfoChem in cooperation with FIZ CHEMIE Berlin developed a (sub)structure searchable, web-based database, the *Chemisches Zentralblatt Structure Database* (20).

In the years 1881 – 1883, the first edition of the Beilstein Handbook, an encyclopedia of organic compounds, was published by Friedrich Konrad Beilstein (21). From the beginning, Beilstein's approach to indexing and categorizing was based on structural diagrams alone, named "The Beilstein System"; certainly a revolutionary concept at that time. His intentions were to make the handbook as comprehensive and critical as possible by listing all known organic chemical species and all their known validated data, including preparations. The printed form of the Beilstein Handbook was discontinued in 1998 with the sixth supplementary series of the fourth edition, which originally was published in 1918. Over the years the ownership of the handbook changed hands several times but the concept stayed the same. The content is now part of Elsevier's Reaxys system (22).

Chemical Abstracts (CA) (23) made its debut in 1907 when William A. Noyes enlarged the *Review of American Chemical Research*, an abstracting service aimed mainly at US researchers that he started in 1895. Over the years, the one-man enterprise grew into the world's principal abstracting and indexing service relying initially on volunteer abstractors. However, because of the ever-increasing amount of chemistry-related worldwide literature to be covered, the use of volunteer abstractors was phased out in 1994. In 1956, CA became

Chemical Abstracts Service (CAS), founded as a division of the American Chemical Society. Publication of the printed version of CA ceased in 2010.

The 20th century saw a host of printed sources specifically designed for information about chemical reactions. Some of these are listed below and described in detail by Engelbert Zass (24). They include the 140 volumes (160,000 pages) of *Houben-Weyl Methoden der Organischen Chemie* (1909–2002, Karger), *Organic Synthesis* (1921–present, Wiley), *Theilheimer's Synthetic Methods of Organic Chemistry* (1946–present, Karger), *Fieser & Fieser's Reagents for Organic Synthesis* (1967–present, Wiley), *ChemInform* (1970–present, FIZ CHEMIE Berlin, Wiley VCH), *Current Chemical Reactions* (1979–present, Institute for Scientific Information, ThomsonReuters), *Comprehensive Heterocyclic Chemistry* (Pergamon Press, 1984), *Encyclopedia of Reagents for Organic Synthesis* (1995, Wiley), and *Strategic Applications of Named Reactions in Organic Synthesis* (2005, Elsevier). Additionally, a large number of books on reagents, named reactions and other transformations is available. All of the listed sources became available as electronic databases (*vide infra*).

The Present

Structure Representation

With the advancement of computer technologies it became feasible in the middle of the 20th century to manage the rapidly increasing amount of data electronically, a subject which later became known as chemoinformatics or cheminformatics (25). This required the development of machine-readable structure representations (26). Several approaches were considered but only linear notations and connection tables attracted attention. The former included the *Dyson Notation*, which was supplanted by the *Wiswesser Line Notation* (WLN) (27), the *SMILES Notation* (28) and very recently the *International Chemical Identifier* for molecules (InChI) (29) and the *International Chemical Identifier for Reactions* (RInChI) (30). While notation systems implicitly encode the topology of a molecule, connection tables encode this information explicitly. Therefore, connection tables rapidly supplanted notation systems. The use of connection tables was first reported in 1957 by Ray and Kirsch (31). Subsequently, designs for faster and more efficient formats were reported over the years (25).

Inherently, structural searches of individual molecules are simpler than the corresponding searches of individual reactions. Not only have reactants and products to be considered but also the reaction sites. This was first recognized by Vladutz (32) in 1963, followed in 1967 by comparison work by Lynch (33). It took nearly 15 years until an efficient and effective method for the detection of reaction sites based on maximum common subgraph isomorphism was reported by Willett (34). His findings served as the basis for the generation of reaction databases and allowed structural searches of the contents. Reports of the first operational systems started to appear in the eighties (35–38) and are described in conference proceedings (39) that also include references to indexing systems. Among the first graphic-oriented, server-client based in-house search systems, **REACCS** (Reaction **ACC**ess System) from Molecular Design Ltd was introduced

in 1983 and later morphed into the Reaction Browser (40) under the client software ISIS/Base. A user-friendly, form-based querying and data-displaying system allowed for accessing a large number (ca. 1 million) of selected reactions from several databases, either individually or in any combination. A hierarchical data structure was the key for efficient structure or data searches.

The availability of structural searches set in motion the development of numerous structure editors for in-house and web-based applications starting in the later parts of the 20th century (41). Besides characterizing a reaction by using structural changes, electronic terms can also be used to describe a reaction. For example, Ugi and his group in Munich carried out long-term studies of matrix representation (Dugundji – Ugi model) for processing chemical synthesis information (42–44). This concept later became the basis of the EROS (Elaboration of Reactions for Organic Synthesis) (45) and WODCA (Workbench for the Organization of Data for Chemical Applications) (46) systems (*vide infra*).

Electronic Publications

Nearly all journals published today are now electronically available online and many of them supply the structural diagrams as connection tables in the supplementary material associated with a given publication. Many of them link primary information (journal) to secondary information (database) to create a fully integrated environment made possibly by the rapidly growing importance of the Internet. The Dymond link (47) in Elsevier publications is one example of connecting primary data with metadata. Another example of linking technology was developed by CAS. ChemPort links users of SciFinder or STN to articles from more than 7,000 electronic journals and a large number of electronic patent documents (48).

Among the handbooks converted into structurally searchable databases, Chemical Abstracts Service developed in the late 70s a text and structure searchable delivery system, called Messenger, which over the years morphed into CAS Online, STN, CASREACT and SciFinder. CASREACT was introduced in 1988 as a document-based, structure-searchable database containing ca. 68 million reactions of which 55 million are single-step reactions. Journals, patents and reference works are covered from 1840 – present. It can be searched on STN and SciFinder. A web-based version of SciFinder was released in 2008. More information is available on the CASREACT Database Summary Sheet (49).

In 1985, the Beilstein Institute initiated the development of a text- and structure-searchable database of the Beilstein Handbook keeping the traditional, hierarchical classification of ca. 400 data fields. The database was first accessible in 1989 on STN and Dialog hosts. However, it turned out that the capacities of the Beilstein database were not fully exploited. This led to the development of CrossFire (21) and its client Commander during the years 1993 – 1995. It incorporated chemical structures and reactions in their usual graphic format and text abstracts, such as abstracts, titles and concepts. Basically, the Crossfire solution treated the Beilstein file as a unit of three databases (substances, reactions, documents). On the CrossFire Web client, the embedded details from a particular search were hyperlinked to a new window containing the object and

further hyperlinks. This, for example, led to the generation of retro-synthetic pathways. Reaxys (22), the successor to CrossFire, was launched by Elsevier in 2009 as a workflow solution for research chemists. It provides links to Scopus (50) and ScienceDirect (51).

Reaction Databases

The first structurally searchable databases based on reactions abstracted from the literature, unlike the document-based *CASREACT* and *CrossFire* databases, started being available in the eighties. These included the following databases from several organizations: *Theilheimer* (Karger), *ChemReact* (InfoChem), *ChemInform* (FIZ CHEMIE Berlin), *eEros* (Wiley), *Science of Synthesis* (Thieme), *Current Chemical Reactions* (ThompsonReuters), *Methods in Organic Synthesis* (Royal Society of Chemistry) and many other, more specialized databases, e.g. *Comprehensive Heterocyclic Chemistry* (Elsevier), *Protecting Groups* (Accelrys), *Solid-Phase Organic Reactions* (FIZ CHEMIE Berlin), etc. (52). Most of these databases can now be accessed on the vendors' website or as in-house databases using vendor-developed front-ends. They serve both occasional or novice and expert users very well. The interfaces take into account users' tasks and capabilities by simplifying the querying process (natural and not rule dependent), providing tools for post-search management of search results (clustering) and facilitating the indexing of data (classification). Some web-based systems allow for the seamless integration of various information sources.

Computer-Aided Synthesis Design

Planning the synthesis or preparation of a compound, small or complex, has been a priority for chemists from alchemists to modern synthetic chemists. They can devise plans that move forwards from readily available starting materials or backwards (retrosynthetic) by recognizing structural features in the desired target that can easily be assembled based on the chemist's knowledge of the literature. These options are the same in the computer age. The first one to suggest the use of computers for designing syntheses of organic compounds in 1963 was Vladutz (31) in his paper on classification and codification of organic reactions. It took another decade until his vision was realized by Corey and Wipke (53) in his seminal paper on computer-aided synthesis (LHASA – Logic and Heuristics Applied to Synthesis Analysis). It should be noted here that this program preceded reaction retrieval systems by 15 years. During the 50 years after the publication of the LHASA paper, many programs were developed to computerize multistep syntheses. Basically, they can be categorized into three main groups: empirical, numerical and formal. Because of the large number of programs in each group, only one or two representatives will be mentioned. Interested readers can refer to several reviews that have been written about the topic (42, 54, 55).

Empirical Approach

LHASA and other members of this group are typical expert systems with a knowledge base and a set of rules. They use diverse reaction libraries to retro synthetically generate a synthesis tree perceived on structural features by applying transforms, rules and schemes. Depending on the target molecule, these trees can grow exponentially if user interaction is not possible. Because of limited user interaction, most synthetic chemists are turned away. A good example is the CASP program (**C**omputer **A**ssisted **S**ynthesis **P**rogram) (56), which was developed by European companies based on the SECS system (**S**imulation and **E**valuation of **C**hemical **S**ynthesis) (57–59). Though several thousand transforms were generated at large expense, chemists never accepted the program. SECS is a successor to the original LHASA program. A very recent example of a rule-based system is the ARChem program developed over the years by Peter Johnson and coworkers (60). A knowledge-based system that works in the forward direction, *i.e.* it predicts a product from a given starting material, is CAMEO (**C**omputer **A**ssisted **M**echanistic **E**valuation of **O**rganic **R**eactions) developed by the Jorgensen group (61). The program which consists of several modules uses mechanistic principles for describing classes of reactions, *e.g.* pericyclic reactions, electrophilic aromatic substitution *etc.*

Numerical Approach

This approach is best characterized by the work of Hendrickson who in 1971 published a paper on “Characterization of Structures and Reactions for Use in Organic Synthesis” (62). This conceptual paper was the first in a long series of papers over the next forty years by Hendrickson resulting in the development of the SYNGEN (**S**Ynthesis **G**enerator) program (63) and the generation of systematic signatures for organic reactions (64). The synthetic strategy in the SYNGEN program is based on convergence and half-reactions. The target molecule is divided into two pieces and each one again into two or more until identical carbon skeletons are found in a catalog of starting materials described by maximum binary numbers obtained from adjacency matrices. A reaction is described by its net structural change at each skeletal carbon atom considering the skeleton and functionality changes (construction and refunctionalization reactions). Construction reactions are a combination of two half-reactions. The exchange of attachments at a skeletal carbon (functionality change) is defined as a unit reaction and is described by a two-letter symbol.

Formal Approach

Research in this area is best described as constitutional chemistry, an approach that is purely algebraic and logic-centered based on BE- and R-matrices (Dugundji-Ugi model) (41). In 1978, Gasteiger and Jochum described the first version of EROS (**E**laboration of **R**eactions for **O**rganic **S**ynthesis) (45) based

on this model. Over the years, particularly in the 90s, many variations using the EROS program as a foundation were developed, including WODCA (**W**orkbench for the **O**rganization of **D**ata for **C**hemical **A**pplications) (46) and RAIN (**R**eaction **A**nd **I**ntermediate **N**etworks) (65). The selection processes of WODCA for synthetic routes are heuristic in nature and allow the quantification of fundamental electronic and energy effects, such as charge distribution, inductive, resonance and polarizability effects, heats of formation and bond dissociation energies. RAIN operates with strictly formal reaction generators. Using the R-matrix of the Dujundji-Ugi model, a graph theoretical representation of chemical reactions was used. Reactants and products are represented by their connectivity matrix and the reaction by subtracting the two matrices mathematically. The algorithm has produced several new verified reactions.

Computer-Aided Management of Reactions

During the last decade, reaction databases, especially CASREACT, Reaxys and ChemReact, accumulated an enormous amount of data that required innovative ways to manipulate them. Fortunately, computer technology and the Internet grew at an even faster pace allowing for the development of more user-friendly, highly interactive programs. Reaction classification played an important role in these developments. This technique, complementary to structure-based retrieval systems, allows for post-search management of large hit lists, encourages browsing, simplifies query generation and provides access to generic types of information in retrieval systems. It is a useful tool for linking reaction information from different sources; it is a source for deriving knowledge bases for reaction prediction and synthesis design and can be applied in the prediction of new reactions. It is useful in setting up automatic procedures for analyses and correlations, in quality control and overlap studies. Describing the many applications in detail would be outside the cope of this chapter. The reader is encouraged to look up some recent publications that describe in detail algorithms for automatic atom-atom mapping, reaction center detection and reaction classification (66–69).

Theses papers contain many references to the methodologies being used. In general, two categories of methodologies have been applied in reaction classification: model- driven and data-driven. In the former, a preconceived model is imposed and in the latter a computer automatically generates a classification by analyzing a set of reactions. In both approaches reaction center detection plays a pivotal role. While model-driven approaches in general only consider the reaction center, data-driven models take into account the surrounding functionality. A good example is the program CLASSIFY (68) developed by InfoChem that is being used in many databases. It uses class codes (fixed numeric strings) to assist chemists in searching reaction databases by allowing fuzzy searches and managing large hit lists through clustering. The program is very useful for checking the diversity and quality of reaction databases (70). Other authors, particularly in the research groups of Gasteiger and Funatsu in Germany and Japan, respectively, added physicochemical descriptors to produce reaction hierarchies. An example

is the HORACE (Hierarchical Organization of Reactions through Attribute and Condition Eduction) system (71) developed by Röse and Gasteiger.

The difficulty of measuring the relationship between two reactions in different subclasses was addressed by Chen and Gasteiger (72) by using Kohonen neural networks (73) and producing two-dimensional classification schemes. The system can also be used for the analysis of databases. Similar work was also carried out by the Funatsu group (74). Though reaction classification based solely on topological features is very much established, improvements in stereochemistry and the use of reliable physicochemical parameters are still needed. Also needed are methods for improved integration with reaction databases and better knowledge acquisition to improve data management.

The Future

It is difficult to predict the future of reaction information. It is obvious that the rapid development of computer technologies will have a major impact on new developments. In order to make the vast amount of data available to researchers in different places, reasonable open access to programs, databases, journals and other sources is essential. Unique identifiers, such as RInChIs, will be used more and more by publishers and database providers to link reactions from different sources. Electronic notebooks (ELN) will become more sophisticated important knowledge sources. For example, a recent publication (75) describes the extraction of reaction sequences from ELNs. Based on the extracted knowledge, a retro-synthesis tool is built in that allows the *de novo* design of compounds very likely to be synthetically feasible. New developments, extensions and improvements of mobile devices will increase the sophistication and usefulness of applications dealing with reaction information.

Reaction data management systems will become more efficient by creating an environment that allows for combining the intelligence and creativity of synthetic chemists with the processing and simulating power of computers and the wealth of information in databases. Despite the little use of computer-aided organic synthesis design programs at present, there is hope that in the future these systems will be more acceptable to synthetic chemists. In order to gain acceptance, systems must take into account the knowledge and intuition of chemists in designing synthetic routes and reaction conditions. At every stage of the program the systems have to be highly interactive and mimic and support the typical planning style of chemists, which generally is not in a straight-forward line.

The future of efficient management and use of reaction information at all levels looks bright as long as available electronic resources serve the synthetic chemist and not *vice versa*.

References

1. Alchemy. <http://en.wikipedia.org/wiki/alchemy> (accessed September 25, 2013).

2. Partington, J. R. *A Short History of Chemistry*; Dover Publications: New York, NY, 1989.
3. Goodrick-Clarke, N. *The Western Esoteric Traditions: A Historical Introduction*; Oxford University Press: Oxford, U.K., 2008.
4. Caley, E. R. *J. Chem. Educ.* **1927**, 4, 979–1002.
5. Caley, E. R. *J. Chem. Educ.* **1926**, 3, 1149–1166.
6. Phillipps, T. *Journal Archaeologia* **1847**, 33, 183–244.
7. Zosimos of Panopolis (Egyptian alchemist). http://en.wikipedia.org/wiki/Zosimos_of_Panapolis_note-2 (accessed September 25, 2013).
8. Abū Mūsa Jābir ibn Hayān (721 – 815). http://en.wikipedia.org/wiki/Jabir_ibn_Hayyan (accessed September 25, 2013).
9. Pseudo-Geber. <http://en.wikipedia.org/wiki/Pseudo-Geber> (accessed October 29, 2013).
10. Robert Boyle. http://en.wikipedia.org/wiki/Robert_Boyle (accessed September 25, 2013).
11. Möstrom, B. *Tobern Bergman: a bibliography of his works*; Almquist & Wiksell: Stockholm, Sweden, 1957.
12. Antoine Lavoisier. http://en.wikipedia.org/wiki/Antoine_Lavoisier (accessed September 25, 2013).
13. Vitalism doctrine. <http://en.wikipedia.org/wiki/Vitalism> (accessed October 29, 2013).
14. Woehler, F. *Ann. Phys. Chem.* **1828**, 88 (2), 253–256.
15. Acetic acid. http://en.wikipedia.org/wiki/Acetic_acid#History (accessed October 25, 2013).
16. Hermann Kolbe. http://en.wikipedia.org/wiki/Hermann_Kolbe (accessed October 25, 2013).
17. Weiske, C. *Chem. Ber.* **1973**, 106, I–XVI.
18. Pflücke, M. *Angew. Chem.* **1954**, 66, 537–541.
19. FIZ CHEMIE Berlin (now Wiley-VCH). <http://www.fiz-chemie.de>.
20. InfoChem GmbH, Munich. <http://www.infochem.de/products/databases/czb.shtml>.
21. Lawson, A. The Beilstein Database. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH Verlag: Weinheim, Germany, 2003; pp 608–628.
22. Reaxys: *Chemistry Workflow Solution*; Elsevier Information Systems: Frankfurt, Germany. <http://www.elsevier.com/online-tools/reaxys> (accessed October 9, 2013); and the chapter by Swienty-Busch et al. in this volume.
23. Fisanik, W.; Shively, E. R. The CAS Information System: Applying Scientific Knowledge and Technology for Better Information. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH Verlag: Weinheim, Germany, 2003; pp 556–607.
24. Zass, E. Reaction Databases. In *Encyclopedia of Computational Chemistry*; von Ragué Schleyer, P., Allinger, N. L., Clark, T.; Gasteiger, J.; Kollman, P. A.; Schaefer, H. F.; Shreiner, P. R., Eds.; Wiley: Chichester, U.K., 1998; pp 2402–2420.

25. Willett, P. A History of Chemoinformatics. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH Verlag: Weinheim, Germany, 2003; pp 6–20.
26. Warr, W. A. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1* (4), 557–579.
27. Smith, E. G.; Wiswesser, W. J.; Addelston, A. *The Wiswesser line-formula chemical notation*; McGraw-Hill: New York, NY, 1968.
28. Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
29. Heller, S. R.; McNaught, A. D.; Stein, S.; Tchekhovski, D.; Pletnev, I. *J. Cheminf.* **2013**, *5*, 7.
30. Grethe, G.; Goodman, J. M.; Allen, C. H. G. *J. Cheminf.* **2013**, *5*, 45.
31. Ray, L. C.; Kirsch, R. A. *Science* **1957**, *126*, 814–819.
32. Vladutz, G. E. *Inf. Stor. Retrieval.* **1963**, *1*, 117–146.
33. Armitage, J. E.; Lynch, M. F. *J. Chem. Soc. (C)* **1967**, 521–528.
34. McGregor, J. J.; Willett, P. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 137–140.
35. Johnson, A. P. *Chem. Br.* **1985**, *21* (1), 59–67.
36. Wipke, T. Exploring Reactions with REACCS. In *Abstracts of Papers, 188th ACS National Meeting & Exposition*, Philadelphia, PA, August 26–31, 1984; American Chemical Society: Washington, DC.
37. Wipke, W. T.; Hounshell, D.; Mook, T. E.; Grier, J. In *Modern Approaches to Chemical Reaction Searching*; Willett, P., Ed.; Gower: Aldershot, U.K., 1986; p 586.
38. Blower, P. E. Design Considerations for a Chemical Reaction Search Service. In *Abstracts of Papers, 188th ACS National Meeting & Exposition*, Philadelphia, PA, August 26–31, 1984; American Chemical Society: Washington, DC.
39. Willett, P., Ed.; *Modern Approaches to Chemical Reaction Searching*; Gower: Aldershot, U.K., 1986.
40. Tseng, S.-S. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1138–1145.
41. Molecule editor. http://en.wikipedia.org/wiki/Molecule_editor (accessed October 9, 2013).
42. Dugundji, J.; Ugi, I. *Top. Curr. Chem.* **1973**, *39*, 19–64.
43. Ugi, I.; Bauer, J.; Bley, K.; Dengler, A.; Dietz, A.; Fontain, E.; Gruber, B.; Herges, R.; Knauer, M.; Reitsam, K.; Stein, N. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 201–227.
44. Brandt, J.; Bauer, J.; Frank, R. M.; von Scholley, A. *Chem. Scr.* **1981**, *18*, 53–60.
45. Gasteiger, J.; Jochum, C. *Top. Curr. Chem.* **1978**, *74*, 93.
46. Gasteiger, J.; Ihlenfeldt, W. D.; Röse, P. *Recl. Trav. Chim. Pays-Bas* **1992**, *111*, 270–290.
47. Lawson, A.; Leonhard, C. Dymond linking: point-and click structure and reaction searching. In *Abstracts of Papers, 221st ACS National Meeting & Exposition*, San Diego, CA, April 1–5, 2001; American Chemical Society: Washington, DC, 2001.
48. Chemical Abstracts Service. *ChemPort*. <http://www.cas.org/fulltext/cas-full-text-options> (accessed March 3, 2014).

49. *CASREACT Database Summary Sheet*; Chemical Abstracts Service: Columbus, OH. <http://www.cas.org/content/reactions> (accessed October 8, 2013).
50. *Scopus*. <http://www.elsevier.com/online-tools/scopus> (accessed March 9, 2014).
51. *ScienceDirect*. <http://www.sciencedirect.com> (accessed March 9, 2014).
52. Zass, E. Databases of Chemical Reactions. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH Verlag: Weinheim, Germany, 2003; pp 667–699.
53. Corey, E. J.; Wipke, W. T. *Science* **1967**, *166*, 178.
54. Barone, R.; Chanon, M. Synthesis Design. In *The Encyclopedia of Computational Chemistry*; von Ragué Schleyer, P., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., Shreiner, P. R., Eds.; Wiley: Chichester, U.K., 1998; pp 2931–2948.
55. Barone, R.; Chanon, M. Computer-Assisted Synthesis Design. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH Verlag: Weinheim, Germany, 2003; pp 1428–1456.
56. Brown, H. W. *Chem. Ind. (Duesseldorf)* **1988**, *40*, 43–44, 48, 50, 53.
57. Wipke, W. T.; Rogers, D. J. *Chem. Inf. Comput. Sci.* **1984**, *24*, 71–81.
58. Wipke, W. T.; Braun, H.; Smith, G.; Choplin, F.; Sieber, W. SECS-Simulation and Evaluation of Chemical Synthesis: Strategy and Planning. In *Computer-Assisted Organic Synthesis*; ACS Symposium Series 61; Wipke, W. T.; Howe, W. J., Eds.; American Chemical Society: Washington, DC, 1977; pp 97–127.
59. Wipke, W. T.; Ouchi, G. I.; Krishnan, S. *Artif. Intell.* **1978**, *11*, 173–193.
60. Law, J.; Zsoldos, Z.; Simon, A.; Reid, D.; Liu, Y.; Khew, S. Y.; Johnson, A. P.; Major, S.; Wade, R. A.; Ando, H. Y. *J. Chem. Inf. Model.* **2009**, *49*, 593–602.
61. Fleischer, J. M.; Gushurst, A. J.; Jorgensen, W. L. *J. Org. Chem.* **1995**, *60*, 490–498.
62. Hendrickson, J. B. *J. Am. Chem. Soc.* **1971**, *93*, 6847.
63. Hendrickson, J. B. *Angew. Chem., Int. Ed. Engl.* **1990**, *29*, 1286–1296.
64. Hendrickson, J. B. *J. Chem. Inf. Model.* **2010**, *50*, 1319–1329.
65. Herges, R. *J. Am. Chem. Soc.* **1990**, *30*, 377–383.
66. Chen, L. Reaction Classification and Knowledge Acquisition. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH Verlag: Weinheim, Germany, 2003; pp 348–388.
67. Grethe, G. Analysis of Reaction Information. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH Verlag: Weinheim, Germany, 2003; pp 1407–1427.
68. Kraut, H.; Eiblmaier, J.; Grethe, G.; Löw, P.; Matuszczyk, H.; Saller, H. *J. Chem. Inf. Model.* **2013**, *53*, 2884–2895.
69. Chen, W. L.; Chen, D. Z.; Taylor, K. T. Automatic reaction mapping and reaction center detection. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2013**, *3*, 560–593.
70. Eigner-Pitto, V.; Kraut, H.; Saller, H.; Matuszczyk, H.; Löw, P.; Grethe, G. Reaction classification, an enduring success story. In *241st ACS National*

Meeting & Exposition, Anaheim, CA, United States, March 27–31, 2011; American Chemical Society: Washington, DC.

71. Röse, J. R.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 74–90.
72. Chen, L. Reaction Classification and Knowledge Acquisition. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH Verlag: Weinheim, Germany, 2003; pp 348–388.
73. Zupan, J.; Gasteiger, J. Neural Networks. In *Neural Networks for Chemists. An Introduction*; Zupan, J., Gasteiger, J., Eds.; VCH Verlagsgesellschaft mbH: Weinheim, Germany, 1993; pp 79–95.
74. Satoh, H.; Sacher, O.; Nakata, T.; Chen, L.; Gasteiger, J.; Funatsu, K. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 210–219.
75. Christ, C. D.; Zentgraf, M.; Kriegl, J. M. *J. Chem. Inf. Model.* **2012**, *52*, 1745–1756.

Chapter 7

The Institute for Scientific Information: A Brief History

Bonnie Lawlor*

**Retiring Executive Director, National Federation of Advanced Information
Services (NFAIS), 276 Upper Gulph Road, Radnor, Pennsylvania 19087**

***E-mail: chescot@aol.com**

The Information Industry has consolidated over the past thirty to forty years through a series of mergers and acquisitions. Today, the dominant commercial players in scientific publishing are the big players such as the Elseviers and Thomson Reuters of the world and the major non-profits are scientific societies such as the American Chemical Society. But during the second half of the last century there were far more creative entrepreneurial players who were developing what would become essential information services. Among others, these included BIOSIS, Derwent, Dialog, the Institute for Scientific Information, Molecular Design, Ltd., Engineering Information, etc. These no longer exist as stand-alone organizations although their products continue under the umbrella of Elsevier, Thomson Reuters, and ProQuest.

This paper will take a look at one of these icons of the past - the Institute for Scientific Information (ISI®) and the innovative chemist and entrepreneur, Eugene Garfield, who created it.

The Institute for Scientific Information (ISI®). A name well known to scholars, researchers, librarians, and information scientists of a certain age around the globe. But a name that has gradually lessened in prominence over the past twenty-two years since the company was acquired by the Thomson Corporation. Why was ISI established and by whom? How did it grow? What was it like to work there during its formative years until it was ultimately sold? And what is its legacy for generations of researchers yet to come?

This brief history of ISI will attempt to answer some of these questions from my perspective - a twenty-eight year ISI employee (1967 - 1995) who had the good fortune to begin and spend the majority of my career in a company that could be accurately described as the Google of its time. It was a wonderful place to learn about publishing and scholarly communication, and many of its “graduates” went on to become major players within the information community. So let’s take a look at ISI - the man behind it, the ideas that started it, and the successful evolution that made it a leading abstracting and indexing service in scholarly and scientific communication - culminating in its purchase in 1992.

The Man

The man responsible for the creation and development of what would ultimately become the Institute for Scientific Information is Eugene Garfield (1925 -). He was born in New York City and studied at the University of Colorado Boulder and the University of California Berkeley before obtaining a Bachelor of Science degree in chemistry from Columbia University in 1949. Garfield’s career began as a laboratory bench chemist and his employer was Professor Louis P. Hammett of Columbia University (father of the Hammett equation). By his own admission, Garfield was not successful in the lab, having had at least two explosive attempts to prepare some picric acid derivatives (1).

His career changed in 1951 purely by happenstance as many such changes do. He was attending the 75th Anniversary meeting of the American Chemical Society (ACS) during the World Chemical Conclave held in New York City and out of curiosity listened in on a session on documentation chaired by James W. Perry. Perry had been involved in looking at the use of punch cards to handle large volumes of chemical information. He chaired the ACS Board Committee on Punch Cards in 1946 and in 1948, along with G. Malcom Dyson, was instrumental in convincing IBM President, Thomas J. Watson, to work on the problem (2). The results of the IBM research was presented at the very session Garfield attended and he was hooked. A casual conversation with Perry ultimately led to a position for which Professor Hammett recommended him as a “hard, but not very original worker” (3). Garfield began as a research assistant at Johns Hopkins University on the Welch Medical Indexing Project under Stanford V. Larkey who needed a chemist who was an expert at subject headings. Along with his research work at Columbia under Hammett, Garfield had done some chemical indexing of their compounds. The Welch project, funded by the Army Medical Library (predecessor to the National Library of Medicine), had begun in 1948. It was one of the first large-scale investigations into the potential of machine-based information systems and had a major impact on all subsequent studies.

Garfield spent two years on the project and the experience would serve him well in the years to come – both in the creation of the three concepts that would serve as the foundation of ISI’s core products and services, and in the development of a network of contacts and friends whom he could call upon as needed throughout his career.

The Concepts

The first concept grew out of his work on improving the currency of the *Current List of Medical Literature*, a type-written contents-page service. Originally this service did not have indexes and was fairly current. When indexes were added, the labor involved in reading and indexing the articles slowed production considerably. While working on the Welch Project Garfield developed machine methods for compiling the *List* and applied the IBM 101 punch card sorter to search the database. Facilitated by his experience on this work and driven by a personal need to keep abreast of all published research related to the Welch Project, Garfield created *Contents in Advance*, an innovative contents-page current awareness tool for articles being published in information and library science. This was a private service created after hours by photographing, not type-writing, the contents pages. And it was a service that he would continue to produce after he left the Project.

The second concept resulted from the time he spent during the Welch Project as a volunteer abstractor for *Chemical Abstracts* and by his work at the Project on searching chemical files. Garfield became interested in learning more about the indexing of chemical compounds and how the timeliness of the process could be improved, for as will be noted later, the abstracting and indexing services that were available in the early 1950's were considerably behind the literature in currency. This interest would ultimately motivate him to create a series of chemical services under ISI's auspices.

The third, and possibly the most influential of the three concepts, grew out of the considerable expertise he was building in indexing methodology. After some time spent indexing articles he came to the realization that the facts stated in research articles were supported by references to prior published research. He began to perceive the bibliographies of scholarly articles as a series of indexing statements. This realization led to the development of the idea that article references could potentially serve as the basis of a new indexing methodology. That idea was confirmed after Garfield received a letter from William C. Adair, a former Vice President with the Frank Shepard Company. This was the company that published *Shepard's Citations*, a listing of all authorities who cited a particular case, statute, or other legal authority. Adair said that his company had considered creating a citator's index in a scientific field, but thought the idea to be impractical. Garfield, who had never heard of *Shepard's*, accessed the publication in a library and, in his own words, had a "Eureka" moment that he would continue to explore after leaving the Project (4).

Finally, when the funding of the Welch Project was coming up for renewal, Garfield wanted to promote the work that was being done and spur interest to keep it alive. With Larkey's support, an advisory committee chaired by Chauncey D. Leake was established and a symposium entitled "Machine Methods in Scientific Documentation" was developed. Three hundred people attended and Garfield gave most of the presentations, leaving positive impressions on attendees and firmly establishing a network of those interested in mechanizing scientific communication.

But the symposium did not save the Project nor did Garfield stay on to see its demise. Larkey disapproved of Garfield working after hours on *Contents in Advance* even though the service was created to keep those *working on the Welch Project up to date on related developments*. Garfield ignored Larkey and continued to produce the service. As a result, he was fired.

As Garfield himself has ironically noted (5), Larkey had taken Hammett's job recommendation to heart and did not expect to hire an innovative assistant. It was Garfield's originality (and stubbornness) that got him tossed from the project, but it was that very originality that allowed him to further develop the concepts of citation indexing, chemical indexing, and contents-page alerts, the three concepts that would eventually serve as the foundation of his company. But that company was still to come, and Garfield left the project to study for a Master's degree in Library Sciences, which he obtained in 1954 from Columbia University. Garfield and other members of the small group involved in the Welch Project would go on to dominate and lead many influential changes in the field of library and information science. But, before continuing our look at ISI, it is important to take time to understand the information environment in the middle of the last century that would drive change within the information industry and ultimately facilitate Garfield's conception and growth of ISI.

The Information Environment

Keeping abreast of the literature had been a challenge for scholars and scientists since the implementation of the printing press in 1440. Attempts to manage the literature began seriously in 1665 when Denys de Sallo, a member of the French parliament, published what is now recognized as the first scholarly journal of the Western World, the *Journal des Scavans*. But the content was not original research, it was actually a form of abstracts. The journal's primary purpose was to catalogue and provide a brief description of the principal books then being printed in Europe, as well as to provide readable and critical accounts of current scholarly writings (6). But by the early 1800's approximately three hundred scientific journals had emerged and scholars were even more concerned than in the past with regard to the volume of information being recorded, stating that they were able to read less than half of the literature related to their research. They had tried to solve the problem by creating discipline-specific journals, hoping to dissect the overall literature into manageable segments, but it did not work. The number of articles continued to grow across all scholarly disciplines, but mostly in the sciences, and it was around this time, continuing well into the next century, that Abstracting & Indexing (A&I) services began to emerge as a means of speeding information discovery. Some examples are:

1817: *Handbuch der Anorganischen Chemie* (Gmelin)

1820: *Pharmacopeia of the United States*

1830: *Pharmaceutisches Centralblatt*

1867: *Catalogue of Papers* (Royal Society of London)

1873: *Shephard's Citations*

1879: *Index Medicus*
1881: *Handbuch der Organischen Chemie* (Beilstein)
1884: *Index Notes* (precursor to Engineering Index)
1889: *Merck Index*
1898: *Science Abstracts* (precursor to Inspec)
1898: *The Cumulative Book Index* (H.W. Wilson)
1907: *Chemical Abstracts*
1926: *Biological Abstracts*

But even the A&I services could not keep up with the growth in the literature, especially in the decades following the end of World War II. For example, by 1960 the annual processing of journal article abstracts at CAS had grown from 8,000 to 104,484, an increase of 1,307% since the process began in 1907 (7). In that same year, *Science Abstracts* processed 21,000 abstracts - a 1,376% increase since that product was launched in 1898 (8). In 1958 Dale Baker became Director of CAS. In his oral history compiled by the Chemical Heritage Foundation, Dale describes the problems that CAS faced during this time period - and I quote: "Our editing and indexing was very slow and we were running late. We were years behind on our indexes. We had a difficult job for four years to get caught up with our indexing and our regular issues. It was a very critical time" (9).

The journal growth rate had jumped from the average 3.3% noted from 1900 - 1944 to 4.68% from 1944 - 1978 (10), and it was believed that machines, specifically the relatively-new computer technology, could be the solution to managing this growth of scholarly and scientific information. Hence the critical importance of initiatives such as the Welch Project and IBM's work on the use of punch cards to handle large volumes of chemical information.

ISI – the Conceptual Years (1954 – 1960)

The time period from 1950 - 1990 would witness significant innovation in the use of machines and ultimately computers to process, manipulate, manage, and deliver information. It was at the start of this era, in the mid-1950's, that the evolution of ISI began in earnest and the concepts that Eugene Garfield developed at Johns Hopkins began to take root. In 1954, armed with the experience and knowledge gained on the Welch Project, his chemistry education, and his new degree in Library Sciences, Garfield began the next phase of his career as a one-man freelance operation. A six-month consulting assignment from Smith, Kline & French brought him to Philadelphia, PA where he worked on machine-based indexes to the pharmaceutical literature (11). His base of operation for several years would be his home, a log cabin located in Thorofare, a rural community outside of Woodbury, New Jersey. By that time he had sold *Contents in Advance* to a library school classmate, Ann McCann, who founded the now defunct publishing firm, Prometheus Press (12). But his interest in this format never wavered and in 1955 he began producing a contents-page service entitled *Management's Documentation Preview* (MDP) (13) based upon journals

covering management behavioral science, and converted a chicken coop on his property to serve as a shop where he did his own printing.

In 1955, Garfield also published his seminal paper, *Citation Indexes for Science*, in the journal *Science* (14), an article that would capture the attention of Professor Joshua Lederberg and other interested scientists.

By 1956 he incorporated under the name Documation, Inc. and around this time Bell Telephone Laboratories heard about *MDP*. They gave Garfield a contract for 500 copies to be printed under the name *Survey of Current Management Literature*. This was a critical moment for Garfield and a story that he has told (and printed) many times. The chicken coop could not handle Bell Labs' print volume and he was forced to turn to the services of a commercial printer who demanded payment of \$500.00 in advance. Even though Garfield was still doing consulting for Smith, Kline & French he did not have the money that was needed. No bank would loan him the money and he was advised to borrow from a personal finance company. The printer reduced the advance to \$300.00 and Garfield went to Household Finance Company (HFC) and quickly was given a check for the original \$500 (he kept \$200 for other expenses). Unfortunately, when he delivered the first copies to Bell Labs he did not think to bring an invoice and he still owed the printer, plus he needed funds for the next print run. The bank still refused a loan and the HFC office that issued the first check told Garfield that \$500 was the legal limit in New Jersey and they could not provide additional funding. However, the innuendo was that perhaps another office could (at that time there were no tracking systems in place). So he went to another HFC office and was given the funds because no one asked if he had received a prior loan elsewhere. Within less than two weeks he received Bell Lab's payment and all was well. The Bell contract ultimately lasted for eight years and Garfield continued to use HFC, not banks, to fund his company operations over the years (15).

It was around this time that Garfield, at the advice of a public relations person, changed the name of his company to "Eugene Garfield Associates - Information Engineers" in order to avoid confusion with Documentation, Inc. an already-established company founded in 1952/3 by Mortimer Taube (an attendee of the symposium Garfield organized while at Johns Hopkins). It should be noted that even this new name came under fire from the Pennsylvania Society of Professional Engineers who contacted Garfield to inform him that it was illegal to call oneself an engineer in Pennsylvania without formal training in the field, but the name stayed put until 1960. He also changed the name of his product to *Current Contents, Management and Social Sciences* (16).

By 1957, a medical librarian at Miles Laboratories, Charlotte Studer Mitchell, suggested that Garfield create a service covering journals in medicine and related sciences similar to that which he was providing for Bell Labs for her organization (17). Soon other pharmaceutical companies such as Lederle Labs and Warner Lambert began to express interest. And in 1958, not that long after creating *Contents in Advance*, the first concept noted earlier that Garfield developed while working on The Welch Project came to fruition in the form of *Current Contents of Chemical, Pharmaco-Medical, & Life Sciences* (changed to *Current Contents, Life Sciences* in 1967). Sales were subscription-based, but only to institutions (individuals could not subscribe) and the minimum order was twenty-five copies

at \$60.00 per copy. There were no indexes, but production was regular, the content was current, and the number of customers grew. 1958 also noted other changes: Garfield moved his offices out of the log cabin to a building at 15th & Spring Garden in Philadelphia, PA, across the street from Smith, Kline & French where he continued to work part-time as a consultant; he hired Marvin Schiller, a student at Penn State University, as a part-time marketing consultant to help promote and sell *Current Contents*; and he hired his first full-time employee, my dear friend Beverly Bartolomeo, who would do a little bit of everything in those early days, from page-layout to office cleaning. She told me that her father drove her to the interview with Garfield, and when he saw the dilapidated building, waited for her outside to make sure all was well.

But there simply were not enough industrial subscriptions to allow the service to remain sustainable long-term and university professors and other individuals began requesting subscriptions. An educational rate was set at \$50 per copy and the University of Wisconsin McArdle Laboratories became the first university subscriber. Direct mail promotions proved successful and by 1960 a second *Current Contents* edition was released, *Current Contents of Space, Electronic & Physical Sciences Including Pure & Applied Chemistry* (later changed to *Current Contents, Physical Sciences*). In that same year Garfield changed the name of his company for the last time. It now became the Institute for Scientific Information (ISI®), providing “a more institutional setting” for a new service that was about to be launched based on the second concept that Garfield developed while at Johns Hopkins. This was a new approach to indexing the chemical literature in the form of *Index Chemicus*®.

The seeds of all of Garfield’s early concepts had now been sown. Those of *Current Contents* had already taken root and begun to blossom while the fruits of the other two would soon emerge.

ISI – The Early Years 1960 - 1970

As noted earlier, in the years following the close of World War II abstracting and indexing services significantly lagged behind the published journal literature. Garfield had been acutely aware of this while at Johns Hopkins and solving the problem for science in general and specifically for chemistry remained one of his personal passions. The problem was reinforced while he worked on a three-year project for the Pharmaceutical Manufacturers Association (1957 - 1960) to index steroid chemical compounds that appeared in the journal literature so that the U. S. Patent Office could more quickly perform journal literature searches while processing new patent filings (18). In parallel to this project he was working on his Ph.D. in structural linguistics at the University of Pennsylvania and his thesis was on an algorithm for translating chemical names to molecular formulae. Garfield had come to the realization that every article he processed for the PMA project included the molecular formulas of the compounds that were being indexed. Since these formulas were quite prominent within the articles, he believed that he could very quickly produce an index to new compounds. He attempted to get *Chemical Abstracts* to adopt the concept, but they turned the idea down.

He applied his concept while doing a consulting job at Smith, Kline & French and tried to convince them as well that he could index the new compounds that appeared in one hundred core chemistry journals and produce a current monthly index for about \$25K. They did not believe him but gave him permission to try the initiative on his own – if it worked they would become a customer. He convinced twelve other companies to put in \$2K each. The group of participants met and suggested that a graphical abstract be included along with the formulas. Garfield agreed and the first issue of *Index Chemicus* was launched in June of 1960 (19), eventually spawning a family of chemical information products services. And while *Index Chemicus* was being launched, the seeds of the *Science Citation Index*® were beginning to sprout as well.

As noted earlier, Garfield's *Science* paper on citation indexes had captured the interest of Dr. Joshua Lederberg. In 1958, the year that Lederberg won the Nobel Prize, he wrote Garfield asking whatever happened to the idea. After some correspondence passed between the two, Lederberg suggested that Garfield apply to NIH (not NSF) for a grant to pursue the idea (NSF only provided service contracts, not grants, to for-profit organizations). A three-year grant (1960 - 1963) was negotiated to produce a genetics citation index and to determine the feasibility and scope of a citation index to the scientific literature. Mid-project the government rules changed and NIH was no longer allowed to offer grants to for-profits. All existing grants had to be converted to service contracts. Garfield's project was handed over to NSF because of their experience with the contracts and the result was a deal to create one thousand copies of the *Genetics Citation Index*. To do this, he had to first create a multidisciplinary database of 1.4 million citations from which genetic-specific citations could then be extracted (that data was taken from journals published in 1961). Garfield recommended that the contract include the production of a multidisciplinary citation index for evaluation by the scientific community. NSF rejected the idea and Garfield took the risk to publish it himself. In July 1963 the *Genetics Citation Index (GCI)* was published. Later that year, after securing sufficient orders to cover the cost, the first edition of the *Science Citation Index (SCI)*® based upon the 1961 data used for the *GCI* was released (20). The following year the *SCI*® was created using the most current journal literature. The product was enhanced with the *Permuterm Subject Index* in 1966 to enhance searchability. This innovative index was (and continues to be) created by pairing all of the significant words in an article title, with each pair becoming a separate entry in the index.

By the mid-1960's *Current Contents*® was making money, but barely supporting *Index Chemicus*® and the *Science Citation Index*®, both of which had not yet broken even. ISI® had grown to about one hundred employees and a level of senior management was in place, including four Vice Presidents, all of whom had serious financial concerns about the company. To resolve these concerns, 20% of ISI was given to some Wall Street investors in return for a half-million convertible debenture. Production costs also had to be shaved. The Vice-Presidents were unhappy with Garfield's leadership and demanded that he resign. He did not and they all left to form their own short-lived company, Information Corporation of America. One of the VP's, Art Elias, became a major player at BIOSIS. Another, Dr. Irving Sher, eventually returned to ISI and until

his death in 1996 contributed significantly to the growth and enhancement of ISI's products and services.

Ironically, it turned out that the Wall Street money was not needed and ISI continued to grow. I joined in April of 1967 as an indexer for *Index Chemicus* and by then the company had re-located to rent several floors of a building near Independence Hall at 325 Chestnut Street in Philadelphia, PA. The following year, the *Index Chemistry Registry System* (ICRS®) was created to allow for computer searching of the chemical structures that appeared in *Index Chemicus*. The system utilized the Wiswesser Line Notation (WLN), which was the first line notation capable of precisely describing complex molecules. A group of the chemical indexers, myself included, were taught the encoding process by Bill Wiswesser. And several of us became active in the Chemical Notation Association (CNA), now morphed into the Chemical Structure Association Trust. Many of the pharmaceutical companies of the day (ICI, Beecham, Glaxo, Sandoz, Hoffman-La Roche, etc.) were using the notation for their in-house files and wanted the ISI files of chemical structures as well. ICRS allowed that to happen.

And as ISI's chemical information services continued to expand so did *Current Contents*. In January 1969 *Current Contents, Education* was launched and March of that same year witnessed the release of *Current Contents, Social, Behavioral, & Management* (these two editions would be merged in 1971). The 1960's also saw the launch of other services such as the *Original Article Tear Sheet Service* (OATS®) (21) that offered a one-day turnaround of original articles cut from journals. Photocopies of the same article, if ordered, would be made from the master copy of the journal (ISI received several copies of each journal issue). This service later became known as the *Genuine Article*. Another service was the *Automatic Subject Citation Alert* (ASCA®) (22), a customized weekly product introduced in 1965 that provided printouts listing articles on a subject specifically requested by the subscriber.

So despite the fear of insolvency and the "revolution" of senior management, ISI did more than just survive its early days. It created ancillary services to meet the evolving needs of its customer base and by the time ISI moved into the next decade, both *Current Contents* and the *SCI* were in the black.

ISI – The Middle Years 1970 – 1980

During this decade ISI continued to enhance its existing product lines, improve operations, and add new services. It was a transition period between the print world and the soon-to-emerge digital age. Computers were heavily relied upon as a key component of production. As someone who had to deliver trays of the newly-keyed IBM cards to the computer room I lived in fear of dropping them since they were filed in a specific sequence. We had to draw a huge "X" across the top of the cards so that if the tray was dropped the cards could be put back in sequence by replicating the "X." The magnetic tapes that were created for production were also used as a delivery channel for the corporate customers who had the requisite hardware. And, since this decade witnessed the advent of the online era and ISI was an early adopter, the tapes were used to deliver all of

ISI's databases to the online host services that had emerged such as Dialog, BRS, SDC, DataStar, DIMDI, ESA, etc.

The *Current Contents* product line continued to expand, with *Current Contents/Agricultural, Food & Veterinary Science* and *Current Contents/Engineering and Technology* both being launched in January 1970 (23). At the end of that same year the coverage of *Current Contents/Chemical Sciences* was merged into *Current Contents/Physical & Chemical Sciences*. The Clinical Practice edition was released in January 1973 (24) (the title would be later changed to Clinical Medicine), and the *Arts & Humanities* edition closed out the decade with its launch in January 1979 (25). Also, Garfield's essays that appeared intermittently in *Current Contents* became a regular feature in the early 1970's. In 1974 it was decided to collect and published the essays in chronological order under the title *Essays of an Information Scientist*. To do this ISI Press was established in 1977, ultimately publishing other titles as well (26).

The citation index product line was expanded with the launch of the *Social Science Citation Index* in 1973 (27) and the *Arts & Humanities Citation Index* in 1978 (28). An offshoot was developed from the citation files in 1973. This was *The Journal Citation Reports (JCR)* (29). It listed the top one-thousand most-cited journals; ranked the journals by impact factor (the average number of citations per article published); provided a detailed list of all journals by which a given journal was cited; and a listing of what journals each of the top journals themselves had cited. At that time each section could be purchased separately. It has arguably become one of the most controversial services that ISI ever created as it has often been used for purposes for which it was not created. The impact factor reigned for years in the print world as *the* metric for evaluating research. Today, there are other metrics such as the Eigen Factor, Plum Analytics, article-level metrics compiled by the Public Library of Science, etc. But the annual release of the *JCR* and its impact factor listings continue to this day to be eagerly awaited by journal editors and publishers.

The chemical information products also expanded during this time period (although they remained a financial vulnerability). In January 1971 the printed *Chemical Substructure Index (CSI)* (30) was released. This was a monthly index to the new compounds that appeared in the weekly issues of *Current Abstracts of Chemistry* and it also had an annual cumulation. *CSI* facilitated the searching for compounds without needing to use chemical nomenclature. Based upon the Wiswesser Line Notation mentioned earlier, the linear notation for each structure was rotated and a separate index entry was created for each substructure - an average of six entries per compound. One could easily locate specific ring systems, all compounds with specific functional groups, etc. This service represented a significant breakthrough at a time when personal computers did not yet exist and the drawing of chemical structures to perform searches was just a dream. The product was also made available on microfilm in 1976. A personal alerting service similar to ASCA® was also released. This was the *Automatic New Structure Alert (ANSA®)* (31) that provided subscribers with printouts of newly-published compounds that met their structure/substructure criteria. But my personal favorite was the launch of the monthly *Current Chemical Reactions™* in 1979 (32). Like *Current Abstracts of Chemistry and Index Chemicus™*, this print product included

flow diagrams and the author's abstract, but focused on new and newly-modified synthetic methods. Four indexes were included – author, address, subject, and journal along with a cross-reference (if appropriate) to *CAC&IC*.

ISI ended the 1970's on an extremely positive note. On October 17, 1978, ground was broken for its new headquarters at 3501 Market Street in Philadelphia, PA, in the heart of the University City Science Center. The four-story building was the first office to be built by the award-winning firm of Venturi & Rauch. Garfield also planned to (and did) build a day-care center across the street from the rear of the building. He also incorporated original artwork in the building, a portion of which had to be accessible to the public, and specifically commissioned murals for the external walls of the day care center (33). The actual move into the new building took place over a series of days in the fall of 1979. By then ISI had nearly 500 employees. The building was built to handle future growth and so empty space on the fourth floor was originally rented to others. One of the first tenants was Richard Buckminster Fuller, the inventor of the geodesic dome. ISI staff all received collapsible cardboard versions of the dome as a memento.

ISI – the Digital Years and the End of an Era (1980 – 1992)

The 1980's kicked off the start of the digital age for the masses with the launch of the IBM PC model 5150 on August 12, 1981 (34). ISI had been using computer technology since the 1960's, but other than delivering information on magnetic tape to a select set of customers, that use was behind the scenes for production. But the PC was about to move computers front and center for in the following year about three million microcomputers were shipped to users within the USA (35).

ISI's first product targeted to the PC user was announced in March 1983 (36). This was *Sci-Mate*, a software package that would allow for online and offline access to and retrieval of information. It was also a database management system, originally conceived for the organization of reprints, but enhanced to manage all sorts of information. However, it would not be until later in the decade that ISI would actually release its own content on magnetic media other than tape for computer searching.

Throughout the early 1980's ISI continued to develop and enhance online versions of all of its information products for computer access. In June of 1984 *Index Chemicus* went online via Telesystemes using the DARC/Questel system (37). In 1986 a much more ambitious chemical information initiative was undertaken in the form of the *Current Chemical Reactions InHouse Database* (38). That year we actively contacted the large chemical and pharmaceutical companies in the US and Europe to determine their interest in having an in-house searchable file of chemical reactions and gauging their willingness to fund its development. We went with an in-house approach rather than online access via a third party vendor because many of the companies were concerned about confidentiality and the proprietary nature of their searches, and they knew that there was a significant cost attached to having large numbers of employees performing frequent searches online. In less than a year more than twenty companies, including 3M, Schering

AG, Monsanto, Pfizer, Ciba-Geigy, Hoffman-La Roche, etc., had all signed up to be part of what was called The Reaction Data Club. They served as editorial advisors and had a say in the development of the database features and functionalities. The software used to support the system was *REACCS* from Molecular Design Limited, the leading chemical software developer at the time. The project was hugely successful and the service was launched in 1987. In that same year ISI released its very first information products on diskette, *Index Chemicus* and *Current Chemical Reactions Personal Databases*, allowing PC access to new chemical compounds and synthetic methods (39). The reaction data was searchable using Molecular Design Limited's *Chembase* and the structure data was searchable by both *Chembase* and *ChemSmart*, the latter software developed by Scott Gould (40) and marketed by ISI. There was a lot of excitement about the diskette products, but sadly, they were before their time. In 1987 researchers still relied pretty much on information specialists and librarians for information retrieval and PC usage was still relatively new. The products were discontinued in less than two years.

The next product to be released specifically for PC access by end users was far more successful, the CD-ROM version of the *Science Citation Index*. After a year of development effort it was made available in May 1988 on two searchable discs (41). And for the first time searchers were able to display related records (the papers that share references with paper that resulted from the search). The CD-ROM version of the *Social Science Citation Index* came out the following year (42).

In addition to creating e-versions of the core products that emerged from Garfield's original ideas, ISI continued to provide new offerings. The *ISI Atlas of Science: Biochemistry and Molecular Biology* was produced in 1981 (43). This covered 102 subspecialties of those disciplines. Each specialty had its own chapter that included a review on the topic written by a scientist in the field, a "cluster" map that showed the relationships among the core documents in the specialty, a bibliography of the core documents, and a list of current papers that cited those core documents. The Atlas was a stand-alone product as well as a support tool for the *Science Citation Index* from the perspective that the core documents could be used as a starting point for searches. In 1981 ISI also began to develop a series of discipline specific indexes (44). Each discipline had four components: a print citation index covering the years 1951-1980, a print annual cumulation, online access to the database, and a monthly current awareness service similar to *Current Contents*. The first, *ISI/BIOMED*, was released in 1981. The second, *ISI/COMPUMATH*, was released in March the following year and the third, *ISI/GeoSciTech*, followed in July of that same year. During the same time period ISI Press expanded under the leadership of Robert Day. ISI Press had published Day's book, *How to Write and Publish a Scientific Paper* in 1979. The book was a huge success and Day, who had served as managing editor of many of the journals published by the American Society of Microbiology and who had also served as Chair of the Council of Biology Editors and President of the Society for Scholarly Publishing, joined ISI in 1980 as the Press' new Director. In 1982 two additional "how to write" books were published for specific disciplines, one for engineering and one for medicine, a book on abstracting was published as was a book on

communication skills for the foreign professional. And last, but not least, with the first edition of *The Scientist* in October 1986, Garfield fulfilled a long-term dream of publishing a newspaper of science, the goal of which was to address the specific professional needs of scientists in general (45).

But ISI's expansion efforts and adaption to the digital age came with a price tag. There was a market perception that diskette and CD-ROM products saved the publisher a lot of money because there were no massive print runs. Nothing was further than the truth. The cost of creating the information did not change. Yes, print runs went away, but customer support costs escalated. No one ever called customer support to ask how to turn the page of a book. But the phone rang off the hook when the diskette and CD-ROM products were launched. In fact our help desk activity grew 581% between 1987 when no CD-ROM or diskette products were offered and 1990 (46). The majority of users (librarians included) were unfamiliar with e-products and were not completely computer literate. User manuals were supplied, but they were left in their wrappers unread. In addition, the cost of software upgrades, enhancements, etc. meant that there was an ongoing investment in the products that had not existed before. Once again ISI management had serious financial concerns and in early 1986 the Board insisted that Garfield bring in a second-in-command to take control of day-to-day-operations (47). I will never forget the day, while Dr. Garfield was out of the office for medical reasons, that the new CEO fired all of the Vice Presidents with the exception of me and quickly built a new level of management. It was a very difficult and unpleasant period of about eighteen months and the financial situation only worsened. Fortunately Dr. Garfield regained control and terminated the CEO, but as he himself has said, a lot of damage had been done (48).

It was around this time that Garfield was introduced to Theodore Lamont Cross who owned a small company entitled "JPT Publishing Group" ("JPT" for the first names of the principals, Joe Palazolo, Paul Neuthaler, and Theodore Cross) and who had previously run the publishing firm of Warren, Gorham and Lamont along with his two brothers. Cross obtained more than 50% control of ISI in 1988 (the prior year he had tried to buy Harper & Row for a reported \$190 million, but lost out to Rupert Murdoch) (49). With the advantage of hindsight, I can see that he began to shape ISI for an eventual sale, but at the time I was pleased that they tried to understand the products and invested in their improvement. Of particular note was the infusion of cash to launch *Current Contents on Diskette* starting with the Life Sciences edition in September 1988 (50). After much discussion and debate, they also supported the addition in 1991 of author abstracts to the diskette editions of *Current Contents/Life Sciences*, *Agriculture, Biology & Environmental Sciences*, *Physical, Chemical & Earth Sciences* and *Clinical Medicine* (51). The release of these new editions was announced with much fanfare at a reception held at the National Online meeting in New York in May 1991.

But in parallel to the growth of some products, a number of initiatives were easily cut since Garfield no longer had control. The CEO mentioned earlier had shut down the ISI Press before JPT came on the scene. JPT shut down the *Atlas of Science* and the ISI Day Care Center. They also sold *The Scientist* to Garfield for one dollar. Then on April 10, 1992, almost four years after JPT took over control, ISI was sold to Thomson Business Information, a subsidiary of the Thomson

Corporation. I remember the day well as I had to stay back from the Spring National Meeting of the American Chemical Society in San Francisco in order to participate in making the announcement to staff. After the announcement, the Vice Presidents met with Thomson staff and we each were asked what we thought of the acquisition. My response was to request that they ask me in a year since there was no way I could know what the acquisition would bring. They remembered my response and a year later, almost to the day, Michael Brown, then President of Thomson, asked the question again. And my response was positive - they had treated us well, we had access to more financial resources than ever, and Thomson ISI (as it was briefly known) again began to grow. That same year, in December 1993, Garfield announced the end of his weekly essays in *Current Contents* and on January 1, 1994 became ISI's Chairman Emeritus, serving as a consultant and member of the Advisory Board and retaining office space in the building that he had built just sixteen years earlier (52). I left ISI just about one year later in January 1995.

In retrospect, I cannot imagine a better place to have started my career than at ISI. It opened a whole new world to me and motivated me to follow a non-traditional path for a chemist - one in scientific publishing. Garfield was and remains a creative force and was in many ways a nurturer of careers. Regardless of age, gender or ethnicity, he supported you if you had good ideas and worked hard. And as a chemist, he was totally committed to those of us on staff who served as volunteers for the American Chemical Society even though in some ways the organizations were competitors. He himself served the current Division of Chemical Information when it was the Division of Chemical Literature, both as an active committee chair and in the presentation of papers (53).

ISI was a crazy place to work in the early days and I have gone into some detail on that environment elsewhere, but I will re-iterate here. "People parked their motorcycles at their desk. The work dress ranged from normal to eccentric. One executive always wore a teddy bear on his belt and another staff member wore baby doll pajamas on occasion (these two streaked together at an ISI party!). When my boss complained about the length (or lack thereof) of miniskirts, the corporate (unofficial) response was that the only dress coded requirement was shoes! The examples are endless" (54). Obviously, it was not a bureaucratic company, but rather it had the look and feel of a family run operation. In truth, throughout the years members of Garfield's family actually worked for ISI. In the log cabin days his wife laid out the pages of *Current Contents* for production. His son Joshua worked on the *Atlas of Science*. His stepson, Peter Aborn, had the largest role at ISI. He was Vice President, Administrative Services at the time ISI was building its own headquarters and was a key player in planning the facility and coordinating the actual move. Peter also was the driving force in the creation and building of the ISI Day Care Center. But to Garfield we were all family. In his essay from December 22, 1975 he printed the names of all of ISI's 336 employees as part of his Happy New Year wish to subscribers, stating that "ISI is people – not paper, systems and machines" (55). To this day those of us who worked there back then still feel that way. And the feeling was palpably clear when Dr. Garfield held a reunion for us at his home on September 25, 2010 in celebration of his 85th birthday.

Thomson sold the ISI building a few years ago to Drexel University and moved to rented space at 1500 Spring Garden Street in Philadelphia, PA. When I visited there just a few months ago I was struck by the irony of it. Here was ISI - back within a stone's throw from the dilapidated building in which it rented space in 1958 and that witnessed the hiring of ISI's first full-time employee. Gone was the frenetic, entrepreneurial, family-like atmosphere that permeated the company. It has, as is natural, evolved into a completely different organization at a later stage of its life cycle. Yes indeed, an era has ended. But Garfield's legacy has not. As I looked around, I realized that hundreds of people owe their jobs to the fruit of his work as Thomson continues to build on what he created. One of Garfield's dreams, a book citation index, was launched at the end of 2011, and a year later a data citation index was launched in response to the increased importance of being able to discover and access data sets. The citation indexes, content-page products, and chemical information services that he conceived and developed will long continue to inform and empower successive generations of researchers around the world. After all, one should not overlook the fact that it was his creative use of citations as an indexing tool that sparked the creativity of Google founders Larry Page and Sergey Brin who cited Garfield in their academic work on PageRank, the algorithm that powers their company's search engine (56). Yes, indeed, his legacy will continue to live on in many forms!

References

1. Garfield, E. *Essays of an Information Scientist*; ISI Press: Philadelphia, PA, 1977; Vol. 2, p 33.
2. Williams, R. V.; Bowden, M. E. *Chronology of Chemical Information Science*; <http://faculty.libsci.sc.edu/bob/chemnet/CC1900.HTM> - 1940 (accessed on June 25, 2014).
3. Garfield, E. *Essays of an Information Scientist*; ISI Press: Philadelphia, PA, 1977; Vol. 2, p 33.
4. <http://www.webofstories.com/play/eugene.garfield/25> (accessed on June 25, 2014).
5. Garfield, E. *Essays of an Information Scientist*; ISI Press: Philadelphia, PA, 1977; Vol. 2, p 32.
6. Lawlor, B. Abstracting and Information Services: Managing the Flow of Scholarly Communication – Past, Present and Future. *Serials Review* **2003**, 29, 199.
7. Wiggins, G. *What is Chemical Information?* Indiana University: August 11, 2002. http://www.indiana.edu/~cheminfo/acs800/soced_wash.html (accessed on June 25, 2014).
8. The History of Science Abstracts, The Institution of Electrical Engineers (IEE), <http://www.theiet.org/resources/library/archives/inspec/1898-1914.cfm> (accessed on June 25, 2014).
9. Baker, D. B. *Interview by Robert V. Williams and Leo B. Slater at Columbus, OH*, 9 June 1997 (Philadelphia: Chemical Heritage Foundation, Oral History Transcript #0160).

10. Mabe, M.; Amin, M. Growth Dynamics of Journals. *Scientometrics* **2001**, *51*, 147.
11. Garfield, E. *Of Nobel Class, Women in Science, Citation Classics, and Other Essays*; ISI Press: Philadelphia, PA, 1993; Vol. 15, p 84.
12. Cawkell, T.; Garfield, E. Institute for Scientific Information. *A Century of Scientific Publishing*; Fredriksson, E. H., Ed.; IOS Press: 2001, Chapter 15, p 151.
13. Garfield, E. *Creativity, Delayed Recognition, and Other Essays*; Vol. 12, p 53, ISI Press, Philadelphia, PA, 1991
14. Garfield, E. Citation Indexes for Science. *Science* **1955**, *122*, 108.
15. Garfield, E. *Essays of an Information Scientist*; ISI Press: Philadelphia, PA, 1981; Vol. 4, p 359.
16. Garfield, E. *Essays of an Information Scientist*; ISI Press: Philadelphia, PA, 1993; Vol. 1, p 13, 438.
17. Garfield, E. *Essays of an Information Scientist*; ISI Press: Philadelphia, PA, 1993; Vol. 1, p 13.
18. Williams, R. V. *An Interview with Eugene Garfield*; conducted July 29, 1997, Philadelphia, PA, The Chemical Heritage Foundation; p 48.
19. Garfield, E. *Essays of an Information Scientist*; ISI Press: Philadelphia, PA, 1977; Vol. 1, p 18.
20. Garfield, E. *Essays of an Information Scientist*; ISI Press: Philadelphia, PA, 1977; Vol. 1, p 192.
21. Garfield, E. *Essays of an Information Scientist*; ISI Press: Philadelphia, PA, 1983; Vol. 6, p 185.
22. Garfield, E. *Essays of an Information Scientist*; ISI Press: Philadelphia, PA, 1980; Vol. 3, p 640.
23. Garfield, E. *Essays of an Information Scientist*; ISI Press: Philadelphia, PA, 1977; Vol. 1, p 68.
24. Garfield, E. *Essays of an Information Scientist*; ISI Press: Philadelphia, PA, 1977; Vol. 2, p 144.
25. Garfield, E. *Essays of an Information Scientist*; ISI Press: Philadelphia, PA, 1980; Vol. 3, p 556.
26. Garfield, E. *Essays of an Information Scientist*; ISI Press: Philadelphia, PA, 1977; Vol. 2, p 173.
27. Garfield, E. *Essays of an Information Scientist*; ISI Press: Philadelphia, PA, 1977; Vol. 2, p 471.
28. Garfield, E. *Essays of an Information Scientist*; ISI Press: Philadelphia, PA, 1980; Vol. 3, p 204,
29. Garfield, E. *Essays of an Information Scientist*; ISI Press: Philadelphia, PA, 1977; Vol. 2, p 473.
30. Garfield, E. *Essays of an Information Scientist*; ISI Press: Philadelphia, PA, 1977; Vol. 2, p 184.
31. Garfield, E. *Essays of an Information Scientist*; ISI Press: Philadelphia, PA, 1977; Vol. 2, p 142.
32. Garfield, E. *Essays of an Information Scientist*; ISI Press: Philadelphia, PA, 1981; Vol. 4, p 12.

33. Garfield, E. *Essays of an Information Scientist*; ISI Press: Philadelphia, PA, 1983; Vol. 5, p 15 (see also the color insert between pages 244 and 245 in Garfield, E. *Towards Scientography*; ISI Press: Philadelphia, PA, 1988; Vol. 9).
34. Orion, E. The First IBM PC came out 30 years ago today. *Philadelphia Inquirer*; August 12, 2011.
35. Friedrich, O. The Computer Moves In. *Time* **1983**, 121 (1), 14–24.
36. Garfield, E. *Essays of an Information Scientist*; ISI Press: Philadelphia, PA, 1984; Vol. 6, p 80.
37. Garfield, E. *The Awards of Science and Other Essays*; ISI Press: Philadelphia, PA, 1984; Vol. 7, p 194.
38. Garfield, E. *Peer Review, Refereeing, Fraud, and Other Essays*; ISI Press: Philadelphia, PA, 1987; Vol. 10, p 83.
39. Garfield, E. *Peer Review, Refereeing, Fraud, and Other Essays*; ISI Press: Philadelphia, PA, 1987; Vol. 10, p 59.
40. Heller, S. R. *Chemical Substructure Searching on a PC*; <http://www.hellers.com/steve/resume/p105.html> (accessed on June 25, 2014)
41. Garfield, E. *Science Literacy, policy, Evaluation, and Other Essays*; ISI Press: Philadelphia, PA, 1988; Vol. 11, p 160.
42. Garfield, E. *Creativity, Delayed Recognition, and Other Essays*; ISI Press: Philadelphia, PA, 1989; Vol. 12, p 256.
43. Garfield, E. *Essays of an Information Scientist*; ISI Press: Philadelphia, PA, 1983; Vol. 5, p 279.
44. Garfield, E. *Essays of an Information Scientist*; ISI Press: Philadelphia, PA, 1983; Vol. 5, p 11.
45. Garfield, E. *Towards Scientography*; ISI Press: Philadelphia, PA, 1986; Vol. 9, p 222.
46. Lawlor, B. Pricing, Marketing, Customer Support, and Legal Implications; *Information Distribution for the 90s*; NFAIS Report Series 1991; NFAIS: Philadelphia, PA, p 91.
47. Knox, A. ISI founder Seeks Ways to Finance His “Historic Mission”; *Philadelphia Inquirer*, June 30, 1986; http://articles.philly.com/1986-06-30/business/26046303_1_isi-current-contents-science-citation-index (accessed on June 25, 2014)
48. Williams, R. V. *An Oral Interview with Eugene Garfield*, conducted in July 29, 1997, Philadelphia, PA, The Chemical Heritage Foundation; p 81.
49. Fox, M. Theodore Cross Dies at 86, A Champion of Civil Rights *The New York Times*; March 30, 2010.
50. Garfield, E. *Science Literacy, Policy, Evaluation, and Other Essays*; ISI Press: Philadelphia, PA, 1988; Vol. 11, p 3311.
51. Garfield, E. *Science Reviews, Journalism, Inventiveness and Other Essays*; ISI Press: Philadelphia, PA, 1991; Vol. 14, p 74.
52. Garfield, E. *Of Nobel Class, Women in Science, Citation Classics and Other Essays*; ISI Press: Philadelphia, PA, 1993; Vol. 15, p 408.
53. Metanomski, V. *50 Years of Chemical Information in the American Chemical Society*; American Chemical Society: 1993; pp 12–13.

54. Baykoucheva, S. Interview with Bonnie Lawlor. *Chemical Information Bulletin*; Vol. 62, No. 1, Spring 2010; <http://bulletin.acscinf.org/node/188> (accessed on June 25, 2014).
55. Garfield, E. *Essays of an Information Scientist*; ISI Press: Philadelphia, PA, 1977; Vol. 2; p 399.
56. Rumsey, E. Eugene Garfield: Librarian and Grandfather of Google, posted July 12, 2010; <http://blog.lib.uiowa.edu/hardinmd/2010/07/12/eugene-garfield-librarian-grandfather-of-google/> (accessed on June 25, 2014).

Chapter 8

The Making of Reaxys—Towards Unobstructed Access to Relevant Chemistry Information

Alexander J. Lawson,¹ Jürgen Swienty-Busch,^{*,2} Thibault Géoui,²
and David Evans¹

¹Reed Elsevier Properties SA, Espace de l'Europe 3,
2000 Neuchâtel, Switzerland

²Elsevier Information Systems GmbH, Theodor-Heuss-Allee 108,
60486 Frankfurt, Germany

*E-mail: J.Swienty-Busch@elsevier.com

An account is given of the history leading to the launch of the chemical information system Reaxys in 2009, its subsequent development until 2014, and outlook for the future. The path leading from the print form of the two major chemical Handbooks of the 19th century through the building of online databases of the late 1980s and the client/server system of CrossFire (1993-2009) is discussed with particular emphasis placed on the importance of technological development in creating user needs that in turn require an ongoing overhaul of the same technology to better serve the market. The evolution of the Gmelin and Beilstein Handbooks from property-centered collections of chemical structures into the premium data sources of chemical information in CrossFire in the early years of the 21st century is one excellent example of this phenomenon, and the subsequent development of CrossFire and its databases to Reaxys is shown as a second inevitable consequence of the same driver. The account closes with a description of some currently evolving trends and the first steps taken in Reaxys to continue this tradition of innovation.

1. Introduction

Reaxys (1) descends from two of the great data catalogs born from the proliferation of scientific knowledge in the nineteenth century. Professionalization and institutionalization of scientific practice in the early 1800s fostered discoveries and technological innovations that bred new questions and lead to further experimentation and developments. Chemistry was no exception. In the words of Pierre-Joseph Macquer (1776): “You know what the condition of chemistry is today: only a child two days ago, it suddenly finds itself in an incredible state of growth, and is changing into a colossus” (2). This explosion in scientific research and the accompanying accumulation of data prompted a drive to collect, organize, and record knowledge in the best information vehicle known at the time—books.

Communication from author to reader was not an easy task. Quite apart from the barriers of national language, standard conventions were not yet in place for the description of either textual (nomenclature) or graphical (structural) representation of chemical entities, or even the measurement of some experimental data, such as melting or boiling point. The use of such printed works therefore involved (and indeed required) special levels of chemical expertise and interpretation of the author’s own methodology, which in turn was explained at length. Despite the didactic approach of the early volumes, these works were not for laypersons; they reviewed chemical research results for an audience of chemists. This user base has not changed with time: over the course of two centuries, catalyzed by advances in technology, some of these printed data compendia evolved into the rich information systems used today, but the user base remains firmly in the area of chemical researchers, whatever other specialization (such as “information specialist” or “medicinal chemist”) they may have attained.

The focus of the printed data repositories in the 19th and 20th centuries was the collection and classification of information into a condensed format, apt for a book. Data extraction from source documents and data cataloging into a book was a time-consuming operation where input was carefully selected and processed. The advent of computer systems in the mid 20th century lifted content restrictions for databases. As a result, focus shifted from the selection and condensation of information to the efficient capture of rapidly growing volumes and diversified sources of information, as well as the development of indexing schemes that made databases searchable. By the end of the 20th century, graphical user interfaces allowed the focus to return to the research chemist. User interface design aimed to create a more natural portal into databases, where queries could be constructed without knowledge of programming or database structure and where search results could be organized, filtered and evaluated by the user.

As shown in Figure 1, the evolution of Reaxys mirrors this historical course. The focus of development efforts shifted from information source, to database structure and access, and then to the user. This evolution begins with the legacy found in the Reaxys core data repository. Built from the authoritative data collection and classification of the Beilstein and Gmelin Handbooks, this rich compendium is organized according to a unique scientific model that emphasizes the convergence of substance, property and reaction data (see Section 2). As these immense printed datasets evolved into electronic databases, extensive

indexing enabled unique and powerful search functionality, albeit only accessible to information specialists and expert users. Improvements in the 1990s brought the database closer to the chemist in the form of the CrossFire system (3) and initiated the erosion of the inherent barrier between a sophisticated information system and its general user base (see also Section 3). Then, an extensive overhaul brought about the conception and launch of Reaxys in 2009. Featuring enhanced functionality for a core audience, later complemented with improvements relevant to any chemist, Reaxys evolved into a system designed to respond to the needs of all potential users, regardless of training (see Section 4). Finally, as we progress toward Web 3.0 (4) in the 21st century, the backend—the collection of processes that interpret user queries and extract results from the database—moves into the limelight. Contextual query interpretation, meaningful linking of data entries above and beyond “synonym” or “is related to”, and matching data to a searcher’s needs without an explicit command are some examples of the expected backend processing that will reduce a complex search workflow to asking a simple question (see Section 5).

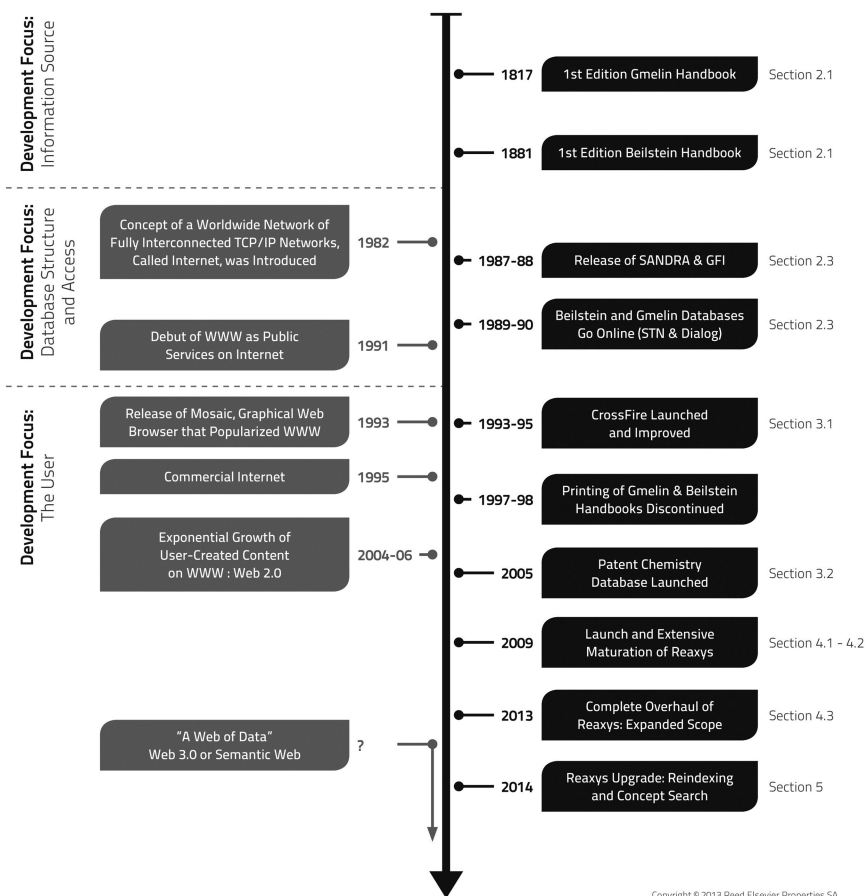
The history of Reaxys summarized in Figure 1 serves as the backdrop to a more detailed review of the diverse elements that have contributed to its distinctive aspects; these will be discussed in the following Sections.

2. A Visionary Organizational Heritage

2.1. The Gmelin and Beilstein Handbooks

The Gmelin Handbook started as the ambitious project of Leopold Gmelin (1788-1853) to gather and publish in one source all known data relevant to chemistry. The first edition was published in 1817. Gmelin underestimated the rapid growth of chemical data at the time and the Handbook was subsequently restricted in the 1850s to inorganic and organometallic compounds. Complementing this compendium was the collection of data on organic substances spearheaded by the chemist Friedrich Konrad Beilstein (1838-1906) that resulted in the Beilstein Handbook, with a first publication in 1881-83. Both Handbooks were massive undertakings. They assembled and systematically classified relevant research findings scattered throughout the primary scientific literature, reducing large amounts of information into readily usable form. By the time printing of the Handbooks was discontinued in 1997 and 1998 respectively, the Gmelin Handbook consisted of 760 volumes plus the 35 books of the Gmelin Formula Index, and the Beilstein Handbook included 503 volumes.

Beyond a collection of data entries, both Handbooks critically evaluated the data destined to be included in the compilation and organized that information into strict, chemically logical groupings. The data gathered in the Handbooks were reliable, relevant and structured. Several hundred factual categories organized under each compound were filled with experimental data (where available) and augmented as new information was published in the chemistry literature. It therefore sufficed to find the compound of interest in one of the Handbooks to obtain a standardized and comprehensive collection of structural, identification, physical, chemical, and reaction data.



Copyright © 2013 Reed Elsevier Properties SA.

Figure 1. The evolution of Reaxys: a timeline from the first publication of the Gmelin Handbook to the most recent update of Reaxys. Courtesy of Elsevier Information Systems GmbH and Reed Elsevier Properties SA, 2013.

2.2. The Power of a Structure-Based Organization

The organization of the main entries (i.e. the recorded compounds) was the key to the utility of the Handbooks and this is also where the Beilstein Handbook differed from Gmelin. The eighth and last edition of the Gmelin Handbook was organized according to the Gmelin System, whereby elements were assigned a System Number and compounds were catalogued under the constituent element with the highest System Number. Entries for each element were combined into a Handbook volume (and supplements) and arranged by increasing order of complexity and number of constituent elements (5). In other words, organization was based on molecular formula. Beilstein devised a classification system for his Handbook that emphasized the structure of a compound. Through a complex set of rules, every compound was assigned a unique Volume Number and System Number, which did not change over time and allowed the user to find

information in the same place in all Beilstein volumes and supplements. Each compound was assigned to one of three broad categories—acyclic, isocyclic and heterocyclic—and then clustered into structure-based subcategories (6, 7).

In using structure as its organizational backbone, the Beilstein Handbook emphasized the link between chemical structure and chemical properties, a concept which was at its infancy at the time of the Handbook's first publication. In an article on the Beilstein Database (8), Lawson describes the visionary nature of this classification system. Each entry into the Beilstein Handbook was a triad of data that emphasized linking citation-validated experimental data to a given chemical structure, as well as the effect of and the means to alter that chemical structure (i.e. a reaction). This structure-based data organization created a unique informational space that accommodates research methodologies inherent to the chemist's thought process: What does it look like? How do I make it? How does it behave? What happens if I change this functional group? And, increasingly important over the years, what is the influence of stereochemistry? As early as 1967, the Beilstein Handbook incorporated the Cahn-Ingold-Prelog priority rules into its classification system to generate the entry name of stereoisomers. Similarly, each compound in the Handbook was implicitly cross-referenced to existing entries of the starting materials used in its preparation.

In short, this chemistry informational space contained extensively interlinked data on substances, properties, and reactions that cannot be found elsewhere, and the scientific model behind it was preserved from the first catalogued entries of the Beilstein Handbook through more than 120 years of experience abstracting information in high data quality and detail. This pioneer work would later be fully exploited by CrossFire, the precursor to Reaxys (8, 9), as described in Section 3, but first, we must look briefly at the first public implementation of the Beilstein and Gmelin Databases on mainframe-based host systems in the late 1980s.

2.3. Digitization of the Gmelin and Beilstein Handbooks

From the viewpoint of market penetration, the Handbooks peaked in the mid 1960s. Many factors lead to a slow decline in subscriptions to the printed works, including language, price, content-currency and usability issues. The usability aspect was particularly acute in the case of the printed Beilstein work, because locating information on individual compounds (pre-SANDRA, see below) required expert understanding of the rules of classification described in the previous Section. The extent of this expertise in the chemist user base had declined over the years, and researchers in the chemical industry especially were encouraged to use the internal services of an expert intermediary; the terms "information specialist" and "end-user" arose to describe this relationship, which became stronger as more computerized systems (internal and external) became available to industrial chemists in the 1970s.

As purchases of the printed compilations declined in the 1980s, the Institutes managing their respective Handbooks began the task of computerizing the data contained in the Beilstein and the Gmelin Handbooks. The Beilstein Online Database and the Gmelin Online Database were made available via the STN and Dialog (10) hosts in 1989 and the early 1990s. This digitization occurred

stepwise. For example, the PC-software SANDRA (Structure and Reference Analyzer) preceded online database access and complemented the printed Handbook by facilitating its use. SANDRA, written by Lawson and launched in 1987, automated the Beilstein organizational rules giving the user the exact location of a compound in the multi-volume Handbook. A Gmelin counterpart, the GFI (Gmelin Formula Index), was made available on STN in 1988 and enabled Handbook entry searches that gave the volume and page number where the sought-after information could be found. Also, content for both databases was placed online in increments. The final versions were an asset to the scientific community, making possible structural and factual searches on 400 separate data fields in Beilstein (8) and over 800 in Gmelin (5).

Making the online databases available through STN and Dialog had its advantages and disadvantages. Both platforms offered very powerful search capabilities and an unparalleled collection of other bibliographic and reference databases. The search software used by each host, however, shaped the implementation of the Beilstein and Gmelin Databases. For example, the structural algorithms used by STN at the time did not allow steric searches (as was possible later with CrossFire). Additionally, knowledge of database organization and Boolean logic were required to perform comprehensive and efficient queries. The correct commands were needed and the complex cost structure made it prohibitive to just “surf” the content. Indeed, in view of the high cost of structure searching, many queries on these hosts were deliberately formulated to avoid graphical queries wherever possible, and often keys such as registry numbers, molecular formula and chemical name fragments were used instead. As a consequence, use of the online databases was limited to information specialists and expert users, and the specific advantages of the data model described above were not fully available until the CrossFire system revolutionized access to the digital data compendia.

3. The CrossFire Revolution

3.1. CrossFire Brings Information to the Chemist

Building on advances in computer operating systems and previous experience in developing independent software for CD-ROM versions of databases, the Beilstein Institute and the newly founded Beilstein Information Systems GmbH created and improved the CrossFire system in the span of two years (1993-1995). Under the CrossFire system, a PC client application called “CrossFire Commander” accessed the Beilstein and Gmelin Databases hosted initially on an in-house server, where user queries were run in real-time. Shortly after the introduction of CrossFire, the databases were moved to a central server and CrossFire Commander accessed the server via the internet. This model was known as CrossFire Direct.

The client/server architecture of CrossFire Direct was hailed by Meehan and Schofield as a “revolution” (9). With the databases physically under central control, maintenance and updates remained in hands of the database management staff, but any chemist could incorporate use of these resources into his or her daily

work. First, access was granted directly from a client personal computer and, with minimum training, chemists could conduct searches themselves rather than depend on an information specialist or librarian to broker the information. Second, the search engines developed for CrossFire enabled improvements in performance by one to two orders of magnitude (*11*), often shortening substructure search times from minutes to as many seconds. As a consequence, information searches became a manageable component of the daily workflow. Third, the cost structure was simplified to an annual subscription based on the number of users. With no time- or search-based charges, the subscription costs for companies, universities, NGOs and consortia were fixed and budget able in advance. Furthermore, this gave users liberty to perform exploratory searches with relaxed search criteria that might produce unexpected results and lead to discoveries or workflows left uncovered by a strict research approach.

3.2. CrossFire Introduces Data Export and Linking

CrossFire also introduced features that gave the user flexibility in processing the results of a search. The user interface had a query level and a display level. A query was communicated to the server, the server probed the data and generated a hitset that was interpreted and delivered in the display level. This afforded a concentrated compilation of results that could be further refined either by performing a search with narrower criteria or by manipulating hitsets for their Boolean intersection. A particularly useful feature of this display level was the ability to download the delivered content. The export module of CrossFire Commander included a wizard to assist in defining export settings and supported Microsoft Excel, Microsoft Word (*12*), ASCII, SDfile, and RDfile formats (*13*). Once users had a workable hitset, they were not tied to the CrossFire system in order to read and examine results. They could walk away with an electronic file of their relevant content and process it for their own use as they saw fit.

Equally important to a search is being able to expand the relevance of a hitset. This can be done, of course, by reformulating a query, but it is more efficient to simply browse related topics connected to the content of a hitset. Hyperlinks afford that functionality and CrossFire made heavy use of them. Through hyperlinking, the user had “point & click” access to additional information contained in the database, expanding relevance within the context of the search results. In this way, the hitset resulting from a structure-based search included hyperlinks to data called up from a reaction or a bibliographic search. This hyperlinking was extraordinary for its time. As Lawson mentions in his discussion of the CrossFire revolution (*8*), in its earliest days CrossFire had more hyperlinks than the entire internet.

The scope of data linking grew as CrossFire evolved. In the late 1990s, mechanisms were incorporated to allow linking out of the system to the original documents referenced in the database so the user could directly view the electronic file. As a next step, the reverse process was implemented (i.e. accessing the database content through a link embedded in an electronic article) so a user reading a chemistry article could call up data on a particular compound or reaction by clicking on a section of text. A combination of the two processes therefore created a reiterative cycle that led to expanded relevance (*8*).

By the beginning of the millennium, CrossFire was a successful and growing business. Although it consisted of two separate databases—CrossFire Beilstein and CrossFire Gmelin—the user interface software probed both repositories with a single query. Missing however was patent content. To close that gap, the Patent Chemistry Database was launched in 2005, which included data excerpted from English language patent publications back to 1976. This opened a vast resource of information (especially for reactions) that is often left untapped because of the ambiguity of patent language and/or the preference for peer-reviewed sources (14). With the addition of patent information, chemists had access to a comprehensive information system with the first manifestations of next-generation research solutions, where ease-of-use plays a central role in development goals. Nevertheless, installation of CrossFire Commander on all workstations was labor-intensive, the user interface was still off-putting for younger chemists, and there was a noticeable time lag between the publication of data and the actual availability of the updated data for the user. Technological advances could eliminate these drawbacks and a revamp of the information system addressing these concerns would lead to the creation and launch of Reaxys.

4. Reaxys

The use of graphical user interfaces and the advent of Web 2.0 features (15) moved the user into the forefront of design changes to information systems. With this transition, user expectations also changed, especially among younger generations. Already featuring mechanisms that facilitated direct access to expanded information relevance (see Section 3.2.), CrossFire was the ideal precursor for an updated system that would respond to the demands of customers. In 2007, feedback gathered from CrossFire users set goals for the extensive refurbishment of the system. The result of two years of development at Elsevier Information Systems GmbH was the first version of Reaxys, launched in 2009. Reaxys matured over the following two years, underwent significant upgrades in 2013, and continues to adapt to the changing needs of a diversifying user base.

4.1. Building on Strengths in Data Linking and Access

The development of Reaxys emphasized two strategic priorities: merging the three databases of CrossFire covering organic, inorganic and organometallic data from journal and patent literature, and constructing a new, accessible user interface. Ultimately, both priorities reinforced existing strengths of CrossFire in data linking that made the database merger axiomatic, and in system access that primed the use of a web portal. As we will see in Section 4.2., CrossFire's strength in structure/reaction content would also contribute to streamlining the user interface.

Given the ample overlap of data fields, merging the databases into one coherent system was an obvious next step in the management of the underlying data compendia. The upshot of the merger was that it placed information from all three databases into the same systematic format so the user could juxtapose,

combine and compare information extracted from journals and patents, thereby increasing productivity and promoting creativity.

Reaxys responded to customer demand for reduced upkeep by providing a web-based interface requiring minimal to no maintenance on the customer side. This was the logical sequel to the client/server architecture of CrossFire. With this new online portal, Reaxys also offered an unlimited number of users in the system and an increased frequency of database updates. Thus, a customer organization with a subscription to Reaxys could grant users access (controlled via IP address or a user account and password) to an enhanced information system covering both journal articles and patents that underwent biweekly updates (compared to quarterly updates of CrossFire).

4.2. Building on Strengths in Structure and Reaction Data

Coupled to the logistics of unhindered access was the provision of an unencumbered user experience. The gradual addition of functions to CrossFire had led to a user interface that was cluttered and confusing. Feedback from users with a range of experience using CrossFire called for a simpler interface without loss of search power. In response, the redesigned user interface was built around the thought and work processes of the CrossFire core user, the preparative chemist. The new user interface was streamlined by merging redundant features and aligning the location of functionalities with a search workflow that begins with a structure or a reaction. According to a review in 2009 (14), the straightforward interface design invited the chemist to start a query directly, without previous training or use of instructions.

Elsevier Information Systems also launched new functionalities that enhanced the strength of Reaxys in structure and reaction searches, and these functions have been preserved through all subsequent updates of Reaxys. Structure searches in Reaxys are facilitated by flexible query formulation. The structure search input form is compatible with multiple structure sketchers for users who wish to draw the structure to be examined. Alternatively, Reaxys can also generate the structure from an entered compound name, saving time and mitigating potential drawing errors. A complementary spectrum of tools allows chemists to build substitution counts into a structure to emphasize the essential components of the structure to be searched, explore the impact of substitutions at designated sites, or obtain an expanded hitset based on similarities of structure or classification. For inorganic and material chemistry, the search engine of Reaxys enables queries using molecular formulas, as well as searches for ligands attached to central metal atoms and for alloys. Substructure filtering on the hitset is also directly available, one of many functions designed to sharpen the focus without reformulating a query. Together, these search functions constitute a toolbox that enables both novice and expert to maximize the utility of the database.

In contrast to CrossFire, reaction searches in Reaxys have matured beyond a simple listing of reactions that yield a given product. Details indexed for over 30 million reactions—from yield to solvent, time and temperature—can be used to formulate a query that results in a list of relevant reactions with experimental procedure and links to source documents. Furthermore, the indexing of reaction

data extracts enough information from a given source for a user to assess whether the purchase of the source article is necessary or if the synthesis can be conducted from the information provided by Reaxys.

The utility of reaction searches in Reaxys is particularly apparent in Synthesis Planner (see Figure 2), where the linking of information from disparate sources into a graphical display delivers an actionable synthetic pathway. The preparation of a substance in Reaxys can be planned step-by-step using reactions or portions of reactions extracted from journals, books, or patents. The user can create the synthesis plan manually by selecting steps from a list of reaction options that Reaxys provides, each with yield, conditions, and reference (14). Alternatively, the Autoplan feature, incorporated in 2013, retrieves multiple complete synthesis plans for selection and subsequent amendment. Each step in these complete synthesis plans is also provided with one or more references.

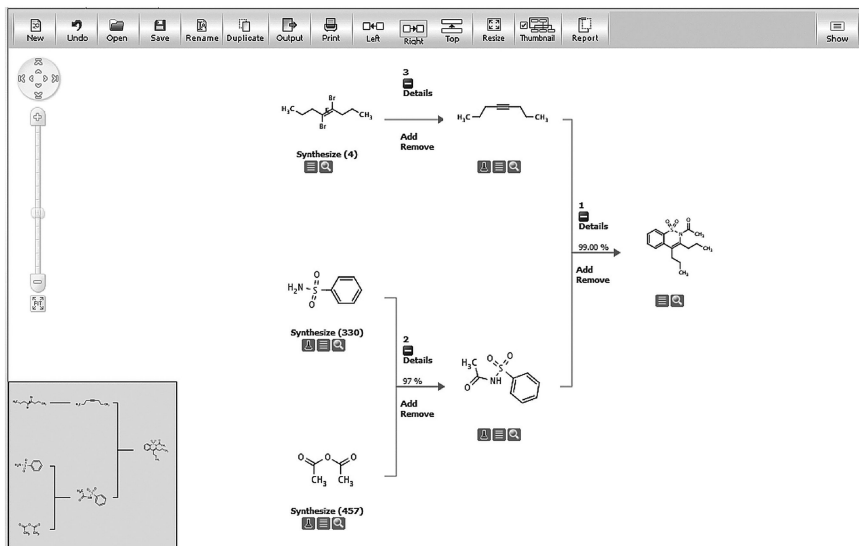


Figure 2. Screenshot of a synthesis generated in Synthesis Planner.

As shown in Figure 2, each step of that reaction is illustrated in the plan with reactants, isolated intermediates, and products displayed as structures. Links from each molecule in the reaction give access to chemical name, synonyms, InChIKey (16), CAS Registry Number (17), references with links to original documents, spectral and other characterizing data with excerpts of the relevant original text. All experimental details of each reaction in the plan are summarized in table form making the user aware of the extent and variability in published results and synthesis plans for each reactant molecule can also be iteratively expanded to extract data from further sources, including collaboration with other vendors such as PubChem (18), eMolecules, ChemACX, and Accelrys ACD (19). As of November 2013, Reaxys had over 22 million substance entries, PubChem over

47.5 million and eMolecules over 6 million. The deduplicated sum of entries amounts to approximately 55 million substance entries.

The extensive linking of data and functions both within the Reaxys environment and with external sources creates a whole that delivers more impactful results than each component element could on its own. The Synthesis Planner is a good example of this principle and, as we will see in Section 5, the future vision for Reaxys builds upon it.

4.3. Expanding Coverage Scope

Enhanced functionality in structure and reaction searches differentiated Reaxys from other chemistry information solutions, but embedded in the database (and perhaps shadowed by the emphasis on structure-based research) was an untapped collection of facts in over 400 extracted data fields for other disciplines. With time, customer demand for expanded content and improved usability for researchers other than synthetic chemists indicated the added value of this data richness. An overhaul in 2013 responded by bringing to the foreground the full extent of the Reaxys database and by augmenting the system's coverage of periodicals. These improvements achieved a characteristic balance between specialized research functionalities and comprehensive content attractive to an expanded audience.

The coverage scope of Reaxys was expanded to include data from 16,000 source periodicals in parallel to the existing core journal articles and patents. Core data is still extracted and processed manually from a selected set of 400 sources using the strict data structure at the backbone of Reaxys, but a far-reaching collection of relevant information from journals, books, conference proceedings, abstracts, and editorials was introduced to complement the core data and cover fields as diverse as biology, physiology, engineering, pharmacology, and environmental science.

The impact of this amplified coverage is two-fold. On the one hand, it broadens the relevance of this research tool within an increasingly interdisciplinary scientific arena: chemistry never was an isolated endeavor, but its ubiquitous presence throughout the natural sciences is stronger than ever given its central role in academic research and in the biotechnology and biomedical industries. In that sense, the expanded Reaxys mirrors current scientific interest and can therefore serve as a platform where multiple disciplines intersect. On the other hand, this interdisciplinarity requires extended reporting and communication features to enable information to be shared among colleagues and between platforms. The Reaxys user can compile details from multiple result views into a single report, annotate the data, and then save the report or share it with colleagues via email.

One logical consequence of expanding the coverage of Reaxys was the need to ensure that the increased volume and diversity of data to be analyzed and abstracted did not diminish the data quality and detail expected of Reaxys. To accomplish this, a highly efficient production system was established, relying on both automatic indexing and computer-assisted manual data excerption and equipped with reiterative quality assurance mechanisms. The focus of the abstracting process remains in the chemistry space, regardless of the source

topic or emphasis. In a first step, documents are automatically tagged for data matching strict relevance criteria consistent with the underlying scientific model of Reaxys: compound data, related properties, facts, preparations and reactions, and bibliographic data. Then the data is passed to a special tool for manual curation, the interactive Excerption Interface (iEI).

As illustrated in Figure 3, the use of iEI grants the abstractor a dual view of both the source, in a reading pane containing tagged information, and the target, in a working pane that displays data fields to be filled and a navigation of the Reaxys taxonomy. Thus, the abstractor can analyze and enter relevant data in real-time. This work is supported by features that facilitate the recognition of relevant information, its correct placement in the excerpted data, and an ongoing quality check via integrated tools that compare entered data against known compounds and locate missing or inconsistent data. These checks during the excerption process are the first line of quality control. Subsequently, two independent and manual quality controls take place on the excerpted data prior to integration into the database. Finally, Reaxys production conducts a quality control of the database integrity prior to loading on the public server.

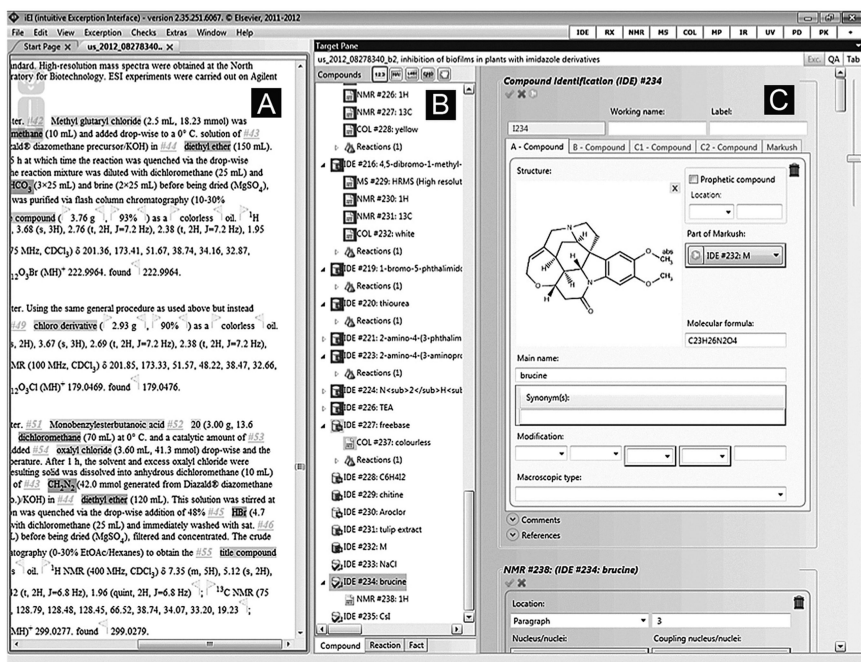


Figure 3. The interactive Excerption Interface (iEI), the specialized tool that supports manual extraction of data for Reaxys, includes (A) a reading pane with tagged information (B) a navigation window with the Reaxys taxonomy and (C) a working pane for data entry. Courtesy of Elsevier Information Systems GmbH and Reed Elsevier Properties SA, 2013.

A second logical consequence of the expanded content of Reaxys is an increased output. To process this increased output, the user requires flexible tools to expand and narrow down queries, as well as filter and sort hitsets. Additionally, facilitating meaningful searches requires that query formulation corresponds to a natural way of asking the underlying question, which is a vital factor in helping the user to focus on relevance in the answer set (as discussed below).

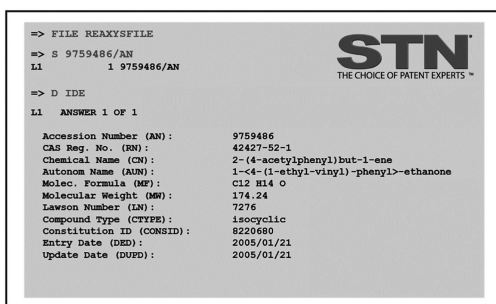
Therefore, the start page design of Reaxys (Figure 4) offers different query themes based on how the underlying question is to be stated: would you like to find information based on a structure or a reaction, based on the name, formula or other compound identifier, or based on standard bibliographic data? The resulting search form (regardless of query theme) can be customized further by adding search fields pertaining to reaction data, physical and spectral properties, pharmacological data, and natural product of origin. Finally, the answer set can be sorted to entries with a maximum relevance according to various criteria (e.g. year of publication).

Taking the example of the trans stilbene derivative “resveratrol” (see Figure 5), starting with a structural search qualified by the trans stereochemistry will be vital if the user is interested in synthetic routes to this molecule or finding physiological or other numeric data. On the other hand, if the interest centers on publications that have this molecule as a central topic for other reasons (e.g. patent aspects), a qualified literature search could be more appropriate. Both of these answer sets will be highly focused. If however, the user chooses to search for a comprehensive set of data based on reports of resveratrol (or its cis stereoisomer) in unspecified context, the answer set would be dramatically increased (over 200 molecular entries exist, including mixtures, salts, and isotopically labeled substances mentioned in over 1000 documents); then, the same filtering options can be applied to rapidly come to a manageable focus (e.g. in Figure 5, the number of literature references was used).

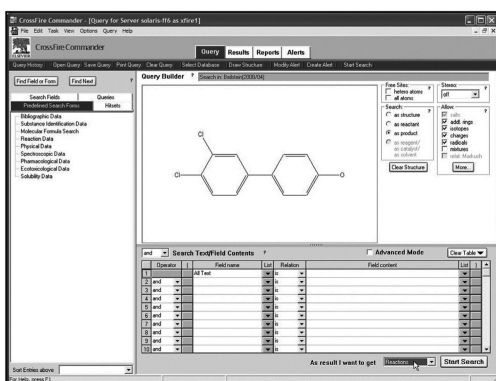
An alternative search on the name “resveratrol” in the context of a literature search would yield an answer set consisting of 71 citations displayed in tabular or grid format with links to full abstract and text where available, and graphical display of all the substances that appear in each citation. In short, query formulation in Reaxys addresses different search approaches and provides the flexibility in filtering to deal with an intuitive search by the novice user or a complex, multifaceted search by the expert.

4.4. User-Centered, Rather than Technology-Centered Development

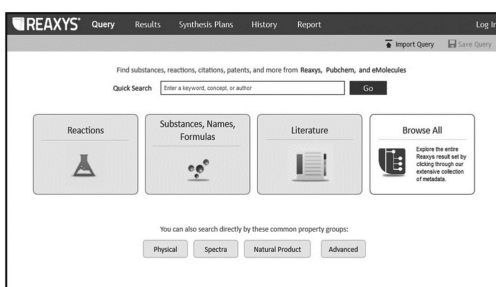
Reaxys inherited from CrossFire considerable sophisticated technology. Nevertheless, each development since the inception of Reaxys has placed the user (and not the technology) at the center of the Reaxys experience. The first five years of Reaxys have been a response to user behavior and needs. Customer feedback was gathered through surveys, market research and by working with key collaboration partners from both academia and industry. The profiles collected, however, were only snapshots of a rapidly changing and increasingly sophisticated user base. The next five to ten years of developments will need to step beyond responding and instead anticipate the information needs of users and thus, the user will need to be an integral component of the evolution of Reaxys.



1989 – STN



1996 – CrossFire Commander



2014 – Reaxys

- Text-based
 - Knowledge of database structure and host-specific commands required
 - Not all query forms or results readily accessible
-
- Graphical client/server-based system
 - Improved access via input forms
 - Structure/reaction searches combined with factual queries
 - Cluttered display of functions
-
- Subject oriented, customizable Web-based user interface
 - Concept search enabled
 - Supporting all types of chemical information needs
 - One-click access to advanced query forms
 - Multiple source databases

Copyright © 2013 Reed Elsevier Properties SA.

Figure 4. Evolution of the user interface from Beilstein Database Online (STN) to Reaxys in 2014 Courtesy of Elsevier Information Systems GmbH and Reed Elsevier Properties SA, 2013.

207 substances out of 1028 citations

Filter by:

- Sub-structure
- Molecular Weight
- Number of Fragments
- Physical Data
- Spectroscopic Data
- Bioactivity
- Natural Product
- Availability
- Availability in other DBs
- Document Type
- Authors
- Patent Assignee
- Journal Title
- Publication Year

Substances (Grid) Substances (Table) Citations

Limit to Exclude Output Print Zoom in Zoom out

Sort by No of References

Structure	Structure/Compound Data	N° of preparations All Preps All Reactions	Available Data	N° of ref.
<p>Synthesize Show Details</p>	<p>Chemical Name: (E)-1-(3,5-dihydroxyphenyl)-2-(4-hydroxyphenyl)-ethene</p> <p>Reaxys Registry Number: 1912434</p> <p>CAS Registry Number: 501-36-0</p> <p>Type of Substance: isocyclic</p> <p>Molecular Formula: C₁₄H₁₂O₃</p> <p>Linear Structure Formula: C₁₄H₁₂O₃</p> <p>Molecular Weight: 228.247</p> <p>InChI Key: UW8DSAVLPHMSZ-OWOJBTEDSA-N</p>	<p>130 prep out of 315 reactions.</p>	<p>Identification Physical Data (210) Spectra (260) Bioactivity/Ecotox (2991) Use/Application (961) Natural Product (123)</p>	933

Figure 5. Hitset from a structure search on resveratrol displayed in tabular form, organized by substance and sorted by number of references.

Therefore, four years ago, Reed Elsevier Properties SA established the Reaxys Club. Yearly, young chemists from around the world submit their work for evaluation by a jury of independent experts and the three most original and innovative researchers are awarded the Reaxys PhD Prize. The winners and 42 finalists are invited to join the Reaxys Club. This international network of the brightest chemical minds is a venue for collaboration, creative brainstorming, idea exchange, and an incubation ground for advances in chemistry. For the Reaxys team, the Club is a vital connection to the current and future user and an “in the trenches” learning forum. Honest feedback about new products and their features from members of this Club is an exciting, albeit sobering, steering force for development efforts. Insight from this open exchange with young users is the basis for the steps taken toward integrating Reaxys into the natural research process of the chemist. This brings us to the final step in this article, bringing it all together. In Section 5 we will move into the Reaxys of 2014 and beyond, where this insight is already coming into play.

5. Steps toward Unobstructed Information Access

As mentioned in Section 1, chemical information sources have always implicitly required users to “understand” the underlying relationships of the system they are using; in contrast, a mirror-image approach could involve the system “understanding” the underlying intention of the user it is serving. Consider the following example scenario: you would like to know what movie you can watch after work and where you can eat dinner thereafter. You search movie theater programs, check their location in online maps, look for restaurants near the theater and read reviews, and consult your traffic-App to ensure you can make it to the theater on time. You visit at least five different websites and scan through several paragraphs of text to come up with the right answer to a rather simple question. In essence, you conduct five different searches and extract relevant information from each hitset. The final answer lies at the intersection of a diverse

spectrum of information sources and types, and identifying that intersection still lies in the hands of the user because only the user can assign meaning to the uncovered data. If, however, information in disparate sources were indexed and networked through relationships, the outcome might be a simple answer with a movie and restaurant choice based on previously noted preferences, a route plan to minimize your time in transit, and maybe a reminder that you have a late meeting scheduled in your calendar. In addition to uncovering the relevant data, the search engine interprets the objective of the query and processes the data to identify the answer intersection (movie and restaurant) and even anticipate the user's need for additional relevant information (a scheduling conflict).

This is the vision of the semantic web (4) and the user experience that Reaxys aims to create as an unobtrusive portal to the richness of its data, without making the user interface complicated. This means that text searches, despite the ambiguity inherent to text query formulations, must become more natural and deliver only context-relevant results. This also means that the user experience must be fluid, adapting to workflows, interpreting the context and underlying meaning of a query, and anticipating additional information that is relevant to the search without explicit commands.

First steps towards this vision have already been taken in three different areas, building on the strengths of the information system:

- indexing and database taxonomies have been rebuilt for better text-based searches
- with its flexible data model, Reaxys has been integrated directly into user environments
- finally, building on extensive intra- and extramural data linking, the taxonomy of Reaxys has been connected to that of other information products; a query in Reaxys can extract relevant data from other scientific domains and thus support interdisciplinary workflows.

These three aspects will be discussed in the following three Sections.

5.1. Indexing and Taxonomy beyond Equivalence and Hierarchical Relationships

In general, data structure at its simplest organizes information in such a way that a particular entry can be found based on a set of rules. The logic underlying the rules generally reflects the functional objective of the data, i.e. how the data will be used. Thus, for example, a dictionary provides word definitions that are easily found because of the strict alphabetical arrangement of the entries. More complex data arrangements invoke subclass relationships, where entries are assigned to supracategorical terms and these, in turn, are assigned to overarching groupings and so on. The result is the hierarchical structure of a taxonomy, which accommodates relationships of synonymy, inclusiveness, and ranking. Users of Reaxys however, look for a much broader spectrum of relationships in data. Therefore, most information research workflows entail multiple searches and identification of the intersection between resulting hitsets which best reflects the

context of the driving research question. As in the “movie and dinner” example above, what is obtrusive about this process (and the source of ambiguity in the search), is that users must translate their question into component sub-queries which can be understood by the search engine; that is, users must adapt to the technology, and not the other way around.

For the envisioned fluid user experience where query construction is intuitive, the interface must accurately translate the query of a user regardless of formulation, and the underlying algorithms must understand complex relationships between database entries to generate a hitset that reflects the way the user thinks. Better query translation and data relationship interpretation enable the system to approximate the answer intersection of the query, just as the movie and restaurant matched for time and preference in our example. Both functions require that data indexing and classification be broadened to include semantic relationships that specify how two or more entities are related.

As a simple illustration of this, (analogous to the controlled vocabulary used by F.K. Beilstein, as kindly suggested by one reviewer) consider the following determination of the relationship between three entities embedded in a single phrase. The phrase “prep of 4-nitro-2-alkylphenols from 2-alkylanisoles in aq. HNO₃” should present no problems for an algorithm armed with a good name-to-structure translator (numerous are currently available) and combined with a minimal vocabulary covering prepositions such as “of”, “from”, “by”, “in”, “with” and nouns such as “prep” and “treatment” and corresponding synonyms. Such words would then **only** be interpreted (and therefore activated in query generation) in the presence of structurally translatable nomenclature terms at the corresponding positions in the phrase. The end result would be a graphical generic reaction query (reactant and product structure entities correlated) combined with a reagent specification. In addition, beyond such basic relationships and others like “is synonymous to”, or “is subclass of”, data can be networked with indexing that reflects cause and effect (“is an adverse reaction to”), correlation (“is present when”), compound uses (“is treatment for”), biochemical pathways, competing hypotheses (“contradicts”), and more. This conceptual classification expands data organization to a highly networked, polyhierarchical structure.

Therefore, at the beginning of 2014, data in Reaxys were indexed with an added set of relevance designators that equate to a chemical dictionary. That is, an entry is additionally tagged with one or more constructs that reflect a chemical meaning and thus, define the context in which the queried words or phrase appear in the source document. Furthermore, tagged compound names in the source are automatically translated into searchable compound structures. This indexing, supported by a chemical dictionary and searchable structures, establishes meaningful connections between previously unstructured data and allows users to filter text-based search results for hits that appear in a context relevant to their research question.

A second layer of semantic relations is built into query processing to accommodate the use of natural language for query formulations. As an alternative to the customizable search forms, Reaxys now has a simple search field where the user can enter search criteria in free form (Figure 4, 2014). Thus, users who do not wish to use the query themes (reactions, substances or literature) can

perform a quick search into the richness of the database using directly the natural query formulation they had in mind. Similar to some consumer search engines, the search algorithms underlying this entry field enable concept searching. Under this function, a phrase such as “average melting point of ethidium bromide” is no longer simply a string of words that are searched independently, but rather a set of keywords linked by operators that assign meaning to the entire phrase. The underlying search and ranking algorithms interpret the phrase and deliver a hitset ranked by relevance to the interpreted meaning.

5.2. Integral Component of the User’s Environment

An unobtrusive research tool is one that is present exactly where and when need arises. In the case of scientists in an organization, this often means when they are working within internal knowledge management systems. The data model of Reaxys has the flexibility to readily integrate with such systems, either at a basic level through an application programming interface (API), or more extensively, where Reaxys itself is adapted to organize customer-generated data and make them searchable.

At a first level of integration (i.e. one that is designed with all Reaxys users in mind), accessing the rich repository of Reaxys is possible directly from several commercially available electronic lab notebooks (ELN). Under this access model, an API embedded in the ELN serves as a one-click portal into Reaxys. The user initiates the query within the ELN, and then enters Reaxys to refine the search and select the output of interest. To complete the circuit, the chosen data are imported into the ELN.

Integration with third-party or company in-house tools and search systems is often a first step towards a deeper integration within an organization-specific data and knowledge management platform. Here, an API embedded in the platform calls up data from Reaxys by generating a query in XML that retrieves data from Reaxys and returns them to the user’s platform. Such technology supports virtual screenings of large sets of compounds or patent numbers, where the information extracted from Reaxys is incorporated into the analysis. Another form of this integration level is an API incorporated into a federated search system where Reaxys is probed for information availability. The user then knows to conduct a search directly in Reaxys.

A third, highly customized, integration level transposes in-house proprietary content into a customer-specific version of the Reaxys platform. The system runs on the customer’s infrastructure and conducts a federated search in the Reaxys database and one or more databases containing the customer’s internal and experimental content. The search results are delivered on two separate tabs but are highly connected via crosslinking. In that sense, the system offers the benefit of the single platform and interconnected data of a warehousing model, as well as the independence of databases with their nuances and specifics afforded by a federated model. Such an integration was finalized for Roche in 2012 (20). The knowledge holdings from a long history of research and development at Roche were organized and indexed to be discoverable through the Reaxys user interface. New internal data recorded in ELNs is automatically converted

into RDfiles and XML files that are used for a nightly update of the integrated customer database. Feedback from Roche indicates that the integrated system is easier to maintain than having to manage multiple platforms, has improved the searchability of internal content, especially that contained within ELNs, and has increased the productivity and precision of scientists. While content integration is a tailored product, each customization generates experience and more streamlined mechanisms for adapting content and data models, which will lead to easier integration of Reaxys into the user environment.

5.3. Interoperability

As the user base of Reaxys diversifies, so does the definition of a successful search because information needs differ depending on user specialty and application of the uncovered data. To deliver a successful and productive search experience, query and data representation in a system must match the mental models of users, and these models can be very diverse. Creating an all-inclusive, “one size fits all” system that attempts to meet all interests would be inefficient and cumbersome, but interconnecting multiple systems, each focused on a given interest would magnify the search power, so that a query in one system could also extract meaningful data from another.

Consider the following example: Reaxys is a research tool for chemists, but pharmacologists are also interested in chemistry data, only in a data model relevant to the drug discovery workflow. Reaxys Medicinal Chemistry focuses on linking structures, bioactivity data and pharmacological targets of substances. Matching data models commonly used in drug development, the search input form and database taxonomy emphasize biological and pharmacological experimental data and resulting hitsets are displayed as a heat map of a substance-target matrix or other user-defined axes (Figure 6). The taxonomies of Reaxys and Reaxys Medicinal Chemistry are coordinated so that a user with access to the two products performs a query in Reaxys or in Medicinal Chemistry and receives relevant information from both. In this way, the search power is magnified because a greater informational space is covered (chemistry, biology and pharmacology), but examination of that space is streamlined because the coordinated taxonomies guarantee relevance of the query results. In line with our example where movie and restaurant were identified by Web 3.0 processing, this interoperability condenses multiple searches into one targeted query that uses the connection between taxonomies to identify the data that are relevant in each database.

Ultimately the addition of more interoperable taxonomies will generate a network of interactive information solutions that cover an extensive informational landscape. At this point, this network of products is most developed in the Elsevier Life Sciences Solutions (21) but may expand into other areas as well. With content augmented by the integration of third party databases (e.g. PubChem, eMolecules) and links to the abstract and citation database Scopus and the electronic literature database ScienceDirect (22), Reaxys is part of a growing ecosystem of interactive products that make available interdisciplinary information at any point within the chemist’s research workflow.

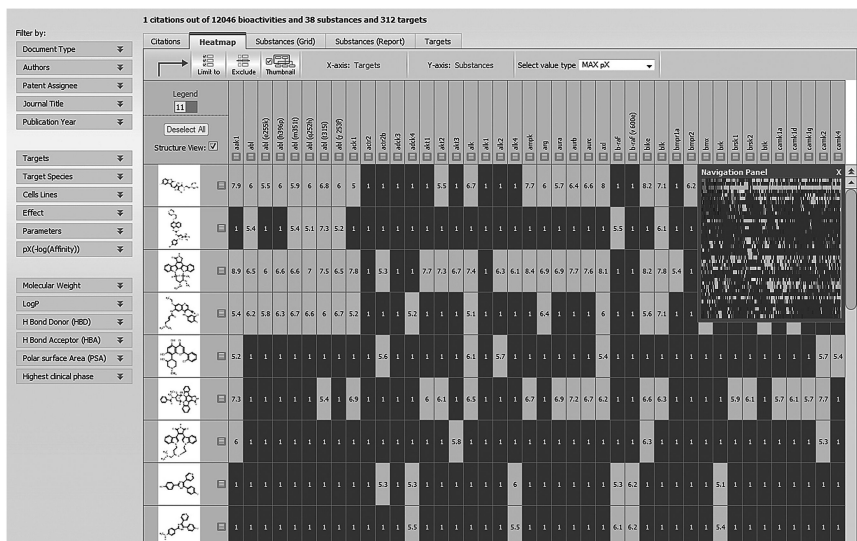


Figure 6. Results from a search in Reaxys Medicinal Chemistry displayed as a heat map of a substance-target matrix.

In summary, the path toward an information system where intelligent, behind-the-scenes engines create a natural “conversation” with the user is doubtless still very long and extremely difficult. Nevertheless, the history of Reaxys and its predecessors has been an accumulation of advances that have, at critical moments, driven change in chemical information science. The evolution over the past five years has therefore set the stage for Reaxys to continue to play an important role in the ongoing innovation required to deal with the challenges of the 21st century in this field.

Acknowledgments

The authors would like to thank Timothy Hctor, Sebastian Radestock, and Anja Brunner for their valuable contributions to the content and preparation of this manuscript.

References

1. www.reaxys.com (accessed March 14, 2014). Reaxys, Reaxys Medicinal Chemistry and the Reaxys trademark are owned and protected by Reed Elsevier Properties SA and used under license.
2. Richter, F. How Beilstein Is Made. *J. Chem. Educ.* **1938**, *15*, 310 (<http://pubs.acs.org/doi/pdf/10.1021/ed015p310>).
3. CrossFire is a trademark of Elsevier Information Systems GmbH.
4. The Semantic Web (also called Web 3.0) is a term first introduced by Berners-Lee, Hendler and Lassila in a Scientific American article to describe

an expected evolution of the World Wide Web (The Semantic Web - Scientific American <http://www.scientificamerican.com/article/the-semantic-web/> (accessed December 18, 2013). The World Wide Web Consortium (W3C; www.w3.org/standards/semanticweb/) now spearheads a collaborative movement that aims to transform the internet into a “web of data” where connections between data carry meaning (semantic content) that can be processed by machines.

5. *Gmelin Handbook of Inorganic and Organometallic Chemistry*, 8th ed.; Gmelin Institute for Inorganic Chemistry of the Max-Planck-Society for the Advancement of Science and Springer Verlag: Berlin, 1998 (<http://library.buffalo.edu/asl/guides/Gmelin-Complete-Catalog.pdf> (accessed March 14, 2014)).
6. Luckenbach, R. The Beilstein Handbook of Organic Chemistry: The First Hundred Years. *J. Chem. Inf. Model.* **1981**, *21*, 82–83 (<http://pubs.acs.org/doi/pdf/10.1021/ci00030a006>).
7. Weissbach, O.; Hoffmann, H. M. R. *The Beilstein Guide: A Manual for the Use of Beilstein's Handbuch Der Organischen Chemie*; Springer Verlag: New York, 1976.
8. Lawson, A. J. The Beilstein Database. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH Verlag GmbH: Weinheim, Germany, 2003; pp 608–628.
9. Meehan, P.; Schofield, H. CrossFire: A Structural Revolution for Chemists. *Online Inf. Rev.* **2001**, *25*, 241–249.
10. STN International, www.stn-international.com (accessed December 18th 2013), and Dialog www.dialog.com (accessed December 18th 2013; now owned by ProQuest), are online database services offering access to hundreds of scientific and technology databases. STN is a registered trademark of the American Chemical Society; Dialog is a registered trademark of Dialog, LLC.
11. Lawson, A. J.; Swienty-Busch, J. Crossfire: A Comparison of Online and In-House Performance. *Online Inf.* **1993**, 378.
12. Microsoft Excel and Microsoft Word are programs of the Microsoft Office Suite. Excel and Microsoft are trademarks of Microsoft Corporation.
13. Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Model.* **1992**, *32*, 244–255. For a description of the Accelrys CTFE format please see <http://download.accelrys.com/freeware/ctfile-formats/ctfile-formats.zip> (accessed March 10, 2014).
14. Goodman, J. Computer Software Review: Reaxys. *J. Chem. Inf. Model.* **2009**, *49*, 2897–2898.
15. Web 2.0 is a term coined in 1999 by Darcy DiNucci (but first popularized in 2004) to describe the cumulative changes to the World Wide Web that enable enrichment of websites through user-generated content.
16. Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI - the Worldwide Chemical Structure Identifier Standard. *J. Cheminform.* **2013**, *5*, 7.

17. For more details about the CAS Registry Number, please refer to <http://www.cas.org/content/chemical-substances/faqs> (accessed December 18, 2013)
18. Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. PubChem: Integrated Platform of Small Molecules and Biological Activities. *Annu. Rep. Comput. Chem.* **2008**, *4*, 217–241.
19. eMolecules is a search engine for chemical compounds with an extensive database of molecules from commercial providers. eMolecules, Inc., 11025 North Torrey Pines Road, La Jolla, CA 92037, USA. Accelrys Available Chemicals Directory (ACD) is a database for chemical sourcing. Accelrys, Inc., 5005 Wateridge Vista Drive, San Diego, CA 92121, USA. CambridgeSoft Available Chemicals Exchange (ChemACX) is a database for chemical sourcing covering suppliers worldwide. PerkinElmer Informatics, 940 Winter St., Waltham, MA 02451, USA.
20. Agnetti, F.; Bensch, M.; Biller, H.; Blapp, M.; Cheikh, B.; Blanke, G.; Degen, J.; Dienon, B.; Doerner, T.; Doernen, G.; Farshchian, F.; Gotzeina, W.; Hilty, P.; Horstmoeller, R.; Jeker, T.; Jones, B.; Kappler, M.; Momin, A.; Regoli, A.; Ribaud, D.; Starck, B.; Stoffler, D.; Weymann, K.; Udupa, P. Intuitive and Integrated Browsing of Reactions, Structures, and Citations: The Roche Experience. Presented at the 245th National Meeting of the American Chemical Society, New Orleans, LA, April 7–11, 2013.
21. Elsevier Life Science Solutions is a collection of online information systems covering the life sciences and their application in industry. For more details, please see www.elsevier.com/online-tools/corporate/life-science-solutions (accessed December 18, 2013).
22. For more details about Scopus, please see www.elsevier.com/online-tools/scopus (accessed December 18, 2013). For more details about ScienceDirect, please see www.sciencedirect.com (accessed December 18, 2013). Scopus and ScienceDirect are registered trademarks of Elsevier B.V.

Chapter 9

Back to the Future: CAS and the Shape of Chemical Information To Come

Roger J. Schenck* and Kevin R. Zapiecki

**Chemical Abstracts Service, 2540 Olentangy River Road, Columbus,
Ohio 43202**

***E-mail: rschenck@cas.org**

Chemical Abstracts Service (CAS), the only organization in the world whose objective is to find, collect and organize all publicly disclosed chemistry, has been a leader in providing scientists with access to chemical information for more than 100 years. CAS relied on a group of globally situated volunteer abstractors from 1907 until the early 1990s. CAS now keeps pace with the explosion in newly disclosed chemistry with more than 500 scientists working at the CAS headquarters in Columbus, Ohio, who are supported in turn by that same number of scientists working in locations around the world. CAS has designed computer applications both for database-building efforts and service delivery. In 1984, STN was developed for professional searchers to access scientific and technical databases. With the introduction of SciFinder in 1995, CAS developed the first chemical information analysis tool specifically targeted to help chemists working in the lab. Since then, CAS has leveraged rapid changes in technology and evolving sources of disclosed chemistry, to fulfill its mission to provide the world's best digital research environment to search, retrieve, analyze and link chemical information. This chapter describes how CAS has adapted to the phenomenal growth in published research to continuously support scientific discoveries and will close with some thoughts about the future of chemical information.

Overview of CAS

In 1907, E. J. Crane established the importance of indexes, not just abstracts, as part of *Chemical Abstracts*, starting with author and subject indexes (1). Since there was little control over nomenclature systems used in the early chemical literature, Carleton Curran and Austin Patterson of *Chemical Abstracts* devised a systematic method of naming substances in 1916 (2). They surveyed organic chemical literature for common practices, established an order of precedence for chemical functionality and instituted the use of inverted index names. Inverted names became popular as a way to group similar classes of compounds in an alphabetical printed index (3). *Chemical Abstracts* came to be recognized as a leader for chemical substance nomenclature development. In 1937, *Chemical Abstracts* published its one-millionth abstract (4).

Around the time of the Seventh Collective period (1962-1966) *Chemical Abstracts* staff was struggling to keep pace with substances reported in the chemical literature (5). Before 1965, structures were hand drawn and the substances were subsequently named. Manual comparisons were done to determine if the incoming substance had been previously indexed. At the same time computer technology was emerging, and Chemical Abstracts Service research staff brought computers to bear on the problem. The CAS Chemical Registry System was introduced in 1965 as an internal production system that replaced the redundant and very expensive task of naming known compounds. Using a unique CAS Registry Number to identify each chemical substance, the system proved to be a future benefit to chemical research, health and safety information, and the communication of chemical information. There are now more than 85 million (April 2014) (6) organic and inorganic substances in CAS REGISTRY, which makes it the world's largest substance database.

Introduced in 1980, CAS ONLINE made it possible for users (primarily information specialists) to search the CAS REGISTRY database (7). Using a command language, users communicated their search strategies to the system. Users with a specific model of an intelligent graphics terminal could select structure features from a menu and then assemble them on the terminal monitor using a graphics tablet and stylus. These terminals could display answers with consistently drawn structure diagrams.

CAS content speeds the pace of scientific discovery through two platforms: STN and SciFinder. In 1983, CAS partnered with FIZ Karlsruhe (in Germany) and was represented in Japan by The Japan Science and Technology Agency (JST) to form an international online network. STN, the Scientific and Technical Information Network, was launched the next year. STN made databases accessible through distributed processing on a global scale. Initially, only CAS databases and Physics Briefs were accessible. Over time, STN grew to include many scientific databases from a range of information providers. STN databases are uniquely integrated so researchers can consult multiple databases with a single query. A new web-based platform, with a project-oriented workflow, and enhanced search power, precision and usability, was recently released and continues to be developed.

CAS introduced SciFinder in 1995 as a research tool to give scientists direct access to CAS databases with no prerequisite to learn a command language (8). With its intuitive, graphical interface, SciFinder simplified the exploration of the world's scientific literature, patents and substance information, making this activity part of the process for scientific research.

CAS recognized the possibilities of the Internet to speed and simplify access to original journal articles and patents. CAS Full Text Options (originally called ChemPort) was introduced to CAS and STN electronic services in 1997. Today it provides access to full-text journal articles and patents from more than 7,400 electronic journals from nearly 360 participating publishers (9). CAS Full Text Options also provides links to electronic patent documents from full-text patents from five offices: USPTO (U.S. Patent and Trademark Office), Espacenet (European Patent Office), SIPO (State Intellectual Property Office of the P.R.C.), JPO (Japanese Patent Office) and KIPRIS (Korea Intellectual Property Rights Information Service).

Addressing the Information Needs of Scientists

In the late 1960s, with the advent of computer technologies, CAS investigated chemical information products and services beyond what was already available in CAS REGISTRY and the CA File on STN. The market drove CAS to consult chemists and information professionals to better understand their needs. Beyond the need for access to chemical substance information and the literature from which those substances were selected, there was a clear opportunity for CAS to provide scientists with much more targeted information. The desire for a collection of chemical reactions that included both standard, trusted reactions as well as new and novel synthetic techniques was front and center among customers interviewed. This was the beginning of a rich suite of additional chemical information currently available to scientists in the CAS databases.

CAS is the only organization in the world whose objective is to find, collect and organize all publicly disclosed substance information. CAS currently covers more than 10,000 active journals (10) and patents from 63 patent authorities (11). This scientific literature and these patents come from 180 countries in 50 languages (12). CAS has developed seven core databases that cover the most current scientific information: chemical substances (CAS REGISTRY), references (CAplus), Markush (MARPAT), reactions (CASREACT), chemical suppliers (CHEMCATS), regulated chemicals (CHEMLIST) and Chemical Industry Notes (CIN).

CAplus covers international journals, patents, patent families, technical reports, books, conference proceedings and dissertations from all areas of chemistry, biochemistry, chemical engineering and related sciences from 1907 to the present. There are more than 38 million records as of April 2014. In addition, over 180,000 records for pre-1907 patent and journal references are available, from sources such as the American Chemical Society (ACS), the Royal Society of Chemistry (RSC) and *Chemisches Zentralblatt* (9). Other benefits of CAplus

include abstracts of foreign language references (patent and journal) that are translated into English. CAPlus also assures patent records, from nine major patent offices worldwide, are available online within two days of the patent's issuance, and fully indexed by CAS scientists in 27 days or less from the date of issue (13).

Voicing a clear need to leave no stone unturned when searching for prior art and freedom to operate, information professionals pushed CAS to develop a database of generic structures selected from patent applications. To address this need, CAS developed MARPAT, a database of Markush structures derived from patent applications. Introduced on STN in 1990, MARPAT was designed as an extension of the information provided in the CAS REGISTRY and CAPlus databases to perform comprehensive patent substance searching.⁸ There are more than one million searchable Markush structures derived from patents covered by CAS from 1988 to the present.

CASREACT was introduced in 1988 on STN and made available in SciFinder since the launch of the product in 1995. CASREACT offers access to current reaction information found in literature covering synthetic organic chemistry. The literature includes journals and patents from 1840 to the present. There are currently more than 58 million single- and multi-step reactions, and more than 13 million synthetic preparations in SciFinder (14).

CHEMCATS, introduced on STN in 1995, is a chemical catalog database containing information about commercially available chemicals and worldwide suppliers. It contains more than 65 million commercially available products, more than 990 chemical catalogs, more than 880 suppliers and more than 27 million unique CAS Registry Numbers (15).

After the passage of the Toxic Substances Control Act (TSCA) by the U.S. Congress in 1976, regulatory officers began asking CAS for access to an electronic version of the TSCA Inventory and other national inventories like the EINECS Inventory in Europe. CHEMLIST, the regulated chemicals database, is available on STN and in SciFinder. It was originally built from data in the 1985 TSCA inventory of more than 308,000 regulated substances (16). It is the most accurate source of substance and regulatory information with validated CAS Registry Numbers and the world's most extensive collection of chemical names, consisting of systematic, trade and common names from 14 national chemical inventories.

Seeking current business information from the chemical enterprise worldwide, CAS introduced a database called Chemical Industry Notes (CIN) on STN in 1989. It was built from 100 trade journals (including bibliographic data, abstracts, indexing and CAS Registry Numbers). CIN offers chemical business news related to production, pricing, sales, facilities, products and processes, corporate activities, government activities and people. Today, CIN contains an estimated 1.7 million records drawn from 80 sources from 1974 to the present, including both domestic and foreign journals, trade magazines, newspapers and newsletters (17).

Trusting CAS for Current and Comprehensive Information

As well as covering chemistry in its broadest sense, the CAS databases are current and up-to-date so chemists can discover information sooner than from other scientific information providers. While the identification and approval process for new projects within research organizations typically requires a comprehensive review prior to moving forward, it still remains possible that, during the lifetime of a project, information can become available that could alter the scope of the project or even ruin it. Specific types of information affecting these efforts include:

- Recent publication of parallel or more advanced research efforts by competitors using the same approach and goals as the current project.
- Recent publication of key processes in the project by academic researchers or companies that limit patentability of the approach and/or enables competitor workarounds.
- New patent filings by competitors preventing freedom-to-operate for key processes in the project.
- Identification of old publications or patents (not identified previously) that limit the patentability of current efforts (i.e., prior art).

It is important that scientists have access to up-to-date information. There is intense competition to publish research first. The sooner the research is published by reliable sources, the more it provides scientists the help they need to plan and generate new scientific ideas and concepts.

In the mid-1960s, as CAS REGISTRY was being designed and implemented, chemists and computer scientists at CAS needed to estimate the pace and size of future growth – how many substances might chemists ultimately synthesize, and how fast? Initial estimates ranged from six to twelve million substances. Some predicted that when chemists had finally synthesized all possible substances; when they had combined all atoms in all synthetically accessible combinations, CAS REGISTRY might reach 25 million substances. While it took CAS 33 years to register its first ten million substances in published literature (18), in December of 2012, just 18 months after reaching 60 million small molecules, CAS registered its 70 millionth substance (19). Where are CAS analysts seeing these new substances? Patents, especially from the Asia Pacific region, have exploded during the past eleven years. In 2012, CAS saw a spike in Chinese patent applications unlike any in its history.

Covering 63 patent authorities, the CAS databases reflect patent activity around the globe through the years. Figure 1 shows Chinese patent growth as a major force in the Asia Pacific region and worldwide. In 2013 alone, the number of patents from the Asia Pacific countries was responsible for more than 67 percent of the patent publications seen by CAS, and China contributed around 65 percent of that region's patent output.

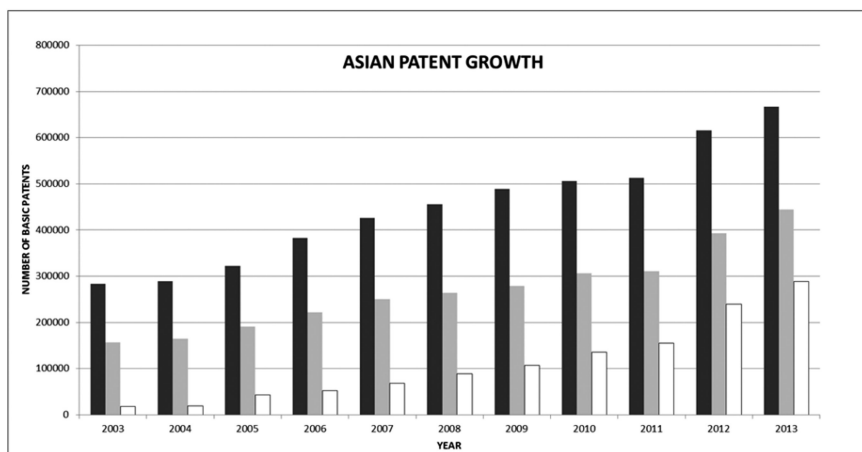


Figure 1. Asian patent growth over the past 11 years. The black bars show total worldwide patent growth; the grey bars show the contribution to worldwide growth from the Asia Pacific region (China included); the white bars show China only. Source: Cplus database.

For drug discovery scientists, knowing what's being patented for freedom-to-operate and intellectual property concerns is important. Every day, CAS scientists add more than 3,000 substances from Chinese patent applications alone. SciFinder and STN searchers have access to this novel patent information up to three months sooner than their competition.

The Future of Chemistry Research

At the inception of any research effort, whether it is a commercial drug development project or the potential subject for a PhD dissertation, researchers need to know what has been done in the past. They must find out what has worked, what hasn't worked and who else is working in the area of research. Before the 1970s, days, sometimes weeks, were spent in the library searching printed *Chemical Abstracts* indexes, and other compendia, to uncover what had been accomplished in the past. Extensive notes documenting the literature search were kept. Original journals articles, if not held in the local library, were acquired through interlibrary loans or document delivery services. Figure 2 is a visual representation of the relative time spent fetching (search and acquisition) relevant chemistry research and original literature versus the time spent reading and absorbing that literature (evaluation).

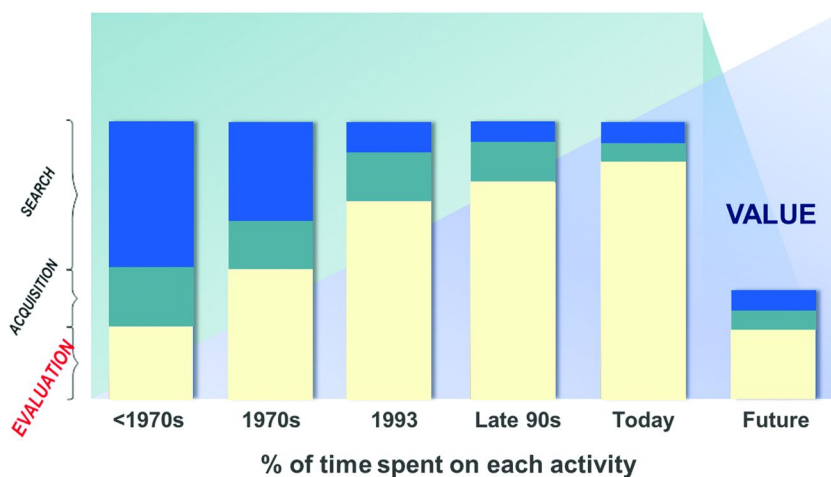


Figure 2. Content innovation and technology have significantly simplified scientific literature searches and provided a new area of opportunity: EVALUATION. Note: This graph is qualitative not quantitative.

With the advent in the 1970s of computer-based searching systems, time spent in the library began to shrink. Not all major reference works were available electronically, so library time was still necessary. Because of the intricacies of online searching systems, researchers often had to explain their questions to information experts who would then query online databases. As the secondary information industry moved through the 1980s and into the 1990s, searching became more efficient. More and more chemical information products were made available in electronic form. The primary literature was beginning to be delivered electronically in formats like PDF. In the mid-nineties, CAS developed SciFinder, a researcher's tool that was simple to use. Chemists were no longer required to understand the nature of arcane printed indexes or the sometimes complex search commands necessary to use online databases – they could search for themselves, find useful answers quickly, and access electronic versions of patents and journal articles – all from their own computer. So, today, the time required to search and acquire scientific information has been greatly reduced. A new problem has arisen – too many answers are resulting from the explosion in worldwide scientific publishing. CAS is currently developing features and functions in its products that take advantage of that content to reduce the time it takes scientists to evaluate a collection of patents and literature articles. The problem that CAS needs to solve now is not getting more answers but getting the best answers.

So what is CAS doing to aid researchers in getting to the most relevant literature quickly? CAS is adding more context to its records so scientists have more information that points to the right answers. Access to comprehensive and timely scientific information is vital. CAS, with its comprehensive, timely and high quality content, helps organizations eliminate or avoid wasted, unproductive efforts by quickly discovering business critical information as soon as possible. The search and acquisition time has been reduced and now CAS is finding ways to drastically cut the evaluation time. Let's describe some of those enhancements.

Experimental Procedures and Reaction Transformations

CAS provides access to more than 58 60 million single- and multi-step reactions and synthetic preparations (20), as well as associated experimental procedures for reactions, through SciFinder. Experimental procedures help scientists find useful reactions and the most relevant publications. CAS provides access to millions of experimental procedures from other sources including English-language translations from German and Japanese patents, the Shanghai Institute of Organic Chemistry, Chinese Journal of Organic Chemistry and Acta Chimica Sinica, hundreds of Springer journals and all ACS Publications journals in addition to English-language patents from the United States Patent and Trademark Office, European Patent Office, and the World Intellectual Property Organization (2000 to the present).

The group by reaction transformation feature in SciFinder saves users time reviewing reaction answer sets by speeding evaluation synthesis options and preferred pathways by grouping single-step reaction answers by transformation type. It classifies answers in a way that is meaningful to synthetic chemists and allows a user to easily manage and evaluate large, comprehensive answer sets.

Bioactivity and Target Indicators

Scientists working in the drug discovery arena, such as medicinal chemists, are experts in diseases, the protein pathways involved in those diseases, and small molecules or biologics that may inhibit, or enhance, protein expression. The essence of drug discovery is in identifying and validating druggable protein targets, designing lead molecules that affect their behavior and decorating that drug lead to maximize its efficacy.

In 2011, CAS began adding bioactivity indicators and target indicators to the small molecules in CAS REGISTRY. Bioactivity indicators are a defined set of approximately 260 bioactivity terms, much like therapeutic indications. A term is assigned to a substance in CAS REGISTRY when there is a high probability that the bioactivity indicator was reported for that substance in a journal article or patent. For instance, Velcade (CAS Registry Number 179324-69-7) is associated with bioactivity terms like antitumor agents and biological radio sensitizers. Target indicators are assigned by the same manner. Thus, Velcade is associated with the target indicators Akt kinase and 26S proteasome. These bioactivity and target indicators guide drug discovery scientists to new uses for known drugs, possible

side effects and the original literature where this pharmaceutical information was reported.

Relevancy Ranking

Relevancy ranking speeds access to desired results for researchers. Users sometimes performed multiple searches and refined them to obtain a more manageable answer set size. By using relevancy ranking in both STN and SciFinder, the best answers are pushed to the top, which leads to fewer follow-up searches.

Conclusion

Access to comprehensive and timely scientific information is vital for the advancement of science. For centuries scientists have routinely published their research; their conclusions may then be reviewed, confirmed and used by other scientists. Discoveries lead to more discoveries and science advances. CAS has been the repository of that research for more than 100 years.

CAS is cognizant of the fact that along with more information available in its databases comes the concern of navigating too many answers. CAS analysts are not only indexing and abstracting the important chemical content in reputable scientific publications including articles and patents, but also offering new content and functionality that aids searchers to quickly winnow a large collection of CAS records down to a useful and manageable set for their research. Recent notable content additions include graphical abstracts, experimental procedures for reactions, experimental and predicted properties, bioactivity and target indicators, citations and relevancy ranking capabilities.

In some sense, CAS has come full circle. The first issue of *Chemical Abstracts*, published on January 1, 1907 (8) contained 502 abstracts. Its purpose was more than raising the visibility of the American chemical enterprise. It was to summarize the growing volume of research papers being published worldwide for quick review. For many years, *Chemical Abstracts* was produced by a team of volunteer abstractors located around the world. Today, although CAS indexes well over a million documents on an annual basis, it continues to do so with the support of a team located around the globe. And, from the CAS customers' perspective, strives to develop database content and features that enable researchers, information professionals and patent searchers to winnow a massive collection of published information down to what's important for the problem at hand...just like what happened in 1907.

Many generations of scientists, information professionals, educators and students have used services from CAS, from printed *Chemical Abstracts* to STN and SciFinder. With knowledge gleaned from the CAS databases, scientists have begun their research efforts knowing what has been done before them, and in time, have contributed their own discoveries. In turn, CAS continues to include those discoveries in its databases.

References

1. Schenck, R. J. *Back to the Future*. Presented at the Fall 2012 ACS National Meeting.
2. Patterson, A.; Curran, C. *J. Am. Chem. Soc.* **1917**, *39*, 1623–38.
3. Crane, E. J. The Chemical Abstracts, Service - Good Buy or Good-by. *Chem. Eng. News* **1955**, *33* (26), 2753.
4. Crane, E. J. Why Indexers Turn Gray. *Chem. Eng. News* **1937**, *15* (8), 175.
5. CAS Report Highlights Progress. *Chem. Eng. News* **1962**, *40* (22), 90–97.
6. REGISTRY counter on the www.cas.org website (accessed April 2014).
7. CAS offers new online service. *Chem. Eng. News* **1980**, *58* (40), 34–35.
8. Shively, E. CAS Surveys Its First 100 Years. *Chem. Eng. News* **2007**, *85* (24), 41–53.
9. <http://www.cas.org/fulltext/cas-full-text-options> (accessed April 2014).
10. <http://www.cas.org/content/references> (accessed April 2014).
11. <http://www.cas.org/content/references/patworld> (accessed April 2014).
12. <http://www.cas.org/about-cas/cas-fact-sheets/registry-fact-sheet> (accessed April 2014).
13. <http://www.cas.org/content> (accessed April 2014).
14. <http://www.cas.org/content/reactions> (accessed April 2014).
15. <http://www.cas.org/content/chemical-suppliers> (accessed April 2014).
16. <http://www.cas.org/File%20Library/Training/STN/DBSS/chemlist.pdf> (accessed April 2014).
17. <http://www.cas.org/File%20Library/Training/STN/DBSS/cin.pdf> (accessed April 2014).
18. Toussant, M. A Scientific Milestone. *Chem. Eng. News* **2009**, *87* (37), 3.
19. <http://www.cas.org/news/product-news/70-millionth-substance> (accessed April 2014).
20. <http://www.cas.org/products/scifinder/content-details> (accessed April 2014).

Chapter 10

Spectra and Searching from Punch Cards to Digital Data

Marie Scandone*

**Bio-Rad Laboratories, Inc., 1500 J.F.K. Blvd., Two Penn Center, Suite 800,
Philadelphia, Pennsylvania 19102**

***E-mail: marie_scandone@bio-rad.com**

Chemical compounds number in the millions. Their molecular structures are unique and the manner in which chemicals absorb infrared energy is unique. Infrared reference spectra are important tools in the identification of unknown infrared spectra. Although reference spectra have been available for years, it was not until the use of computers and software programming that their true power was realized. They are important tools in science and chemical information that are available to researchers in academia, industry and government. Infrared spectra will be traced from a very humble beginning to the necessity that they are today in solving routine analyses in the laboratory. Over the years, the absorption peaks in infrared spectra have proven to be as distinctive as a human fingerprint and have provided a means of identification, classification, or verification of chemical compounds.

Introduction

From the perspective of Sadtler Research Laboratories of Philadelphia, which later became part of Bio-Rad Laboratories, Inc., this chapter focuses on infrared spectra and the importance of this technique in scientific research. The Sadtler mission was and remains to provide increasingly efficient solutions for the analytical consumer that combine the means necessary to analyze and access analytical data with the ability to communicate knowledge from that data.

From forensics to polymer chemistry, drug analysis to food chemistry, industrial research to art preservation, the infrared technique has proven to be a

reliable and effective tool. Today, with hand-held instrumentation, it is a quick and easy identification and verification method that is as important today as it was in its early years.

The Present

With an enormous database of spectral data, chemical structures, and chemical and physical properties, a system to store and manage dissimilar data was needed. Software then became an important and increasingly efficient solution for the analytical informatics consumer and provided the ability to create, manage, and communicate knowledge from those databases. Reference spectra were still important, but the breadth and depth of the collections necessitated better handling of the data.

Researchers use spectral search software along with spectral databases to identify unknown substances and verify the composition of synthesized materials in a number of applications and industries. First, precision instruments measure a substance and produce a spectrum, which is expressed as a graph showing a series of peaks and valleys that is specific to the sample material. That spectrum is then compared with a reference database of the measured spectra of known substances. If a matching spectrum is found, the material in question can be identified.

Today, with advances in computer technology, researchers can search 200,000 to 300,000 spectra and match an unknown spectrum query to the measured spectra of known compounds in a second or two. Only a generation ago, the same task would have taken days. Chemists can now automatically process spectra to improve search results and use a variety of search algorithms to find the best results.

The software can also perform spectral subtraction of multi-component spectra or complex mixture analyses that suggest the components that may comprise a mixture. This advanced computer analysis attempts to find all combinations of reference spectra in the libraries that, when combined in the correct proportions, result in a minimal difference between the query and composite spectra. The user can search for two or more components. The result is a series of composite spectra, each accompanied by the individual component spectra that comprise the composite spectrum as well as the residual spectrum (the difference between the query spectrum of the actual mixture and the spectrum that is the composite of the spectra of the individual components). The composite spectra are ranked by how closely they resemble the query spectrum. The speed of this process has been highly optimized, and numerous tests have confirmed its accuracy.

The Early Years

The history of infrared (IR) reference spectral databases is a fascinating journey through time. Today, thousands of spectra can be searched with the click of a button without any knowledge of how those experimental spectra were

generated. It was not always that easy, nor did anyone realize how important the data would become.

Before World War II, no infrared spectrophotometers were commercially available. If a researcher wanted to experiment with infrared data, an instrument had to be built by hand. These custom-made instruments consisted of a light source and mirrors, and as can be imagined, produced inconsistent results. Most of these studies were conducted in university laboratories, and the results were often questionable.

During the war, there were a variety of programs in place using infrared for the analysis of organic molecules. The U.S. Synthetic Rubber Program accelerated the use of infrared spectroscopy, which in turn prompted a need for better instrumentation. Dr. Arnold O. Beckman and Dr. Howard Cary were commissioned to build an infrared spectrophotometer. They were chosen because they were successfully able to produce ultraviolet spectrometers. The result was the IR 1, a single beam instrument, and the spectra that the instrument output represented the transmission of the infrared source that was not absorbed by the sample (1).

Another project during the war effort was headed by Richard S. Perkin and Charles W. Elmer, who were amateur astronomers. Their interest in optical design led to their manufacture of optical instruments. In 1944, Perkin and Elmer introduced another single beam instrument, the P-E Model 12 IR Spectrophotometer, and like the IR 1, it produced spectra with background absorptions. The data generated by these instruments were in the form of charts. This made the information very difficult to interpret, since only those who knew how to read the charts could actually locate a peak and make an assignment.

After the war, Perkin, Elmer and Beckman continued to generate interest by writing articles and praising the results of the new instrumentation. By that time, instruments were being produced commercially to measure infrared spectra. In 1947, there was editorial about the Baird Double Beam Infra Red Spectrophotometer in *Analytical Chemistry*. This instrument was designed by the Dow Chemical Company under the direction of Dr. Norman Wright and could output a spectrum that was very clear and readable. This instrument provided a new method of infrared analysis where the spectra obtained had flat baselines. Another plus was the ability to cancel out atmospheric bands resulting in spectra with absorption bands that were presented uniformly, from one end of the wavelength range to the other (2). This technological innovation provided reproducible results that would be required for standard reference data.

At the time, the only reference infrared spectra available were in a small collection of infrared spectral data of hydrocarbons from the American Petroleum Institute (3). In 1947, however, Sadtler Research Laboratories, an analytical laboratory that used the Baird instrument for chemical analysis and had a large supply of chemicals in-house, offered the first commercially available infrared spectra collection of fifteen spectra, on 7" x 18" cards, the size of the Baird chart. (See Figure 1.) These cards were cumbersome and hard to handle. Users would routinely powder their hands before using the cards to ensure that they did not stick together from perspiration and handling. Nevertheless, these cards proved that commercial instruments could provide readable and reproducible results

and that a library of infrared reference spectra could aid in the identification of unknown chemical compounds.

As the collection of cards grew, researchers moved to McBee cards (4) with holes and notches around the edges for sorting. (See Figure 2.) A hole represented the presence of an infrared absorption band while a notch represented its absence. These cards were sorted by employing knitting needles to locate all the chemical compounds with similar bands or functional groups. A researcher would take a number of cards, put them in a sorter and use the knitting needles to locate specific holes in the card that represented specific spectral absorbance regions. The cards would then be shaken, and only those cards that met the criteria of the search would be separated from those that did not match. Those selected cards would then be examined. For example, if the researcher was trying to identify an aldehyde, cards with absorptions in the desired regions would be located by inserting knitting needles into the appropriate holes in the cards. There was still the problem of separating the cards, for which talcum powder was employed to keep the cards apart.

During this time, there was even an Infra-Red Punch-Card Committee (5) whose original function was to survey existing punch-card systems. Its purpose was to produce a “standard” punch-card that would “facilitate the exchange of infra-red data between laboratories.” The committee examined all the types of cards available and designed two cards, a “bibliography” card and a “compound” card. The “bibliography” card contained the reference number, the subject field, an apparatus field, and the year of publication. The face of the card listed the author, title, journal, and Chemical Abstract Number. The “compound” card included a field for the wavelengths of the absorption maxima, a field for functional groups, a field for melting and boiling points and a field for the number of carbon atoms. The face of the card displayed the name, the empirical and structural formulas, and certain information abstracted from the literature. The back of the card showed the spectrum, as well as a table of the wavelengths of the absorption bands. This was the first attempt to standardize spectra and accompanying property information. However, the attempt was short-lived. The cards were too hard to use in the lab, and the concept ultimately failed.

By 1949, there was a shift to IBM cards. With this shift, more data could be incorporated into the card. This included the molecular formula, the location of the absorption, and a serial number. This number could be linked to a chemical name in numerical and alphabetical order and printouts accompanying the cards were sold.

During the 1950's, PerkinElmer was working on a double-beam instrument. Paul Wilkes engineered the first P-E Model 21 IR, which made IR analysis a routine laboratory tool (6). PerkinElmer approached Sadtler Research Laboratories, which had built up its own collection of infrared reference spectra. PerkinElmer wanted their instrument in every commercial and academic laboratory, but they had to show that consistent results could be achieved in every laboratory using their instrumentation. The first two PerkinElmer double-beam instruments were the prototype instrument and an instrument used for sales calls. Sadtler received the third PerkinElmer instrument that was manufactured and started to run spectra as fast as it could.

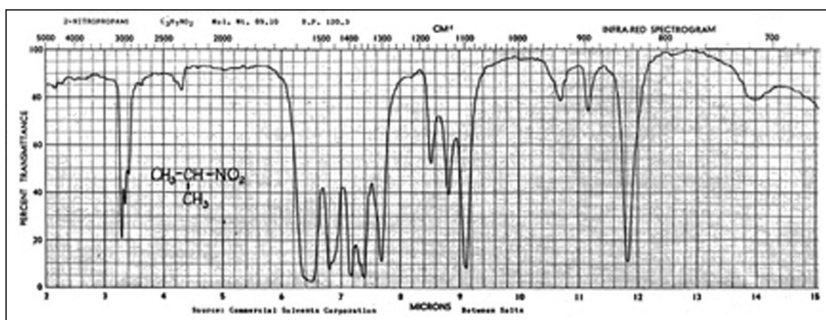


Figure 1. Example of an individual infrared prism spectrum distributed as a flash card by Sadtler Research Laboratories. Reprinted with the permission of Bio-Rad Laboratories, Inc. Copyright 1949 Sadtler Research Laboratories, Inc.

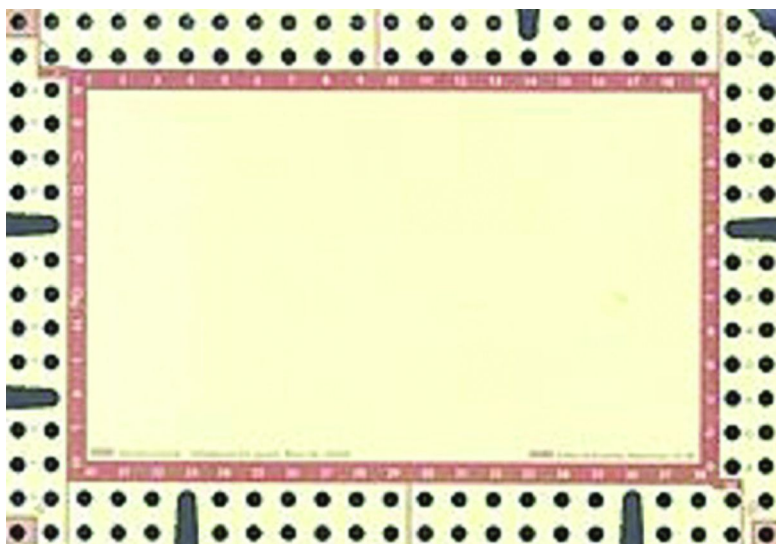


Figure 2. Example of a blank McBee card with holes punched for functional group location.

At about this time, the National Bureau of Standards considered producing infrared reference spectra, but Sadtler already had a large collection of infrared spectra that it was distributing (6). Before the National Bureau of Standards started to collect and publish its own infrared spectra, the “Sadtler Standard Spectra” were born. The government did not want to compete against a commercial endeavor, so it decided not to produce infrared spectra. In 1955, Sadtler collected all the spectra that it had run to date and printed its first collection of ten thousand spectra. (See Figure 3.) There were continual small upgrades to the collection until a user suggested that an optimal addition to the collection would be two thousand spectra annually. After that, Sadtler published two thousand new

infrared reference spectra per year, and the collection continued to be printed until 1996, after which only digital products were created.

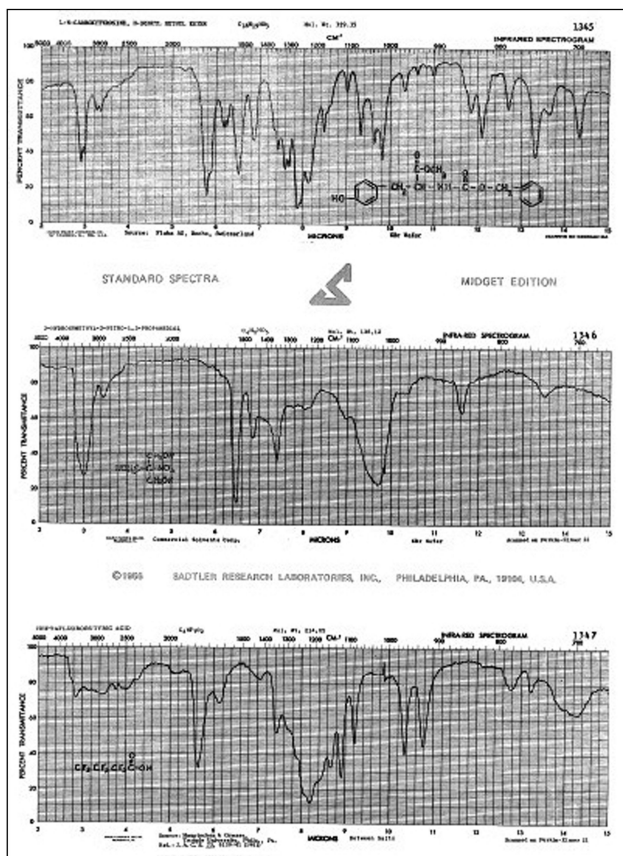


Figure 3. Example of a page of the first Sadtler collection of spectra that appeared in the "Green Books". Reprinted with the permission of Bio-Rad Laboratories, Inc. Copyright 1955 Sadtler Research Laboratories, Inc.

This first collection contained prism spectra, which were measured from 2 to 15 microns. This was the standard for many years as chemists became comfortable using reference spectra to identify or classify their spectra. It was generally accepted as one of the best techniques for the identification of unknown compounds as a band-for-band match of the IR spectrum of an unknown compound against an IR reference spectrum provided the most positive method of analysis available at the time.

As the technology improved, grating or dispersive infrared instruments were being used to measure infrared spectra. (See Figure 4.) This instrument was similar to a prism instrument since it had a light source and mirrors, but the grating was constructed to separate the wavelengths of light and direct each of them through a slit to the detector. Each wavelength was then measured with

the slit monitoring the spectral bandwidth and the grating moving to select the wavelength being measured. The amount of light of a particular wavelength that was absorbed by the sample was measured by adjusting the reference beam until its intensity was equivalent to that of the beam transmitted through the sample.

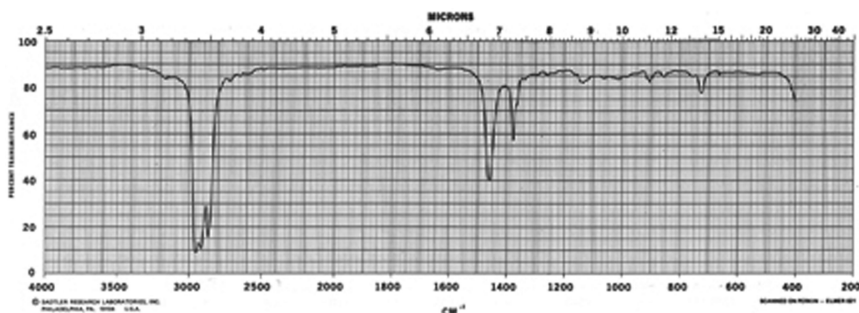


Figure 4. Example of an infrared grating spectrum. Reprinted with the permission of Bio-Rad Laboratories, Inc. Copyright 1969 Sadtler Research Laboratories, Inc.

The X-axis, or peak position on the spectrum, provides information about the wavelength and is usually presented in wavenumbers. It can also be represented as reciprocal centimeters (cm^{-1}). Typically, the wavenumber range is 4000 to 400. The Y-axis or peak intensity provides information about how much the sample absorbs the energy, the units can be in Percent Transmittance or Absorbance. Percent Transmittance ranges from zero to 100% while absorbance ranges from zero to infinity.

As the Sadtler collection grew, a system was needed to locate specific spectra and identify unknown spectra. The most widely used index was the Sadtler Spec-Finder index (7), which was built around numeric tables of infrared spectral absorptions. The Spec-Finder listed each compound's absorption peaks, with the strongest band in a separate column. The user identifying an unknown compound would look at the spectrum of the unknown, determine the highest peak, go to the page in Spec-Finder index listing all the spectra with the same strong band, and scan the list until a compound with peaks identical to the unknown spectrum was found. Each region from 2000 to 400 wavenumbers listed one peak per 100 wavenumbers if there was significant absorption. The regions from 4000 to 2000 wavenumbers listed one peak per 200 wavenumbers if there was significant absorption. As a final check, a chemist could then look up the spectrum in a printed volume. Without the numeric Spec-Finder tables, a researcher would have faced many hours of comparing the unknown with the spectra in the entire Sadtler collection. With the Spec-Finder tables, the job was reduced to minutes (8).

In 1967, the first attempt to computerize the system was made by Sadtler. Researchers could search through the magnetic tape library of 50,000 spectra in 32 minutes, conducting dozens of searches simultaneously. The IBM System /360 Model 30 was used (3). Instead of matching just the strongest absorption bands, the computer could check every absorption peak, as well as other chemical

characteristics. Considered was molecular formula, molecular weight and chemical composition or classification.

The availability of these reference spectra made it possible for chemists to identify and verify their chemical compounds as well as functional groups. It was this access to reference data that made infrared analysis an important tool in laboratories around the world. Because the molecular structure of a chemical is unique, the manner in which it absorbs infrared energy is also unique. That is why a spectrum becomes a reliable “fingerprint” that can be used to classify or identify a chemical compound.

All these attempts to identify unknown infrared spectra received a boost with the introduction of the laser-referenced, computer-controlled, FT-IR Spectrometer. (See Figure 5.) Spectral identification had always been a time-consuming process. With the introduction of computers, the quality and speed of sample processing had improved, but spectral identification still took time. As infrared reference databases continued to increase in size, however, so did the problems in the management of that data.

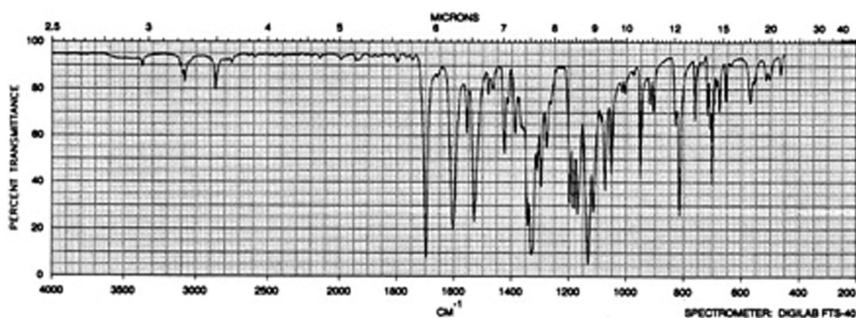


Figure 5. Example of a FT-IR spectrum. Reprinted with the permission of Bio-Rad Laboratories, Inc. Copyright 1996 Bio-Rad Laboratories, Inc., Informatics Division.

Spectral Collections

There are a number of infrared spectra collections available to scientists. Some databases can be used with search software on an infrared instrument for identification while others can only be visually reviewed for comparison. The quality of the spectra may vary but they can assist in the verification of chemical compounds.

Some of the numerous available infrared spectral databases:

Bio-Rad Sadtler Databases (9)
Bio-Rad Laboratories, Inc.
(ca 230,000 spectra)

Spectral Database for Organic Compounds, SDBS (*10*)
National Institute of Advanced Industrial Science and Technology
(AIST), Japan.
(ca 52,500 spectra)

SpecInfo on the Internet (*11*)
John Wiley and Sons, Inc.
(ca 30,000 spectra)

NIST Webbook (*12*)
National Institute of Standards and Technology
(ca 16,000 spectra)

Nicodom Infrared Spectral Databases (*13*)
Nicodom Ltd.
(ca 140,000 spectra)

Aldrich Spectral Databases (*14*)
Sigma-Aldrich Company
(ca 54,000 spectra)

SciFinder (*15*)
Chemical Abstract Service, A division of the American Chemical Society

The Coblenz Society (*16*)
Spectral databases

The Future

Seasoned spectroscopists who could identify a spectrum from sight are retiring. As more of these positions will be replaced by non-spectroscopists, the next generation will rely more heavily on spectral expert systems to identify and classify their chemical compounds and less on their own expertise. Similarly, the management of infrared spectral reference databases will move from systems to archive and warehouse spectral data to tools that help identify and evaluate information. Users, who do not remember a time before home computers and mobile phones, will require a more simple, more intuitive, yet more powerful spectroscopic search expert system for unknown identification that utilizes all of the spectral databases, software technology, and expert knowledge available to provide the most complete answers possible. The technology is just becoming available which utilizes spectral intelligence but there is still more that can be done.

The old adage “content is king” remains true: the more spectra that are available to the user, the higher the probability that an unknown chemical compound can be identified from its infrared spectrum. Commercial infrared reference databases as well as user-built databases provide the greatest opportunity

to identify or classify a chemical, whether it is a pure compound or a mixture. In the end, of course, the user must review the results before an absolute confirmation can be made.

References

1. Sadtler, P.; Sadtler, T. *Appl. Spectrosc.* **1985**, 39 (6), XIX–XXII.
2. Wilks, P. A. *Spectroscopy* **2001**, 16 (12), 14–15.
3. *IBM Computing Report for the Scientist and Engineer* **1968**, IV (3), 7–9.
4. National Research Council (U.S.). *Committee on Modern Methods of Handling Chemical Information* **1964**, 324–330.
5. *J. Opt. Soc. Am.* **1950**, 40 (8), 547–548.
6. Personal Recollections of Philip Sadtler.
7. Shaps, R. H.; Sprouse, J. F. *Ind. Res. Dev.* **1981**, 2, 168–173.
8. *Issues in Science and Technology Librarianship*, Summer 2003, www.istl.org.
9. *Bio-Rad Sadtler Databases*; Bio-Rad Laboratories, Inc. <http://www.knowitall.com/> (accessed October 21, 2013).
10. *Spectral Database for Organic Compounds*; SDBS National Institute of Advanced Industrial Science and Technology (AIST), Japan. <http://sdb.srioddb.aist.go.jp/> (accessed October 21, 2013).
11. *SpecInfo on the Internet*; John Wiley and Sons, Inc. <http://www.wiley-vch.de/stmdata/specinfo.php> (accessed October 21, 2013).
12. *NIST Webbook*; National Institute of Standards and Technology. <http://webbook.nist.gov/chemistry/> (accessed October 21, 2013).
13. *Nicodom Infrared Spectral Databases*; Nicodom Ltd. <http://www.nicodom.cz/> (accessed October 21, 2013).
14. *Aldrich Spectral Databases*; Sigma-Aldrich Company. <http://www.sigmaaldrich.com/> (accessed October 21, 2013).
15. *SciFinder*; Chemical Abstracts Service, A division of the American Chemical Society. <https://www.cas.org/products/scifinder>.
16. The Coblenz Society. <http://www.coblenz.org/education/spectral-databases>.

Chapter 11

Teaching Chemical Information for the Future: The More Things Change, the More They Stay the Same

Judith N. Currano*

Head, Chemistry Library, University of Pennsylvania, Philadelphia,
Pennsylvania 19104

*E-mail: currano@pobox.upenn.edu

Teaching chemical information is a part of the job of every chemical information professional, as well as many chemistry professors. Over the years, many different styles and formats for imparting information skills to students have evolved. While some individuals, particularly in earlier years, think that a dedicated course is the best way to teach students to use the literature, the trend in the early 21st Century has been towards course-integrated instruction. This article provides a contemplative look at some core ideas that all instructors should bear in mind when attempting to teach chemical information skills in the classroom and concludes with a number of classroom activities that incorporate these skills.

The Past and Present of Chemical Information Instruction

The Need To Read

Chemistry has always been a very literature-heavy discipline. In addition to the need to prove that a project is novel, the expense of performing chemical experiments and the dangers inherent to chemical research drive the savvy chemist to the literature before they enter the lab. Prior to using an extremely expensive reagent, one wants to make sure that the reaction to be performed has a high likelihood of success, and before mixing two chemicals, one needs to determine

whether or not the combination is safe. As a result, chemists spend a great deal of time reading the primary and secondary literature; a 2002 publication by Carol Tenopir and Donald King indicates that chemists spend more time reading than any of the other scientific disciplines that they studied and read more articles per person than all but the medical researchers (1).

This “need to read,” coupled with the continuing expansion of the number of publications available to be read in a given year, means that chemists require a reasonably high degree of expertise in searching the literature. The American Chemical Society’s Committee on Professional Training (CPT) has, for some years, included a well-appointed library and training of students’ in information retrieval as part of their evaluation of chemistry programs for degree approval. Section 7.2 of the 2008 guidelines (2) gives a brief overview of the information skills that undergraduates should achieve prior to graduation.

Students should be able to use the peer-reviewed scientific literature effectively and evaluate technical articles critically. They should learn how to retrieve specific information from the chemical literature, including the use of Chemical Abstracts and other compilations, with online, interactive database-searching tools. Approved programs must provide instruction on the effective retrieval and use of the chemical literature. A specific course is an excellent means of imparting information-retrieval skills.... Integrating the use of these skills into several courses is also an effective approach (2).

In addition to this statement within the CPT Guidelines themselves, CPT published a supplement describing these skills in greater detail (3).

Because of the importance of chemical information education, brought to the forefront by the CPT Guidelines, many organizations have sought to assist instructors in planning curricula for classes and training sessions. The ACS Division of Chemical Information (CINF) Education Committee presented a workshop at some ACS National Meetings entitled, “Teaching Chemical Information,” which they last ran in 2007. The Special Libraries Association Chemistry Division released a white paper on suggested information literacy competencies for undergraduate students in 2007. In 2011, they produced a second edition, working with the CINF Education Committee (4).

Surveys of Chemical Information Training in Colleges and Universities

A brief foray into the literature indicates that, historically, there have been three main ways in which chemical information skills are taught to students in undergraduate and graduate chemistry programs. Students can be given a lecture on using the information tools as part of another course, usually in support of some assignment or work done in that course; an instructor can devote a cluster of lectures in a practical skills seminar course to using the literature; or an institution can offer a required or optional course in chemical information. Several individuals and groups have published surveys of institutions’ methods of training students in chemical information, and Arlene Somerville reviews

the results of many early surveys in her 1985 publication (5). According to her comparisons, the preferred method of imparting chemical information education to students at PhD-granting institutions prior to the late 1960s was by means of a dedicated course; in fact, in a 1953 survey by Jahoda, 32 out of 60 schools surveyed reported having a course in “chemical literature,” and 20 of these schools required the course of their students (5, 6). This is a significantly higher percentage than indicated in an informal review of course catalogs performed by Soule in 1932, which discovered that only 20% of the 100 catalogs examined contained a course in “bibliochresis (7),” despite the fact that Soule reports that industry executives cited knowledge of and proficiency in using the chemical literature as being highly desirable skills in potential employees (7). After the late 1960s, the number of courses at PhD-granting institutions sloped down into the thirty percentages, while the number of courses at BS and MS schools remained high (5).

With fewer students taking courses dedicated to the chemical literature, instructors and librarians have clearly turned to other methods of instructing students in the use and retrieval of chemical information. The CINF Education Committee performed a survey of 331 institutions in 1984, to learn how chemistry departments were training their students (8). Although, of the 218 schools that responded, only 32% offered a dedicated course in chemical information, 63% of the institutions indicated that they integrated chemical information instruction in other courses in their curricula. The numbers had risen by the time the Education Committee repeated the survey in 1993, and, of the 390 departments that responded, 41.5% offered dedicated courses, while 76% indicated that they integrated chemical information into at least one course (9). The most recent survey, performed in 2005 by Garritano and Culp, determined that things had changed very little in twelve years; of 249 institutions that responded to the survey, 37% offered a dedicated course in chemical information, and 74% incorporated information skills and training into other courses (10). Despite Jahoda’s assertion in 1953 that “only [devoting an entire course to chemical bibliography] gives the student sufficient training,” (6) instructors clearly favor the course-integrated approach, even at institutions where a dedicated course is offered.

Methods and Models for Teaching Chemical Information

The methods and models by which individuals and institutions teach chemical information techniques to their students are as diverse as the institutions themselves, and they tend to be published in two basic categories of articles: articles that describing the syllabus, goals, and implementation of a dedicated chemical information course, and articles that describe the integration of chemical information skills into other courses in the curriculum. The latter type ranges from articles describing how chemical information skills are integrated into a comprehensive course of study to articles that present assignments or search examples used in specific classes. Chemical information educators have also published bibliographies of chemical information or chemical information instructional resources; most notable among them are Carol Carr’s two “Teaching and Using Chemical Information” bibliographies, now somewhat outdated (11);

Gary Wiggins' Clearinghouse for Chemical Information Instructional Materials, which is currently being overhauled as part of the *Chemical Information Sources* WikiBook (12); and XCITR: eXplore Chemical Information Teaching Resources, a repository that educators can use to share their tools and resources with one another (13). In addition to writing articles and book chapters, chemical information educators frequently present their strategies and techniques in symposia sponsored by the Divisions of Chemical Information and Chemical Education at ACS National Meetings, the abstracts for which can be accessed through a SciFinder search on the desired topic.

The articles referenced in this paper are examples selected from the huge body of chemical information literature to demonstrate the particular aspect of instruction under discussion. Interested readers can easily identify more articles of interest through a quick or extensive search of major chemistry and general science databases (14). Browsing the articles that appeared in the "Chemical Information Instructor" column from the *Journal of Chemical Education*, edited first by Arlene Somerville and then by Andrea Twiss-Brooks, also provides an interesting look at the history and evolution of chemical information instruction. One can accomplish this through a search of the *JCE* Web site for the exact phrase "Chemical Information Instructor" (15).

Dedicated Courses

Many of the papers describing dedicated courses in chemical information over the years present syllabi for the courses, and some go so far as to include sample homework queries. As a group, the courses are taught in-person by either chemistry faculty or chemistry or other science librarians, and the students learn through lectures, demonstrations, assigned readings, and, most importantly, searching assignments. It is interesting, although probably not surprising, to note that the desired learning outcomes of the courses have changed very little over the years. For example, the following paragraph comes from Soule's 1932 description of the objectives of the University of Michigan's chemical information course:

First, each student is expected to become familiar with the standard reference books and know how to use them. Second, he must know the routine of consulting the literature down to date and be reasonably sure that his search will reveal the papers having an important bearing on the problem being investigated. Third, he is helped in the difficult task of critically evaluating the literature. This includes the ability to read an article intelligently and write a brief summary covering the essential points. Finally, how to keep abreast of the times is given due consideration (7).

67 years later, Ricker and Thompson, in describing the chemical information course offered at Oberlin College, indicate that they begin their course syllabus with an almost identical statement of goals:

The course is designed to familiarize you with the major sources of information, to help you to learn how to assess the information that you obtain, and to develop skills in presenting structural and numerical information in chemistry (16).

A dedicated course has the advantage of presenting information resources in context with one another, giving the students a comparative look at a group of tools, many of which may appear, at first glance, to be interchangeable. It gives the instructor time to introduce a wide variety of techniques and resources and gives students the opportunity to develop and implement transferable skills. Finally, the regular meetings and assignments allow students to develop a relationship with the instructor, as well as to become more proficient searchers through extensive practice. The disadvantages to the dedicated course model are two-fold; a school's chemistry curriculum may be too full to allow the incorporation of one into the suite of courses offered, and the fact that the practice searching is happening in a course devoted to chemical information can place the skills learned in a vacuum, with students being uncertain how to apply them to actual research problems.

Course Integrated Instruction

The challenge for instructors of chemical information has been to teach the subject in the correct context for students to gain the optimal level of skills and understanding. Based on the survey results presented in the previous section, the trend today is to integrate chemical information skills in other courses. Course-integrated instruction (CII) may take the form of a single encounter with undergraduate or graduate students, many of which support specific "information-intensive" assignments in the course. For example, Pence presents a five-fold chemical information assignment geared at non-science majors; students must use the literature to validate or invalidate an urban legend, determine the connection between chemistry and their chosen major, compile facts about pollution in their hometowns, research a specific chemical's hazards, and write a short risk-benefit analysis (17). Locknar and coworkers describe the integration of chemical information skills into a large first-year chemistry course, in which the faculty instructor used library-licensed databases in class to locate references related to various topics covered, and the librarians constructed homework assignments and brief, online videos to demonstrate specific information skills that the instructor wanted his or her students to learn (18). Other models of CII, such as the system of course integrated instruction across the chemistry curriculum at the University of Rochester presented in 2003 by Somerville and Cardinal (19), are much more extensive and involve meetings with the students in particular courses throughout their undergraduate study. Walczak and Jackson present a particularly interesting scenario from the analytical chemistry course at St. Olaf, in which students use a role-playing approach to examine a number of situations from the perspective of various departments within a corporation. Through the course of the semester, the students are asked to work in teams

composed of individuals playing each of the four assigned roles to complete assignments, including four assignments specifically related to the chemical information literacy competencies identified as critical by faculty (20).

The advantages to CII, then, are the ability to place the information skills in a more “real world” context, showing the students how information resources can work with other chemical research techniques as part of the process of doing chemistry. The obvious disadvantage is that, without careful work on the part of both the course and chemical information instructors, CII can turn into the obligatory “library lecture,” in which a librarian, knowing that this may be the only chance he or she has to instruct the students, races through all possible resources that they may need for the rest of their course of study.

Looking to the Future: General Principles of Teaching Chemical Information

In 2012, CPT released a white paper, including some proposed changes to the guidelines for undergraduate education. The section on student understanding and use of the chemical literature was one of the sections earmarked for revision. The proposed guidelines emphasize formal instruction in chemical information retrieval, as well as the implementation of information skills in other areas of the curriculum.

A critical student skill is the ability to efficiently and effectively retrieve information by searching the chemical literature. Among the types of searches that students must be able to carry out are searches by keywords, authors, abstracts, citations, patents, registry numbers and structures/substructures. Students must have ready access to databases that allow them to complete these searches and must be able to assess the quality of the search outcomes. Students must be able to read, analyze, interpret, and cite the chemical literature as applied to answering a chemical question. The development of student skills for searching and utilizing the chemical literature must be accomplished through formal instruction and reinforced through undergraduate research or through projects incorporated into the curriculum (21).

This focus on skills, rather than on specific resources is somewhat at odds with the way in which traditional information instruction has been performed. However, since computerized information systems became accessible to the general public in the 1980s and 1990s, chemical information classes have relied extensively on active learning, using the information tools in the classroom itself. It is relatively easy to adapt a traditional “point and click” demonstration in class to be a truly interactive session, in which students experiment with resources in a controlled setting and leave with a new-found information retrieval skill, rather than familiarity with the mechanics of a single tool.

Technology has changed the way we teach chemical information skills to our students... But should it have?

Nobody would disagree that the Internet and, more specifically, the Web, have completely revolutionized the way in which chemical information is used and distributed. As printed sources are replaced by electronic versions, researchers become increasingly accustomed to getting the information that they want delivered to their fingertips quickly, without putting much effort into the process. Coupled with this, students who have grown up in the “information age” are used to employing electronic search systems continually and are quite adept at quickly locating items ranging from restaurant reviews to retailers who sell extremely esoteric merchandise. Since the mechanics of searching for and retrieving information have become so simple and effortless, inexperienced researchers are lulled into a false sense of security. They feel confident in their ability to locate information, while lacking many of the skills needed to locate *everything* that they need. The techniques that serve them well in their day-to-day lives may not transfer effectively to the complicated (in their minds, arcane) organization of the chemical literature.

When approaching chemical information instruction, therefore, it is important to remember that, despite the changes in technology, the way in which one constructs a good search has not changed at all in the past hundred years. Over the fifteen years that I have been teaching students to search, I have changed my style greatly, and this change has been reflected in the syllabus of my course and the topics that I select when asked to teach guest lectures in other people’s classes. The first semester that I taught chemical information to graduate students, I did the “traditional librarian thing,” and I presented a weekly series of resource-based lectures. Week one dealt with Franklin, Penn’s online library catalog, and I introduced the resource and then led the students through a series of example searches using it. I taught them how to use Boolean operators, truncation, subject headings, the works. The students had homework assignments to reinforce the concepts that I reviewed in class, and most earned perfect or near-perfect scores. I got phenomenal course evaluations, and, when the semester was over, many of the students decided that the most effective way for them to discover whether or not Penn had access to a particular book or journal was to e-mail me. Clearly, there was either something wrong with our library catalog or something wrong with my approach to teaching them to use it. Despite the fact that I haven’t seen an OPAC that I like to date, I decided on the latter (after all, they had all demonstrated mastery of the material on their homework), but I didn’t know *what* the problem was. After all, I was explaining things clearly in a library-approved method. I was employing visual, auditory, and kinesthetic learning techniques to great effect, and I was failing to teach my students anything except that I was a great searcher.

I didn’t realize what was wrong with my approach until I was team teaching a class with a librarian colleague. We had a single, sixty-minute lecture in someone else’s class, in which to introduce the students to everything that they would need for the literature review of an extensive project. My job was to teach them the theory of finding chemical information, so, I went through the methods of

identifying substances for information retrieval and talked a great deal about the “great science of trial and error” that is searching the literature. Because my teaching style has always been very participatory, I relied on the rather taciturn group of students to think about substances and suggest methods in which they could identify them, taking more than my allotted share of class time. This left my colleague with the challenging task of demonstrating many different search strategies and sources in a very short amount of time, moving a mile a minute in order to get through the planned content and quickly losing the class’s interest, which had been lukewarm at the beginning of the hour. As I watched my colleague struggle, the problem that been dogging me in my own class was thrown into sharp focus; I, myself, had fallen into the classic librarian blunder: my colleagues and I focused on teaching the students how to use particular tools, rather than helping them to learn how to find information.

When I stand in front of my graduate class for the first time every semester, I explain to them that learning to find information is like learning to do mathematics or learning to speak a foreign language. This is not a subject at which you can become an expert through book learning and cram studying. The only way in which to learn to find information is through practice; as a result, I have always given a large amount of homework in my classes. If I solve a problem for the students, it is abundantly clear to them that the strategy and applications I choose to use are appropriate for the situation; choosing an appropriate resource from scratch and designing search parameters that will find all relevant information with a minimum of irrelevant results is much more challenging because of the sheer number of tools and strategies from which to select. After my course-integrated instruction epiphany, I decided on a completely different approach to teaching, both in my own course and in those of my colleagues. Students are relatively adept at figuring out the mechanics of using databases and search systems, and, using crowd-sourced tips and techniques, they are able to figure out how to input a search that will yield results. With a few exceptions, I have moved away from dwelling on the keystrokes and clicks that will cause a database to retrieve relevant hits. Instead, my goal has been to teach the students how to think about information and introduce them to a core set of transferrable skills that they can use to successfully use the sources to which they have access. This method is grounded in a series of general principles, which, for the most part, are format, platform, and system independent. To whatever extent possible, I try to introduce all five principles into every instruction session that I do, be they stand-alone sessions in other people’s classes or a semester-long, required course.

Principle 1: If it is not there, you cannot find it. If it is there, you need to know what to call it.

The basics of information retrieval are actually quite simple. A resource has access to a certain body of information, which I like to call a “universe,” and it is unaware of anything outside of this universe. I frequently find that I need to remind students of a fairly fundamental fact, “If it is not there, you cannot find it.” When one performs a search of the literature, one is attempting to match a word, phrase, value, structure, or reaction to something that is within the universe

of information being searched. If there is an exact match, the information can be located. If, however, the information input does not find a match within the source's universe, the search will fail. Despite the fact that many modern search systems possess a certain degree of "intelligence," autostemming results to find word variants or mapping terms to controlled vocabulary and then searching for the controlled terms, one still needs to match the search criteria to information actually present in the source. I encourage students to think about the concepts that they wish to research and play a game of "psych out the author/indexer." During class sessions, when I teach search design, I encourage my students to brainstorm as many possible ways of describing a concept as possible. This gives a wealth of possible search terms from which to choose, and the terms that they will ultimately select and employ can vary depending on the source selected for the search.

Principle 2: Information systems take you literally... Except when they don't... And even then, they do!

Early in any class dealing with information skills, I declaim my first law of information retrieval, "Generally speaking, information systems take you literally." For the most part, you need to be aware that the comprehensiveness of your search results is entirely dependent on the information that you put into the system. For example, if you are searching for information on dichloromethane, you will retrieve all records that call the substance by that name, but you will retrieve no records that call the substance only methylene chloride. Not only is it important to do a comprehensive brainstorming exercise for an exhaustive search (one that retrieves all possible information on a topic), you must take care that the context in which you are using your search terms and the ways in which you are combining them are consistent with the information that you want to retrieve. I constantly need to remind my students that, no matter how intelligent a search system looks, it is only going to do what you tell it to do. Therefore, your instructions must be clear, and they must accomplish your desired goal. As a result, it is critical to understand the organization and scope of a source and, as much as possible, to understand a little bit about the search algorithms and behind-the-scenes work that the systems are doing. This leads to Principle 3.

Principle 3: If you want to use a source effectively, you need to understand its scope and organization.

In J.K. Rowling's *Harry Potter and the Chamber of Secrets*, one of the characters offers some words of wisdom that seekers of information would be well advised to bear in mind: "Never trust anything that can think for itself if you can't see where it keeps its brain" (22). The goal of information providers is to create user-friendly search systems. There are a large number of vendors in the current arena, as well as a limited number of dollars to be spent on information systems. Recently, institutions have been scrutinizing the use of information systems, cutting those that fall beneath a threshold "cost per search" value. As a result, vendors seem to be increasingly marketing to end users, rather than librarians, and most end users want to be able to jump into the tools and begin to

retrieve helpful results immediately without spending hours learning appropriate search syntax. On the positive side, this has led to tools with appealing designs, “easy to use” features, and time-saving information analysis and processing tools. Newer search engines perform natural language searching, automatically parsing strings of text that resemble sentences (“the science of the flavor of wine,” for example) and stemming words to find variant terms, making traditional Boolean operators and truncation symbols seem obsolete and needlessly complex to learn. These “ease of use” features, however, come at a cost; it is hard to structure an exhaustive query when one does not know exactly how the system is working behind the scenes to search. In other words, it is impossible to determine what may have been missed or omitted.

Whenever I teach students to use a specific source, therefore, I spend a great deal of time explaining, to the best of my knowledge, its history, organization, and any eccentricities of which I am aware. Many students think that a way of searching that works in many tools will work in all tools; for example, when you type several terms on a line, you will get results that contain all of them. However, a student who attempts this strategy in SciFinder could miss information if the system groups two or more of the terms together into a “concept.” Knowing how SciFinder defines “concepts” and learning to use prepositions to get either items with the two terms “closely associated” (as well as knowing what a close association is) or items where both terms are anywhere in the record can help a student either broaden or narrow a search. Specialized Boolean operators like SAME and NEAR/*n* in Web of Science or W/*n* and PRE/*n* in Scopus can save students time when filtering search results for relevance, but they can also be overly limiting. However, a clear understanding of the presence and nature of these tools will allow the student to decide the breadth of the search at the outset and can save time in the long run.

Principle 4: All information sources are not created equal.

In recent years, I have encountered a staggering number of students who lack a fundamental understanding of the differences between different sources of information. Many focus on the “window dressings” of the tools without being aware that they are searching completely different data sets. A post-doctoral fellow in chemistry once told me that he would stop using SciFinder when all of the features that he found most valuable were available in Web of Science because he found the Web of Science database easier to use. He was shocked when I told him that the two systems were built using completely different databases and that not all of the journals and other document types monitored by one could be found in the other.

I now spend a great deal more time teaching students to evaluate search systems and the information that they contain. When discussing the databases themselves, we talk about their content first and then the relative strengths and weaknesses of their search options and interfaces. For example, in the physical chemistry section of my chemical information course, the students learn about six different databases that index and abstract the journal literature: SciFinder, Inspec, Compendex, Web of Science, Scopus, and MathSciNet. We look at the

subject coverage of the databases, and discuss how the topic of the query can dictate the selection of database. Where the subjects overlap, we look to see if there is a distinguishing search feature that will make one tool easier to use than the others for the search in question. Finally, I am sure to stress that, when they require an exhaustive search, it is necessary to search all indexing and abstracting sources whose scope encompasses the subject of the query.

I find it critical, however, to acknowledge the sources with which the students are already familiar and that they want to use for their work. To vilify Wikipedia because its authorship is not clearly delineated would be an exercise in poor public relations, particularly since some of the content is peer reviewed or verified. Rather than telling the students not to use a popular but suboptimal tool, I place the tool in its proper context, explaining its strengths and weaknesses alongside the tools that I want the students to learn to use. By explaining what the popular tool is best at doing, I can hint at its limitations and then fill the gaps with other resources designed for that purpose.

Principle 5: To choose a source effectively, you need to understand the information landscape.

The process of searching for information is not the primary goal of most chemistry students; they require the information that they seek in support of some outcome that they wish to produce. They want to run a reaction, write a paper, or understand the properties of the materials with which they will be working in lab. Therefore, they do not want to spend a great deal of time searching for new tools or information retrieval skills and are likely to attempt to use whatever tool is already familiar to answer whatever question needs addressing at the present time. This is not a sound information practice, given that an unfamiliar tool may more effectively and efficiently help the student to locate the needed information; fortunately, it is a practice that can be avoided by instructing them early in their careers about the breadth and organization of the chemical literature, as a whole. I do this at the very beginning my advanced undergraduate and graduate level classes, taking them through the chemical research process, from idea to publication, and showing where the various types of literature are most useful.

A group of chemists starts with a great idea, which they discuss with colleagues and refine through searching the literature and reading applicable papers. When the idea is fully developed, they develop some feasible methods and apply for funds. With funds in hands, the actual chemistry begins, but it rarely works the way that the group has anticipated the first time through. The researchers require additional information to explain the failures and optimize the successes until, at last, they are satisfied and ready to prepare a publication. They find themselves going back to the most basic of literature, placing the new research in its scientific context, and ensuring that all of their claims can be substantiated. The new manuscript, after moving through the scholarly publication process, becomes part of the cannon of literature, ready to inform someone else's new idea.

Depending at which point a researcher is in the research process, he or she is going to require different types of information. We discuss the fact that the information retrieval process is bookended by broad, tertiary, review literature

like books, encyclopedias, and review articles. These are the best places to go for background information and are used again when trying to contextualize the novel findings. During the methodology and experimental stages, more of the primary research articles come into play, supported by handbooks of substance and materials properties, spectral databases, and, more frequently, data deposition databases like the Protein Data Bank and PubChem. Throughout the process, secondary indexing and abstracting sources and search engines help the scientists to become aware of the publications that are relevant at any given point.

When teaching a short class in someone else's course, I generally present a list of sources that the students will find helpful in their quest for the type of information that they need to produce their deliverable. In my own, more extensive course, I spend the rest of the semester focusing on the tools that they will use for the remainder of their careers. By teaching a student to think about the stage of his or her research and the exact nature of the information needed, I hope that they will think more critically about the tools available to them and which is most likely to contain or direct them to the information that they require.

Teaching for Retention and Lifetime Learning: Pedagogical Techniques

Teach Relevant, Transferrable Skills, and Use Resources To Demonstrate Those Skills

I generally spend a good amount of time teaching students about search strategy in a vacuum from the sources. With current Web technologies, interfaces change rapidly, sometimes even overnight. In fact, these changes are so rapid and, occasionally, so subtle that I cannot count the number of times when I was surprised during class on Wednesday morning by changes to the resource that were not present on Tuesday afternoon when I prepared for the session. Teaching chemical information, clearly, is not for the faint of heart! Since the information tools are guaranteed to change at least once and perhaps as many as four or five times between my teaching the class and the students' using the tools in their research, I think that it is more important to teach them to "think like a database" than it is to teach them clicks and keystrokes in a particular tool.

Teaching students to think is a laudable goal, but urban legend has it that students will not do any activity unless they receive a grade for it, and they will not pay attention to any material that does not appear to have an immediate application. To be perfectly honest, I do not know if either of these statements is actually true, but I suspect that, for some students, both are solid facts. As a result, it is imperative to ensure that all of the techniques and thought processes being taught relate to material that the students are covering in class or address questions that they are asking in the course of their research. In many ways, it is easier to relate material to concrete "information needs" in a course-integrated instruction session where the information instruction usually supports a particular project or assignment than it is in a dedicated chemical information course. However, even within dedicated courses, one can align the techniques and examples to the students' interests. At the beginning of my semester-long graduate course, I have

the students list two or three scientists, whose work interests them, as well as giving me three to five topics of chemistry that they find exciting. I use these individuals, their research interests, and the research interests of the students to decide which particular skills to highlight in each class, as well as to formulate class examples and homework assignments.

Another good way to teach students to think is to present a resource that will be potentially useful for a task that the students need to accomplish in their class and use that tool to demonstrate a more abstract principle. For example, I will tell the class that the technique of the day is learning to use controlled vocabulary to broaden a search. We discuss what a controlled vocabulary is and how one can leverage it in information retrieval. I present a list of appropriate resources that include controlled vocabularies. Finally, we use one or two of the tools to perform searches using the built-in thesaurus or controlled vocabulary. These searches demonstrate the use and the limitations of these techniques. In later classes or sessions, when I need to introduce another tool that has a controlled vocabulary, I invoke the previously taught material and the resource previously demonstrated to draw parallels. In a full-semester course, I do this with most of the techniques that I teach.

- *Complex search syntax:* Web of Science, Scopus, keyword searching in the local catalog
- *Controlled vocabulary and thesauri:* Inspec, Compendex, MEDLINE, local catalog
- *Natural language searching:* SciFinder
- *Simple substructure and reaction searching:* SciFinder and Reaxys
- *Complex substructure searching and user-defined R-groups:* CAS REGISTRY via STN or Reaxys

The trade-off to employing a skill-based approach, rather than a resource-based approach is that it may not be possible to demonstrate all of the amazing capabilities of each resource taught. As information professionals, we are excited by the myriad of refinement, analysis, and export options available in our major indexing and abstracting tools, and we expect our students to be equally excited by them. It is not possible to highlight every feature, and the students can frequently find many of these options for themselves. I tend to provide the class with appropriate documentation for the tool demonstrated and rely on a group exercise or individual homework assignment to introduce additional options.

Keep It Small, Keep It Active, and Minimize Redundancy

The most tempting pitfall for a librarian, particularly when teaching an isolated lecture in someone else's class, is to attempt to teach the students too many things in a short time. When invited to be a guest lecturer, one's gut instinct is to think, "Oh, my goodness. I have forty-five minutes with these students, and I may never see them again! I need to teach them everything that they need to know for the rest of their careers!" This is not possible, and, as we have already discussed, it is doomed to fail if, indeed, students are not interested in learning

anything for which they cannot see a clear and immediate application. A careful conversation with the faculty member in charge of the course or program of study will help you to select the most critical skills to teach the students. From here, examine the tools that can be used to illustrate these skills. Your choice of tools will frequently suggest additional skills that can be added without overloading the students. For example, if I am demonstrating complex search strategy using Penn's online catalog, I display a search result and show the students the information contained. I draw particular attention to the LC Subject Headings that have been applied to the record and highlight the fact that the controlled terms are applied to all books on the same subject. This allows me to introduce index terms and subject headings *in situ*, and the students can leverage this skill to click across from a book of interest to other, relevant titles. I try to limit the number of "side trips" that I take, and I continually return to the point of the lesson, being sure to summarize everything that we have covered at the end so that the students are not overloaded by the end of the session.

The temptation to teach everything in forty-five minutes frequently spawns a second pitfall; in order to cover all of the desired material, one must dump the information into the students through the means of a lecture. Either of these problems alone is enough to lose the attention of a class, but when taken together, they are a recipe for disaster. I have moved away from a "lecturing" style to more of an "editorialized brainstorming" style for giving students large clumps of information. For example, rather than telling students that the three "universally" accepted Boolean operators are "and", "or", and "not," I will ask the students to tell me different ways of combining two terms. I ask them to explain the ways in which they function, and then I highlight the key points from the student's response and supply any additional information needed for full comprehension.

"Active learning" is currently in vogue in teaching information retrieval techniques. Using brainstorming sessions instead of lectures is one form of active learning; I have also found the use of controlled exercises to be helpful when there is time to use them. Brainstorming can be a very challenging technique to implement, however; an inactive class or a class that goes off on tangents can quickly derail an exercise. As a result, I find it important to give the students appropriate guidance before and during the session, while editorializing their contributions in such a way as to suggest additional material.

In my own course, I give a group assignment every week, which the students work during and immediately after class, and a homework assignment, which is worked individually prior to the next class. During the sessions that I do with undergraduate, organic chemistry students, I have the students work in pairs or trios to find information about an organic substance that I assign. I have found that, when the students discuss the information in groups after learning to use a source, it reinforces the concepts. I try to have all of my exercises structured around a lifelike scenario, along the lines of, "Imagine that you need to find information on X" or "Your boss has asked you to send a recommendation about whether or not you think your company has freedom to operate in area Y." Trying to tailor the scenarios to what the students in the class will actually need to do engages them further and gives them practice that is as close as possible to what they will see "in real life." However, when doing any kind of exercise, one must be careful to

provide prompt feedback. The longer it takes for me to correct the assignments or go over the techniques used in the group work, the more likely the students are to forget the material or, if they have practiced less optimal skills than the ones around which the assignment is structured, they are more likely to remember these less-effective methods.

Redundancy is a tricky problem, particularly when one is delivering most of his or her chemical information instruction as single lectures in other people's classes. It is always appropriate to remind students of skills that they have already learned, but one should avoid completely reteaching the material. A little redundancy helps students learn; if the majority of the material has already been introduced, however, the students will stop paying attention and miss the new information being presented. Tailoring the material to specific deliverables that the students will be producing through the course of the class can help to minimize redundancy, but it is not a sure-fire way of eliminating it all together. When possible, I recommend having one person coordinate, teach, or oversee course integrated instruction within a program. If all instructors record exactly what topics are being taught in each course, this information, coupled with an understanding of the progression of students through the courses in the curriculum will enable everyone to reinforce previously-taught concepts and be aware of the gaps in the students' education.

Avoid the “Ooh! Shiny!” Syndrome

Like all educators, librarians are always looking for “the next great idea” in information instruction and instructional technology. I have attended many conferences and meetings centered around topics like, “How can we use clicker technology to enhance our instruction sessions.” New technologies are exciting, and it is natural for individuals who read extensively on educational topics to think, “Hmmm. Can I use that in my classroom?” The problem comes when we discover that we are using technology strictly for the purpose of using technology. To avoid this pitfall, remember that the point of the instruction session is to teach the students to do something, not to entertain them with toys and gizmos. In order to learn a fact, the students need to connect something new with something that already lives in their brains; therefore, think about the best ways to make those connections. Once you have considered the subject matter and generated a list of learning outcomes that you want the students to achieve, then examine the existing technology and decide what would best enhance and promote student learning. In general, I would have to say that my favorite piece of teaching technology is a writing surface. I don't care what the surface is; it could be a whiteboard, blackboard, or tablet, or even a piece of paper, but it allows me to draw diagrams, make connections, fill in blanks, and stress things through repeated underlining and circling.

Students Are Confused about How They Can Appropriately Reuse Material

Students are taught, from a very early age, that plagiarism is wrong. In the social-media intensive world in which they live, however, I have detected

some misunderstandings of exactly what plagiarism is. Within social media sites, individuals share, repost, retweet, and generally repeat information that they have found particularly interesting or helpful. In some cases, it is easy to determine the original source of the information; in other cases, it has gone around so many times (“post this as your status if you support my cause...”) that the original source is lost in the mists of pixels. The multicultural aspects of university education in the United States also present challenges to students attempting to appropriately use and reuse previous work; intellectual property laws and norms vary in other countries, yet we expect our students to have already learned to conform to our requirements by the time they reach the university. One of the most frequent requests that my colleagues make when they ask me to teach in their classes is, “Please tell my students how to appropriately reuse material.”

I have discovered that it is wise to take nothing for granted when introducing the topic of appropriate reuse of others’ intellectual property. Instead of laying down the law, asserting the University of Pennsylvania’s code of academic conduct, and telling the students “How Not to do it (23),” I engage them in a conversation about why it is important to document sources. We discuss the fact that scientific rewards are based on a scientist’s impact on his or her field, which is measured by citations, as well as the fact that referring back to a published article will give the reader more detail about the research of interest. We close with the fact that a citation puts a certain distance between oneself and the information referenced; if the information turns out to be incorrect, a reader is directed to the original source of the fallacy, and the validity of one’s entire research project is not called into question. Only after discussing these reasons do we move on to the honesty side of the equation.

The concept of not being legally able to reproduce one’s own published information is another thing that does not resonate with the students. This topic usually arises in graduate classes and with students preparing their dissertation. As a result, we have started introducing appropriate mechanisms for reuse of published material, including requests for reuse and appropriate documentation within one’s article or lab report, to undergraduate students, as well. When explaining to doctoral students the reasons that they cannot use their own figures or journal articles in a dissertation without checking if permission is required, we highlight the fact that, in chemistry, one is obligated to publish material in only one place, and that the student may very well have transferred copyright for the article to the publisher, meaning that the article is no longer theirs to use as they would like.

Essential Information Skills All Chemists Should Learn

Every chemist, regardless of subdiscipline, should be able to search effectively using text and using structure. As a result, I teach these skills from an early stage, often starting in the upper-level undergraduate curriculum. Text searching can take a variety of forms, including searching using words, formulae, and numbers. For example, most students are unfamiliar with the Hill Order for writing molecular formulae, in which all like atoms in the formula are grouped

together and are ordered in the following way: for an organic substance, carbon comes first, then hydrogen, and then all other atoms in alphabetical order; in an inorganic substance, all atoms are listed in straight alphabetical order. The fact that one can search in a numeric field using a range of values or even an inequality also takes many young chemists by surprise. I introduce formula and basic structure searching in the organic chemistry laboratory classes, and I teach students to profile substances by property ranges in the advanced undergraduate and graduate classes; however, I find it critical to begin to teach good “word” searching techniques at the freshman level and to introduce substructure searching before the end of the undergraduate curriculum. The following two exercises have proven to be beneficial to the students’ understanding of those two concepts.

Deconstructing and Reconstructing a Textual Query

This skill deals with the ability of a student to translate their information need into a search statement that can be used in a tool of choice. We go through a four-step process to convert a vague desire for information into something that a database will be able to use to retrieve information.

State the Information Need

Although this seems like a very obvious and easy step, it is often the part with which students struggle the most because their tendency is to ask the question that they think can be answered, rather than the question that they actually want answered. To get them thinking along the correct lines, I sometimes ask them to address their question to some all-knowing entity like the computer in Star Trek. An example that I use with my graduate class is, “Computer, give me research articles about the chemistry behind the flavor of beer.”

A well-articulated query gives you two things. First, it presents a set of concepts that, when combined, will give you highly relevant results. However, it should also present some context for the query and the desired information. In the example above, we know, not only do the students want to know about beer flavor, but they are looking for primary literature dealing with this topic. Explicitly stating the context at the outset can help a student settle on a source more quickly and efficiently.

Break the Query into Constituent Concepts

Once the students have clearly articulated what they need to know, I have them begin to break down the query into its constituent concepts. I define a “concept” as any component of the query that could form a stand-alone search term. When searching for the chemistry behind beer flavor, the three basic concepts to be combined are “chemistry”, “flavor”, and “beer.”

This is also challenging for the students to do, and I usually have to prompt them by telling them how many concepts I have in mind and, occasionally, by

suggesting one of them. To them, the query only deals with one concept, the chemistry of beer flavor. In order to perform an exhaustive search, however, this is a very necessary step.

Brainstorm a List of Terms That Could Be Used To Describe Each Concept

At this point in the class, I generally lean on the first and second principles of finding information: you are searching for material that has been published (in other words, it exists), but, in order to retrieve it, your search statement needs to match the way in which it is described in the article, its index terms, or its abstracting. I divide the board into columns, and, at the head of each column, I write one of the concepts that the students have defined from their query. I then encourage the students to play, “psych out the author” and brainstorm all possible terms that could be used to describe or indicate each concept. This allows us to talk about the scope of the search, as well as thinking about the fact that different authors and indexers may use different terminology to describe the same concept. We discuss methods of changing vocabulary selected to broaden or narrow the search, as well as seeking potential synonyms for our search terms. In the beer query, if one wanted to find books that include information about beer flavor, it might be useful to include the search term “alcoholic beverages.” Because books tend to be more general in nature and are indexed using rather broad subject headings, a book about alcoholic beverages may, indeed, contain a chapter related to beer and its flavor. Since the students are looking for research articles, using such a broad search term may or may not be advisable. In some cases, the students end up subdividing search terms, giving them a rather complex tree of interchangeable terms. For example, beer could be described as a “fermented beverage” or a “malt beverage;” a beverage could be described as a “beverage” or a “drink;” there are many forms of the word “fermented;” etc.

Employ the List of Terms To Generate a Sound Query, Using Appropriate Syntax for the Search Tool of Choice

This is a two-step process, and it goes back to using the “Library Equation” to select an appropriate source.

- Determine the source to be used
- Evaluate all potential search terms in light of the chosen source.

Once the source is selected, the searcher should examine its indices, search capabilities, and help or about files to determine the best syntax to use. Employing appropriate Boolean logic and truncation can be key in an electronic tool; understanding the controlled vocabulary and indexing can also be helpful in attempting to select the best possible search terms for the resource.

Deconstructing Molecules To Build Substructures

Substructure searching is even more challenging for students to grasp, in part due to the fact that organic chemists use so much short-hand to describe the structures of their molecules. When students see a six-membered carbon ring, they automatically assume that it is cyclohexane. They do not realize that, in substructure mode, a database sees each carbon as being connected only to two other carbon atoms, and it will therefore retrieve molecules with any kind of substitution, provided the backbone ring system is present as drawn.

To teach substructure searching at the undergraduate level, I use the visual of a molecular template being drawn on a sheet of glass. The glass is held over every structure in the database, and, if it exactly overlaps with a structure, that substance is retrieved as a hit. Any substance whose structure deviates from the template is discarded. At the graduate level, I use this visual as a starting point, but then I introduce the idea of connection tables, explaining how the computer interprets a structure as a series of connections between nodes, and, in order for a substance to be retrieved, its connection table must contain the connections requested in the substructure.

Once students are comfortable with exactly what a substructure search is and why they might want to perform one, I next encourage them to deconstruct and reconstruct the molecule in much the same way they deconstruct and reconstruct a textual query (28). When designing their substructures, I remind them to balance the amount of time they spend designing and inputting the query with the amount of time they want to spend examining their results; for a comprehensive query, I recommend a more general substructure with more result management, while for a directed query, I recommend a highly tailored substructure.

Draw the Core of the Molecule

The molecule's core should be the section that interests the student the most, chemically. I encourage the students to select a core that is small enough to get them what they want, while being large enough to prohibit unwanted hits from being retrieved. In some cases, I recommend that they think about whether they would be interested in answers containing more than one core, and, if so, to draw both. It may be possible at a later stage to merge the two, depending on the capabilities of the search system that they wish to use.

Ask Questions about Atoms

After determining the core, I ask the students to begin asking questions about atoms. First, I tell them to examine all the atoms in the core and decide whether or not they would like to permit variance at any position. Then, I tell them to look at the places where the atoms do not exhibit a fully-satisfied valence and ask several questions. First, can the site be further substituted? If it cannot, they should draw

hydrogen, and if it can, they must ask themselves how. If it can be substituted by any atom, with no restrictions on atom identity or bond order or topology, they should leave it alone. If there are restrictions on either the allowed substituents or the ways in which they may connect to the parent molecule, they will need to use variables or R-groups, combined with bond variability, to describe that which is allowed.

Ask Questions about Bonds

The next series of questions deals with bond order and the general shape of the molecule. Again, I first direct their attention to the core of the molecule, asking them if there is any variability of bond order that will be permitted. I next introduce the concept of topology, asking whether an atom or bond is permitted to be part of a ring or a chain. Looking at the bond orders and topologies of atoms and bonds that are drawn is relatively simple. The problems arise when one attempts to specify bond orders and topologies of undrawn substituents. I teach the students tricks using R-groups and variables; for example, we construct an R-group consisting of hydrogen or A, any atom but hydrogen, and attach it to a given position with a topologically-specified bond, allowing any atom on the periodic table to appear at that position but restricting the topology of the connection (28)^c. Depending on the level of the class, these R-groups become more and more complicated. For undergraduate students, I find it sufficient to introduce the concept of topology and leave it at that, while for synthetic chemistry graduate students, we concoct complex searches that restrict topology in some sites of a molecule, while allowing others open to a variety of configurations.

Construct an Appropriate Query Using the Tools Available in the Database of Choice

The last part of any substructure class is the “point and click” section. While I generally attempt to avoid this type of instruction, I find it impossible to do so when teaching substructures. Most of the major structure databases have different structure editors, and, since minimalism is in vogue at the moment, their “simple” user interfaces hide or obscure some of the most helpful features. Therefore, after leading the students through the molecular analysis questions listed above, I take them through one or more example searches using the “resource of the day,” usually a tool that I have chosen to demonstrate the specific substructure techniques that I wish to demonstrate (SciFinder if I have a set that I wish to analyze in various ways, and Reaxys to demonstrate the use of user-defined R-groups, to name to examples).

Classroom Activities Demonstrating General Principles and Pedagogical Techniques

The following classroom activities demonstrate one or more of the general principles and techniques described in the previous two sections. “What do chemists read, what do they write, and what should they believe”, “Using the Library Equation”, and the activities dealing with substance identifiers and substructures are more specific to chemical information education, while “Monty Python and the Search for Information”, “Ask the *Star Trek* Computer”, and “Deconstructing and Reconstructing a Text Query” can easily be applied to information sessions in other disciplines, and we at Penn have used these successfully with a variety of different engineering classes, as well as in chemistry courses. These can serve as a starting point for developing additional classroom activities and lessons, bearing in mind the need to keep students engaged through brainstorming, activities, and limited redundancy.

What Do Chemists Read, What Do They Write, and What Should They Believe?

This is an exercise that I employ in classes of all levels, from freshmen through graduate students. I challenge the students to brainstorm a list of ways that scientists communicate with one another or learn about developments in their fields. In some classes, this is the second in a two-part discussion of the chemical literature as a whole, which begins with the literature in the context of research in the chemical sciences. As they are listing the different methods of communicating, I group them in three columns according to the level by which the information is removed from the primary laboratory work. We end up with a column of primary sources (journal articles, conference papers, patents, reports, e-mail, etc.), a column of secondary indexing and abstracting sources and catalogs, and a column of tertiary, review literature.

The second step is to create a group of criteria to evaluate whether a piece of information is worth reading, and it involves a second brainstorming exercise. At the end of the exercise, we have a list of criteria that attempt to gauge the accuracy, relevance, authority, and objectivity of the information (24). After the students have finished constructing their list, we discuss the relative strengths and weaknesses of each criterion as a method of evaluating information and attempt to place them in context with one another so that the students can use a collection of criteria together to gauge the overall worth of a source. The students tend to be engaged in this exercise, and it is relatively easy to solicit class participation, so, it is generally a popular class.

We then return to the lists of narrative literature, mainly in the primary and tertiary columns (we discuss the evaluation of secondary sources separately) and apply our newly-acquired methods of evaluating information. Based on what we learn, we determine which of the sources that the students proposed are the most timely and which tend to be more or less reliable than the others. At the end of the day, however, the students do not end up with a list of sources or reference types that are “good” and a list of sources that are “bad;” instead, they learn that all information sources could potentially contain useful information. By applying the various evaluation criteria, they learn to prioritize their reading in the same way that examining the coverage and search features of two databases allows them to prioritize their searching. After all, the fact that an article appears in a reputable journal and is highly cited does not make the information that it contains correct, and, despite the fact that an article is incorrect, if others are spending a great deal of time discussing it, one should probably know exactly what it says.

Monty Python and the Search for Information

This exercise encourages students to think about the information need itself. For cheap laughs, I began using a Monty Python analogy in an engineering ethics class when teaching information searching and evaluation. Before crossing the bridge in *Monty Python and the Holy Grail*, Arthur had to answer three questions: “What is your name?”, “What is your quest?”, and “What is the wing-speed velocity of a common swallow (25)?” I encourage my students to ask three questions, as well:

What is your aim?

Articulate your end goal. This should include the purpose for which you need the information. For example, “I believe that the following substance will be a good catalyst for a particular type of reaction. I wish to make and test the catalyst, and I plan to apply for an NSF grant to fund this research.” The end goal will help you to determine the degree of specificity and comprehensiveness needed in your searches, as well as a potential audience for any resulting documents or publications.

What is your quest?

Determine the specific piece of information that you need at this stage in your process. In other words, answer the question, “For what am I searching right now?” For example, someone with the goal of making and testing a catalyst may want to locate some synthetic preparations of similar molecules and some articles describing their activity so that they can develop a plausible experimental plan and convince a grant review board to fund the research.

What tools exist to help you with this search?

In addition to identifying the information resources to be used in each quest associated with the overall aim, students should learn to identify the search techniques that can be used to greatest effect. In fact, identifying the search techniques and requirements ahead of time can help a savvy searcher select a tool from the ever-increasing supply of information resources.

Ask the Star Trek Computer

In a graduate program in library and information science, aspiring librarians are taught to take an extremely general query and make it more specific with each succeeding search. By casting a wide net and drawing it in, one hopes to prevent any potentially useful information from escaping. However, the average student may not have the need or the patience to sift through every single hit that could potentially be useful; what he or she requires is the one piece of information that exactly fits the bill. I use the *Star Trek* computer analogy in conjunction with the Monty Python exercise described above when teaching students to determine the scope of a query and the degree of comprehensiveness required.

After a student has determined his or her end goal, I encourage him to think about whether, to achieve that goal, he requires a comprehensive or a specific search. If the end goal is very broad (“I want to ascertain that no other researcher has employed this methodology to make this type of substance.”), I recommend the traditional, iterative method of searching for information. If, however, the end goal is more specific (“I need to remove this protecting group in the presence of a sensitive functional group.”), I suggest the following. “Imagine that you are addressing the computer from *Star Trek*. This computer has a limitless databank and can tell you anything that you want to know. It is intelligent; you talk to the *Star Trek* computer the same way that you would talk to a person, and it processes your request the same way that the human brain would process it. How would you ask it to find what you want?”

By thinking in terms of the way in which she would articulate a query to an actual person with limitless data at his or her disposal, the student learns to think, not in terms of the question that she thinks can be answered, but in terms of the question that she actually wants answered. Rather than focusing on the mechanics of entering search terms into a database, she is forced to determine what information she needs without confining herself to those functions that a specific tool can provide. Articulating the query in “plain English” can then lead her to think about the possible resources that might contain the desired information. Only after she has identified the best possible tool should she begin to consider the degree of specificity that needs to be built into the query within that tool.

I have also discovered that, since implementing this visual in my graduate class, students who come in for reference assistance are more likely to give me the background of their query. This helps me to contextualize the information need, and I find that I am more quickly able to suggest a source and a search strategy that will be effective in answering the question that the students need answered.

Use the “Library Equation” To Transform Known Information into Desired Information

Students of chemistry are accustomed to seeing balanced equations and reaction schemata, so, I incorporate this visual into my teaching of source selection. The search for information can be likened to a chemical process in which one transforms a piece of known information to a piece of desired information by acting upon it using a source (Figure 1) (26).

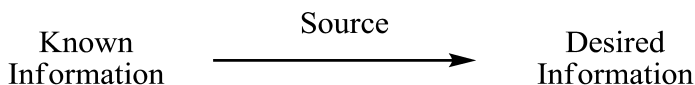


Figure 1. The “Library Equation,” by which a resource converts known information into desired information. Adapted with permission from Finding physical and chemical properties of substances: A myriad of access points, presented as part of Chemical Information Sources, Requests, and Reference at the SLA Annual Conference. This image is protected by a Creative Commons License and is used with permission of the author (26b).

For greatest effect, the source should be indexed or searchable by the known information and should contain the desired information. I use this example most frequently when teaching organic chemistry students how to locate physical properties of substances. I tell them to assume that they have the name of a substance and they wish to locate its melting point. They should select a resource that is searchable by name and that contains melting points.

There are many cases, however, in which this simplified approach to finding chemical information breaks down. Perhaps one has a structure and the best source of the desired information is a printed book. In the example above, if the substance name that the student knows does not appear in the synonym list of its record, the student will not be able to retrieve its physical properties. In cases like these, it may be necessary to daisy-chain sources, finding an intermediate piece of information that can then be used in the source that contains the target, and yielding the following reaction schema (Figure 2).

For example, if searching for the known name of a substance yields no hits in the desired reference source, one could use a different source to locate its formula or CAS REGISTRY Number, which one can then search in the original reference source.

This approach makes the class chuckle or roll its collective eyes, but I have found it to be very effective when thinking about information retrieval, particularly when faced with a group of novices. Too frequently, I have seen students who know how to use a single source well, who know that it includes the information that they want, and who are completely flummoxed by the fact that they cannot find what they need after the first search. The fact that substance identifiers are not always conserved from source to source makes its use in finding physical and chemical properties critical, and the students generally find it helpful in practice.

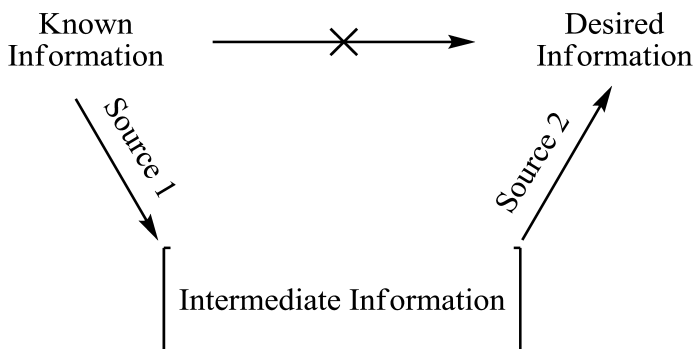


Figure 2. An enhanced version of the “Library Equation,” in which the first source is used to locate information that can be used in a second source to locate the desired information. Adapted with permission from *Finding physical and chemical properties of substances: A myriad of access points, presented as part of Chemical Information Sources, Requests, and Reference at the SLA Annual Conference*. This image is protected by a Creative Commons License and is used with permission of the author (26b).

Identifying Substances for Information Retrieval

Thinking about substances and their properties is frequently challenging for students. I recently had a conversation with a group of graduate students in my chemical information course about a homework question that I had assigned. The question asked them to find some physical properties information of a blue topaz, and they were convinced that I had not given them enough information to answer the question. I asked them what “topaz” was, and they told me it was a gemstone, it was blue, it was attractive, it was a mineral, etc. I tried again, asking them what the word “topaz” was, and they told me it was a noun. After several further attempts involving a variety of substances, they finally realized that topaz was the name of the substance, that chemical name was an available search field in the database they had chosen, and that they were able to find the requisite properties without any additional information.

Before teaching students to locate physical properties, I engage them in a brainstorming session about all of the ways that they can identify substances for information retrieval. I fish for the following identifiers (27):

- Chemical name
- Molecular formula
- CAS Registry Number
- Structure
- Properties
- Spectra

The students frequently come up with other items on the list, including functional groups or units, potential uses, reactivity, etc. Name and formula

usually come up quickly, but only about half of the undergraduate classes I have taught were familiar with CAS REGISTRY Numbers. In order to help the students along when they get stuck, I ask them how they would identify a person. A photograph or picture equates to the substance's structure. Physical description brings up some of the properties. Spectra can be likened to fingerprints, and CAS REGISTRY Numbers to an individual's Social Security Number.

Once we have a list of all of the identifiers, I break them into two groups. Leaving out spectra, which can be tricky to classify, I characterize the first four (name, formula, CAS RN, and structure) as "identifiers," to be used when trying to find information on a specific substance, and the remainder, I classify as "profilers," or things that can be used to locate a collection of similar substances. We spend the rest of the activity focusing on the four identifiers, discussing the strengths and weaknesses of using each as an entry point into the literature and describing the way that one can daisy-chain resources in order to start with the only identifier that you know (topaz, for example) and end up with the information that you desire.

Conclusion

At the end of the day, there are three things that I hope I have taught every student. The first thing is that the chemical literature has a specific organization and specialized entry points and that, although it is possible to find some information without learning to use them, one is much more efficient and effective if one spends a little bit of time learning "the rules." Second, I encourage the students to think about the question that they actually want answered, not the question that they think can be answered. Finally, I remind them of the "twenty-minute rule;" because the literature of chemistry is complex, it can be difficult to locate information quickly. If they are still having problems finding what they need after twenty minutes of searching, they should consider asking for help. Librarians simply love helping students with searches, and our doors are always open.

I am, however, fooling myself if I think that much has changed. When reading Jahoda's description of some of the comments received in his 1953 survey, I was struck by the following: "The majority of schools elaborating on their course explained that it consisted of lectures followed up by written assignments, application of course material in the library. Students in one school complained that they were overworked for a two-credit course" (6). Although I have moved to a curriculum that focuses on skills and uses resources to teach these skills, I still teach students the skills through lecture, directed questioning, brainstorming, and in-class examples and exercises and reinforce the material through group and individual assignments that force them to practice what they have learned. And, even if instructors' methods have changed subtly in the past sixty years, student attitudes have not. I believe that I saw some very similar comments about workload on my own course evaluations in the past five years. Truly, much has changed, but more has stayed constant!

References

1. Tenopir, C.; King, D. W. Reading behaviour and electronic journals. *Learned Publishing* **2002**, *15*, 259–265.
2. American Chemical Society Committee on Professional Training. *Undergraduate Professional Education in Chemistry: ACS Guidelines and Evaluation Procedures for Bachelor's Degree Programs*; American Chemical Society: Washington, DC, 2008; p 30. <http://www.acs.org/content/dam/acsorg/about/governance/committees/training/acsapproved/degreeprogram/2008-ac-s-guidelines-for-bachelors-degree-programs.pdf> (accessed November 19, 2013).
3. American Chemical Society Committee on Professional Training. *Chemical Information Skills*. <http://www.acs.org/content/dam/acsorg/about/governance/committees/training/acsapproved/degreeprogram/chemical-information-skills.pdf> (accessed November 21).
4. Special Libraries Association Chemistry Division; American Chemical Society Division of Chemical Information. *Information Competencies for Chemistry Undergraduates: the elements of information literacy*, 2nd ed.; 2011. <http://chemistry.sla.org/wp-content/uploads/cheminfolit.pdf> (accessed November 21, 2013).
5. Somerville, A. N. Chemical information instruction of the undergraduate: a review and analysis. *J. Chem. Inf. Comput. Sci.* **1985**, *25* (3), 314–23.
6. Jahoda, G. University instruction in chemical literature. *J. Chem. Educ.* **1953**, *30* (5), 245.
7. Soule, B. A. Book ability. *J. Chem. Educ.* **1932**, *9* (11), 1940.
8. Somerville, A. N. Perspectives and criteria for chemical information instruction. *J. Chem. Inf. Comput. Sci.* **1990**, *30* (2), 177–81.
9. Somerville, A. N. Chemical Information Instruction in Academe: Recent and Current Trends. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 1024–1030.
10. Garritano, J. R.; Culp, F. B. Chemical information instruction in academe: who is leading the charge? *J. Chem. Educ.* **2010**, *87* (3), 340–344.
11. (a) Carr, C. Teaching and using chemical information: an updated bibliography. *J. Chem. Educ.* **1993**, *70* (9), 719–726. (b) Carr, C. Teaching and using chemical information: annotated bibliography, 1993–1998. *J. Chem. Educ.* **2000**, *77* (3), 412–422.
12. Chemical Information Sources. http://en.wikibooks.org/wiki/Chemical_Information_Sources (accessed January 8, 2014).
13. XCITR: eXplore Chemical Information Teaching Resources. <http://www.xcitr.org/> (accessed January 9, 2014).
14. Searching for information on this topic can be challenging, due to the many ways in which the words “chemical information” and “libraries” can be used in the chemical enterprise. The following searches proved most successful, although the results required extensive filtering and review. SciFinder: chemical information (literature searching) of instruction (teaching) (education); Web of Science: Topic=((“chemical information” or “chemical bibliography” or (chemi* SAME literature SAME search*)) AND (instruct* OR teach* OR educat*)). Once one locates some interesting

articles, citation pearl growing and cited reference searching can yield additional results.

15. Twiss-Brooks, A. University of Chicago, Chicago, IL. Personal communication, 2014.
16. Ricker, A. S.; Thompson, R. Q. Teaching chemical information in a liberal arts curriculum. *J. Chem. Educ.* **1999**, *76* (11), 1590–1593.
17. Pence, L. E. A chemical information assignment for nonscience majors. *J. Chem. Educ.* **2004**, *81* (5), 764–768.
18. Locknar, A.; Mitchell, R.; Rankin, J.; Sadoway, D. R. Integration of Information Literacy Components into a Large First-Year Lecture-Based Chemistry Course. *J. Chem. Educ.* **2012**, *89* (4), 487–491.
19. Somerville, A. N.; Cardinal, S. K. An integrated chemical information instruction program. *J. Chem. Educ.* **2003**, *80* (5), 574–579.
20. Walczak, M. M.; Jackson, P. T. Incorporating Information Literacy Skills into Analytical Chemistry: An Evolutionary Step. *J. Chem. Educ.* **2007**, *84* (8), 1385.
21. American Chemical Society Committee on Professional Training. *Proposed Changes to the ACS Guidelines and Evaluation Procedures for Bachelor's Degree Programs*, January 2013. <http://www.acs.org/content/dam/acsorg/about/governance/committees/training/guidelines-white-paper.pdf> (accessed November 22, 2013).
22. Rowling, J. K. *Harry Potter and the Chamber of Secrets*; Scholastic: New York, 2000.
23. Dickens, C. *Little Dorrit*; Oxford University Press: New York, 1983.
24. Currano, J. N. Hunting and gathering: Locating information on the cusp between science and legislation. Presented at the 244th ACS National Meeting & Exposition, Philadelphia, PA, United States, August 19–23, 2012; CINF-11.
25. Gilliam, T.; Jones, T., dir. *Monty Python and the Holy Grail*; Columbia TriStar Pictures, 1974.
26. (a) Currano, J. N. Designing an effective training session. Presented at Teaching Chemical Information, ACS National Meeting, 2007. (b) Currano, J. N. Finding physical and chemical properties of substances: A myriad of access points, in *Chemical Information Sources, Requests, and Reference*. Presented at SLA Annual Conference, San Diego, CA, United States, June 7, 2013.
27. Currano, J. N. Qualitative analysis in the library. Presented at 239th ACS National Meeting, San Francisco, CA, United States, March 21–25, 2010; CINF-4.
28. (a) Currano, J. N. Think like a database: Substructure searching in the classroom. Presented at 232nd ACS National Meeting, San Francisco, CA, United States, Sept. 10–14, 2006; CINF-037. (b) Currano, J. N. Deconstructing molecules in an organic information course. Presented at 233rd ACS National Meeting, Chicago, IL, United States, March 25–29, 2007; CINF-087; (c) Currano, J. N. Searching by structure and substructure. In *Chemical Information for Chemists: A Primer*; Currano, J. N., Roth, D. L., Eds.; Royal Society of Chemistry: Cambridge, 2014; pp 109–145.

Chapter 12

Public Chemical Databases and the Semantic Web

Martin A. Walker*

**Department of Chemistry, State University of New York at Potsdam,
Potsdam, New York 13676**

***E-mail: walkerma@potsdam.edu**

Progress in chemistry is increasingly data-driven, and there is a growing need for the public databases reviewed in this chapter. As we move towards a "semantic web" where our chemical information is data-rich, open sharing via public databases will become an intrinsic part of our work.

Introduction

We live at the beginning of an "Information Age" that is changing how we conduct science and science education. We now have chemical papers in our hard drives, rather than in filing cabinets or on library shelves. Chemical data are now often available in machine-readable forms, ready to be processed in our computers. Our students are able to access far more information, within seconds, than students a generation ago.

However, we are failing to reap the full benefit of the World Wide Web for sharing information. In principle we now have access to much of the world's chemical knowledge; in practice, much of that knowledge is either buried inside unreadable files or hidden behind subscription paywalls. Authors submit data-rich manuscripts to journals, but little of that remains machine-readable after publication, and the chemical meaning is mostly lost. Our governments can (apparently) decipher terrorist intent in cellphone records and emails; our online suppliers can read our buying and viewing preferences; yet the data from our professional work remain inaccessible, even to most scientists. This hurts the scientific enterprise, since it means that most results are never critically evaluated or even curated, and research is often unnecessarily duplicated.

Another issue is that of access. Our traditional model for sharing scientific information was designed to serve a small community of elite scientists. Yet today there are many more stakeholders demanding access to that information: Educators and their students, technology-based companies, and even the general public. All of these groups provide the infrastructure that the scientific community needs in order to work – well-educated workers, new technology and taxpayer funding – but they are shut out from much of the scientific knowledge that is generated. At the same time as the Internet is making information more accessible, many journal subscriptions and databases are pricing themselves beyond the reach of all except the elite universities and major corporations.

The World-Wide Web offers many opportunities that are overlooked at present. We can make scientific data available *en masse* for our students to use, analyze and contribute to, so they can be better trained for their future careers. We can find fraud and error in our data, and thereby improve the reliability of those data. We can analyze information in new ways, even bringing in data from outside our speciality, in order to make new connections that may lead to breakthroughs. We can make science more transparent, allowing the educated public to see for themselves the raw data behind drug claims, as well as toxicological data and health information.

This chapter will describe the current state of public chemical electronic databases, and how those databases serve the chemical community and the wider public. It will also show how the data in those databases can be used to build a truly accessible, readable and interconnected collection of information embedded with chemical meaning – a “semantic web” – that will allow chemistry to flourish in the future.

A Brief History of Chemical Data

When chemists began to publish their discoveries openly and systematically, during the Age of Enlightenment, they started a revolution that is still underway today. Science began to require that experiments be reproducible, and empirical data freely shared.

In the modern era, the amount of scientific information has grown immensely, but computers have given us a way to handle that information, and the Internet allows it to be shared rapidly. However, our systems of managing results and data are still rooted in a print publication model, which must evolve if we are to take full advantage of the electronic tools we have available.

Databases offer a way for scientists to work with collections of data, to observe larger trends or to “see the big picture.” Ideally, the database should be compiled directly and comprehensively from the literature, adding new information at the moment of publication, but this ideal has rarely been achieved in practice, except behind a publisher’s paywall. Extracting the data after publication via text mining, structure recognition, etc., is clearly less than optimal, as it is inefficient and it leads to database errors.

At present, it is often necessary for data to be extracted from the literature manually, an approach that requires a great deal of work by technically literate

people. The data may be numeric or text-based in nature, or they may relate to chemical structures or processes. Supplementary material is provided with many journal articles without charge (even when the article itself has a cost), but the data are usually in a PDF file that is not easily read by machine. Scientific data cannot be copyrighted, but publishers have been accused of trying to control access to raw data held in their publications (1). It should be noted, however, that some commercial cheminformatics companies such as Accelrys/DiscoveryGate and Leadscope do contribute data to public databases such as PubChem.

Why Public?

Traditionalists may argue that a network of proprietary databases can serve the community well, and provide a source of income for maintaining and expanding content. Chemical Abstracts Service (CAS) is one well-known example of an information provider with a broad collection of subscription databases. Aside from abstracts of chemistry documents, the service also carries databases of chemical structures, reactions and patent information. This operation requires a large labor force to maintain the high quality and comprehensiveness that CAS is known for, and this must be paid for; also, surpluses from CAS are also used to fund other ACS programs and activities. In the case of commercial information services such as Reaxys (Elsevier), which boasts a vast amount of experimental data, the income from subscription databases is also used to generate profits for shareholders. Databases such as those of CAS and Reaxys (Elsevier) are impressive, but their closed nature limits their value to the scientific community (2). (For further comparison of CAS files and Reaxys, cite chapters by Buntrock and Rusch in this volume.)

Stewart Brand famously once said that “*information wants to be free*” (3). Rudy Baum, then editor at *Chemical & Engineering News*, commented on this in 2006: “*I don’t actually believe that information wants to be free. I don’t think information gives a damn. I think cheapskates want information to be free*” (4). Not only does this conflate free (no cost) with free (*libre* or unfettered), it also completely misses the point. The value of information depends on it being shared. It is quite true that people – including chemists – are naturally attracted towards sites that are free (no cost). This explains why the Alexa ranking of ChemSpider (free) is similar to Chemical Abstracts Service (subscription only) (5), despite the fact that the latter offers a much more comprehensive service.

The central point is that scientists need free (*libre*) and open access to complete data in order to make full use of those data. The community needs to flag dubious results and to confirm new findings; this may be done directly through community curation (as on Wikipedia or ChemSpider) or indirectly through traditional publication. Data must be evaluated, compared and re-used to make new connections and discoveries. Complete access to data ensures that the right connections are made, important results are not missed, and negative results are not “buried.”

In the same editorial Baum argues that “*those who create valuable information deserve to be compensated for it.*” For the scientists who create data the best

compensation is wider recognition, which happens most easily when the data are traceable and freely available to all. Public databases also guarantee that the results of research can be accessed by the public, whose taxes often paid for the work.

The principal aim of a public database, as defined here, is to serve the needs of the public rather than to earn a profit; as such it is committed to openness and free (unfettered) use of its data. The “public” in this context may mean the specialist chemical community (academic, industrial and government) or the general public seeking information about common household products. A public database does not simply serve chemists on a limited budget, but it also accumulates a mass of open data that benefits the entire community.

The Present: An Overview of Public Chemical Databases

Government has a clear interest in meeting the needs of the public, and in sharing data from research at government laboratories or sponsored by government. US government databases (such as those shown in Table I) have become important resources, and European governments are now working together to develop databases such as the “Chemical Entities of Biological Interest” (ChEBI) database. Professional *chemistry organizations* such as RSC serve the community with open databases such as ChemSpider. Some databases are started and maintained by *academics*, such as ZINC, which serves the medicinal chemistry and biochemistry communities. Finally, Wikipedia is an encyclopedia operated by and for the *general public* via a nonprofit organization, and its editors include several chemists who have built up what is equivalent to a small database of chemical information. All of these types of databases are represented in Table I.

Table I. Summary of selected public chemical databases (as of Oct. 2013)

<i>Database</i>	<i>Type</i>	<i>Record type</i>	<i>No. of records</i>
PubMed http://www.ncbi.nlm.nih.gov/pubmed/	Govt.	Biomedical literature refs.	23 million
PubMedCentral http://www.ncbi.nlm.nih.gov/pmc/	Govt.	Biomedical papers full text	2.7 million
Europe PubMedCentral http://europepmc.org/	Govt.	Biomedical abstracts + papers	28 million abst + 2.6 million full text papers
PubMedCentral Canada http://pubmedcentralcanada.ca/pmcc/	Govt.	Biomedical papers full text	2.6 million
PubChem main substance database http://www.ncbi.nlm.nih.gov/pcsubstance	Govt.	Substances	> 100 million

Continued on next page.

Table I. (Continued). Summary of selected public chemical databases (as of Oct. 2013)

<i>Database</i>	<i>Type</i>	<i>Record type</i>	<i>No. of records</i>
PubChem compound database http://www.ncbi.nlm.nih.gov/pccompound	Govt.	Validated structures	47 million
PubChem bioassay database http://www.ncbi.nlm.nih.gov/pccassay	Govt.	Bioassays	> 200 million outcomes
ChemSpider http://www.chemspider.com/	Soc.	Substance	>29 million
NIST Web Book http://webbook.nist.gov/	Govt.	Substances	>16000 (IR, ion energetics) >33000 (MS)
ChEBI https://www.ebi.ac.uk/chebi/	Govt.	Substance ontologies	>35000
ChEMBL https://www.ebi.ac.uk/chembl/	Govt.	Biomedical data	1.3 million
CAS Common Chemistry http://www.commonchemistry.org/	Soc.	Substances	8000
Wikipedia http://www.wikipedia.org/	Non-profit	Encyclopedia articles	~10000 substances, also other topics
EPA ACToR (environmental & toxicity data) http://actor.epa.gov/actor/	Govt.	Substances	> 500,000
HSDB http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?HSDB	Govt.	Toxicology data	~5000 substances
NIOSH Pocket Guide (hazard info) http://www.cdc.gov/niosh/npg/	Govt.	Substances	670 + access to other CDC data
ZINC (drug screening) http://zinc.docking.org/	Acad.	3D structures	>21 million
DrugBank http://www.drugbank.ca/	Acad.	Drugs	6811
Crystallography Open Database (COD) http://www.crystallography.net/	Govt/ Acad	Crystal data	~250,000

This list of databases is by no means comprehensive, but it includes most of the larger and more sophisticated databases available in late 2013.

Merely collecting data is not enough; it is also important to ensure that those data are correct. For public databases, this is done through curation (updating and fixing errors) and validation (a formal procedure designed to verify that certain data

are correct). “Crowdsourced” resources such as Wikipedia are heavily curated, and these can also carry out validation efforts. Machine-built resources such as PubChem depend more on machine-curation (for example, checking that carbon atoms do not have more than five bonds), though some such as ChemSpider do also use community curation.

The International Chemical Identifier (InChI) as an Open Standard

The development of the International Chemical Identifier (InChI) and the related InChIKey has been vital for public chemical databases, by providing an open standard for the machine representation of chemical structures (6). Unlike SMILES and other earlier systems, InChI is clearly non-proprietary (7). It was initiated and then endorsed by IUPAC and NIH; it also has a clear system of versions that ensures that all databases can represent the same structure with the same unique identifier. InChI began in 2005, and it has since become a popular method for representing organic structures, with many structure drawing tools now including InChI generation as a standard feature. Now often include a “search the internet” feature, where a structure can be selected and this triggers a Google search of the InChIKey. The InChIKey is a hashcode version of the InChI that is more suitable for use in web search engines (7). With many important chemistry sites using the InChIKey, the results of such a search are now quite useful (8); chemists can now “Google” a chemical structure they have just drawn.

Co-founder of the InChI project, Stephen Heller, claims that all major chemical databases (public and commercial) now use the InChI, and that more than 100 million InChIs have been indexed (9). The InChI project is slowly expanding into inorganic, organometallic and other structures, as well as reactions. It is maintained by the InChI Trust (10).

Using Public Data

These chemical data may be used many different ways. A researcher may be looking for toxicological information or binding data on a class of substance he or she is working with. An information specialist may be looking for prior art while preparing a patent application. A scientist may be mining the data across several databases, looking for solubility trends.

In the author’s own field, chemical education, it is clear that open access to chemical information is valuable for students. Wikipedia has shown how chemical knowledge can be presented in a way that it can be easily understood by students. Wikipedia gives a context for students to understand how a topic fits into the subject, and the data it provides allow students to enrich their papers and lab reports. Meanwhile, the RSC’s Learn Chemistry Wiki shows how open chemical data (including spectra) from ChemSpider can be used to build simple substance pages inside a chemical education website, in order to make the data useful for high school and undergraduate students.

PubChem and PubMed

In 1997, the US National Library of Medicine (a division of the National Institutes for Health, NIH) launched its PubMed website which allowed the public easy access to MEDLINE, a valuable public database of abstracts from the medical literature. By the year 2000 it was receiving 250 million searches annually (11). In 2000 PubMedCentral was launched, providing the full text of medicine-related journal articles, including many chemistry papers. This has since grown to around 2.7 million articles (12). Both sites are designed to be integrated with each other, and with other US government databases.

Most publishers make their abstracts available on the Web at no cost, but full text articles traditionally require a subscription. In 2008 NIH introduced its Public Access Policy, which requires all NIH grant recipients to make their work publicly available within twelve months of publication, and to provide a manuscript for PubMedCentral (11). In February 2013, the White House announced the new *Office of Science and Technology Policy* (OSTP), which applies the same approach to all government agencies with a scientific research budget of over \$100 million, and requires open access within twelve months of publication. These agencies are now collaborating with academic publishers to develop the necessary infrastructure to comply with the new policy, via a new initiative called CHORUS (Clearinghouse for the Open Research of the United States) (13, 14). This is part of a worldwide trend towards open access publication, for example with Europe PubMedCentral, which is now used for much of the research supported by European funding agencies such as the Wellcome Foundation (a major nonprofit agency in the UK) (15). Likewise, the Canadian Institutes of Health Research (CIHR) has an Open Access Policy (16) that requires its grantees to make manuscripts public within twelve months, and these are accessible through PubMedCentral Canada.

In 2004, NIH and the US National Center for Biotechnology Information (NCBI) launched PubChem, a public database of chemical substance information (17). Despite an attempt by the American Chemical Society in 2005 to shut down PubChem (18), the database has flourished and now boasts over 100 million substance records supplied by over 200 data depositors (19, 20). Substance data are provided by chemical vendors, academic publishers, IBM (US patents) and ChemSpider (see below); biological test results come from both commercial and academic sources. PubChem also offers nearly 200 million outcomes via its PubChem BioAssay database, and validated chemical structure information on nearly 50 million compounds through its PubChem Compound database (21). Validation is only done by machine; structures are checked, as well as name-structure associations. A more sophisticated automated machine-check is to be rolled out in spring 2014 with a focus on authoritative chemical names, to fix structural errors such as stereochemical ambiguity (22). The organization offers data downloads, linked from its main page, to promote free reuse of content. Substance records offer access to related compounds, such as various metal salts for an acid.

Substance searching is quite powerful; the interface allows not just searching by name, but also structure and substructure. When the desired substance record is

found, it may carry a large amount of information, such as vendors, biological and biomedical activities, environmental fate, as well as academic literature references and patents. A typical record is shown in Fig. 1.

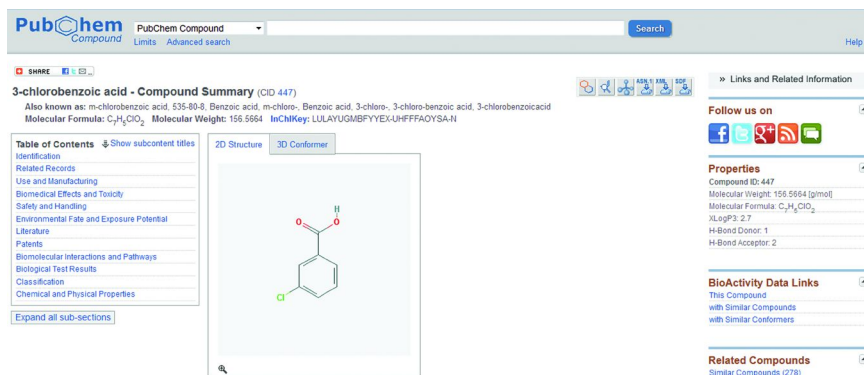


Figure 1. A compound record in PubChem.

PubChem perhaps demonstrates most powerfully the value of making data openly available. Use of bioassays and 3D structures in drug design was reviewed by Li *et al.* (23) Chen, Wild and Guha have demonstrated how PubChem bioassay information can be used for developing new drugs through polypharmacology (24). Reymond's group at the University of Berne developed a "searchable map of PubChem," also to facilitate drug development (25). As PubChem grows and drug design software improves, this approach to virtual screening will only increase in importance.

Another open public database, ChemBank (26), is maintained by the Broad Institute and funded by the National Cancer Institute (NCI), with a particular focus on cancer screening. It includes screening data not typically included in other databases. NCI also maintains an aggregator site, the Chemical Structure Lookup Service, which accepts structure searching and delivers a list of clickable IDs from several databases, equivalent to 46 million unique substances. However, as of 2013 some links in the database seem to be outdated.

ChemSpider

ChemSpider began in 2007 as a private project by Antony Williams and a small group of cheminformaticians, who saw the possibility of developing an open cheminformatics website organized around chemical structures (27) and using the InChI as an identifier system. It uses computational tools to produce a set of chemical substance pages that integrate information dynamically from over 400 sources, and most data are linked to those original sources. The site provides substance-specific information on physical constants (both experimental and predicted), spectra, relevant chemical literature, metabolism data, vendors,

etc. These substance records may be found by searching text (name, identifier) or structure or substructure. There is also some advertising. Part of a typical record is shown in Fig. 2.

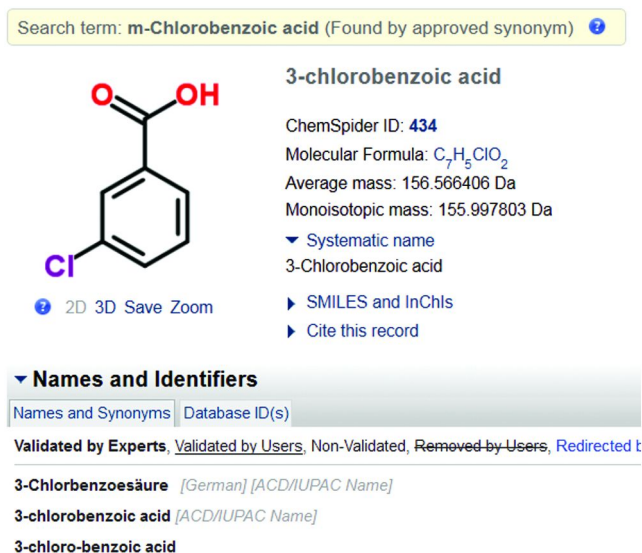


Figure 2. Part of a substance record in ChemSpider. CC-BY-SA 3.0 license

ChemSpider uses community curation to identify errors in the main database. A useful feature of the site is a set of manually curated synonyms that all point to the same InChI, allowing chemists to identify a compound from an unfamiliar name. Spectra are mostly generated on the fly from the original data, allowing the user to zoom in to examine specific details.

In 2009 ChemSpider was acquired by the Royal Society of Chemistry (RSC), and the database has since grown to over 29 million unique substances. The new ownership also allowed a lot of chemical data held by RSC to become searchable. In 2010 the SyntheticPages website, providing practical laboratory procedures, was merged with ChemSpider to become ChemSpider Synthetic Pages. ChemSpider also acts as a service provider for other projects, providing a lookup service for InChIKeys, and a chemical deposition service for the European OpenPhacts project (28). It supplies chemical data for other RSC projects such as LearnChemistry (29). ChemSpider staff has worked with the Wikipedia community to cross-check information and identify errors in both sites (30).

NIST

The US National Institute of Standards and Technology (NIST) has long served the scientific community by providing a set of standard methods and

definitions. It long provided chemical information in print, and this was the foundation of what has become a popular chemistry site, the NIST Chemistry Web Book (31). This is organized by chemical substance, and it provides chemical and physical properties, IR, UV and mass spectral data for thousands of compounds (32). The site also gives thermochemical data for substances and gas reactions. The WebBook is searchable via structure, formula and physical property values.

ChEBI and ChEMBL

The Chemical Entities of Biological Interest (ChEBI) site (33) provides a public database and ontology of chemical entities, with a special focus on small molecules. These molecules are “either products of nature or synthetic products used to intervene in the processes of living organisms” (34). It is run by the European Bioinformatics Institute (EBI), part of the publicly funded European Molecular Biology Laboratory (EMBL). The “Advanced Search” feature includes many options, including structure searches, which typically return a substance record describing key identifiers and properties. A unique feature of the ChEBI database is its ontology, which connects an entry in a formal way to other entities that are closely related. This provides a highly reliable way to perform similarity searching. An ontology entry for 3-chlorobenzoic acid is shown in Fig. 3.


ChEBI Ontology 	
Outgoing	3-chlorobenzoic acid (CHEBI:49410) has functional parent benzoic acid (CHEBI:30746)
	3-chlorobenzoic acid (CHEBI:49410) is a monochlorobenzoic acid (CHEBI:51967)
	3-chlorobenzoic acid (CHEBI:49410) is conjugate acid of 3-chlorobenzoate (CHEBI:19984)
Incoming	3-chlorobenzoate (CHEBI:19984) is conjugate base of 3-chlorobenzoic acid (CHEBI:49410)

Figure 3. Sample ontology in ChEBI. CC-BY-SA 3.0 license.

Another EMBL database is ChEMBL (35), which provides bioactive data for drug discovery. Searches can be done by ligand, target or drug, and substructure searching is available for ligand searches.

Wikipedia

Wikipedia (36) is a general encyclopedia written by “crowdsourcing” – millions of contributions from the general public submitted via the site’s wiki (a user-editable website). The site is maintained by the nonprofit Wikimedia Foundation, which also maintains other sites such as Wikimedia Commons (which holds almost 19 million media files, mainly pictures) (37). All content is released

under a Creative Commons BY-SA license, which allows it to be freely used (even commercially) as long as the source and license are clearly acknowledged, and new distribution is under the same license. The site is currently ranked as the sixth most popular website in the world by Alexa, with an estimated 365 million readers worldwide (38). A chemistry article such as “Sulfuric acid” typically receives over 100,000 page hits per month (39), and even a specialist article such as “Gold(III) chloride” normally receives at least 3000 page hits per month (40).

Some, especially more than five years ago, have criticized the “amateur” nature of Wikipedia contributors (41). However, when the content is measured objectively it often stands up well to comparison with supposed “authoritative” sources; for example, a 2009 survey of toxicologists rated Wikipedia close to WebMD for reliable toxicity information, well ahead of mainstream media outlets (42). This author has personally met several of the most active chemistry contributors, and in fact all were chemistry students or professional scientists (sometimes retired), often with a chemistry Ph.D.

Although Wikipedia does not aim to be a database, it has become the world’s largest encyclopedia, with 4.4 million articles in the English language Wikipedia alone. The dBpedia project is in fact aiming to organize the data from Wikipedia into a searchable database (43). Wikipedia contains enough chemical information to make it comparable to a small database; for example, there are over 11000 articles tagged by the Chemicals WikiProject (which covers chemical substances) (44), and 9000 of these use the ChemBox template to display substance properties in a table (45).

The chemistry content is edited and maintained by a small but dedicated group, organized through subject-based WikiProjects covering topics such as Chemistry (general topics), Chemicals (substances), Elements, Polymers, and Pharmacology. Of these, the most important are probably the chemical substance pages (around 10,000 articles), which are maintained by WikiProject Chemicals. These pages contain data inside a ChemBox, reorganized in around 2008 to facilitate maintenance and machine readability. A typical ChemBox contains the a picture and/or a structure, common identifiers, physical properties, structural and thermochemical information, and hazards. These data are compiled from many different sources, but they must meet Wikipedia’s criteria as a “reliable source” (46). Some of the data in the Chembox – mainly identifiers – have been through a validation process to ensure their accuracy. This involved checking identifiers against a master list provided by the source of the chemical identifier (e.g., CAS for registry numbers), and then deep linking to the appropriate database record. Validated content is indicated by a green check mark, which turns to a red X if the validated content is edited (47). A supplementary data page (linked from near the bottom of the ChemBox) is used to keep more specialist information off the main article page, and such pages include things such as thermodynamic data and spectral information (48). A section of a ChemBox is shown in Fig. 4.

Aside from substance articles, the Pharmacology WikiProject oversees more than 8000 articles on drugs (including illicit substances) (49); the Molecular & Cellular Biology WikiProject also maintains over 20,000 articles (50). There are also chemical reactions and processes, chemists (biographical articles), equipment, equations, and general chemical concepts.

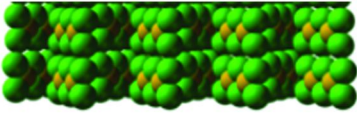

	
IUPAC name	[hide]
Gold(III) chloride	
Other names	[hide]
Auric chloride Gold trichloride	
Identifiers	
CAS number	13453-07-1 ✓
PubChem	26030
ChemSpider	8036939 ✓
UNII	15443PR153 ✓
ChEBI	CHEBI:30076 ✓
RTECS number	MD5420000
Jmol-3D images	Image 1 
SMILES	[show]
InChI	[show]
Properties	
Molecular formula	AuCl ₃ (exists as Au ₂ Cl ₆)
Molar mass	303.325 g/mol
Appearance	Red crystals (anhydrous);

Figure 4. Part of a Wikipedia ChemBox. CC-BY-SA 3.0 license

The Wikipedia chemistry editors also collaborate with outside groups. As mentioned above, there has been a longstanding relationship with ChemSpider to curate and validate data on both sites (30). CAS has shared data on around 8000 chemicals with Wikipedia chemists, and this led to both the validation of CAS Registry numbers on Wikipedia, as well as establishment of the CAS Common Chemistry website in 2009 (51). IUPAC has worked with Wikipedia, most recently adding and correcting polymer chemistry definitions within Wikipedia (52).

Wikipedia has several features that make it unique among open information source (53), because the data are almost completely entered by hand, rather than by machine, and they are very heavily curated – something otherwise only available on high-cost subscription databases. This has two effects in substance articles: First, because the information is entered by hand, structures are drawn manually and “fresh” rather than copied over from other databases, and data may often be typed in and checked against a print source. This means that errors for a particular substance that may have “infected” many databases are frequently (though not always!) avoided in Wikipedia. Second, the high page traffic and easy editing

mean that when errors do occur in Wikipedia they are often found and corrected quickly.

Another special feature of Wikipedia is the openness, transparency and customizability of the site. There is a complete edit history available for every page, which allows older versions of an article to be viewed, and for editor contributions to be checked. The WikiTrust extension for Firefox can be downloaded, allowing a user to quickly check the reliability of every word on any article page (54). Articles can be evaluated for their importance manually (using WikiProject ratings found on article talk pages) or by using the “EI” (external interest) feature in the Wikipedia 1.0 “release version” assessment tool, developed by editors to assist in compiling article collections (55).

Environmental and Safety Databases

The US Environmental Protection Agency (EPA) operates several databases. The Chemical Data Access Tool (CDAT) offers the general public a chance to examine documents deposited by organizations in compliance with the Toxic Substances Control Act (TSCA). Searches may be done via a chemical name, CAS number, company name, or document number.

Toxicology information may be found using the EPA’s toxicology databases, organized through the Aggregated Computational Toxicology Resource (ACToR) (56). They include ToxRefDB (57) (based on results of animal studies) and ToxCastDB (based on toxicity forecasting) (58). The EPA’s DSSTox database aims to use advanced structure-toxicity relationships to build “a public data foundation for improved structure-activity and predictive toxicology capabilities (59). Although many parts of NIH’s TOXNET require a license, the Hazardous Substances Data Bank (HSDB) is public. It claims to provide “comprehensive, peer-reviewed toxicology data for about 5,000 chemicals.”

The National Institute for Occupational Safety and Health (NIOSH) has long maintained its Registry of Toxic Effects of Chemical Substances (RTECS), based on data extracted from the scientific literature. However, since 2001 RTECS has been maintained by a for-profit database company, currently Accelrys (60, 61). Access to some NIOSH safety information is available, however, through the NIOSH Pocket Guide to Chemical Hazards (62). Chemical supplier MSDSs and toxicology information can be located using the University of Vermont’s “Safety Information Resource Inc.” (SIRI) website, which is supported by advertising (63). Many other public resources are available covering chemical safety, but a detailed description lies beyond the scope of this chapter.

Drug-Related Databases

As well as the government sites described above, there are several academic websites that have become important information sources.

ZINC is “a free database of commercially-available compounds for virtual screening” (64), to be used for ligand discovery. It is operated by the University of California at San Francisco and is supported by NIH. It contains over twenty

million compounds, searchable by structure, biological activity, physical property, vendor, catalog number, name, or CAS number (65). It is aimed at a user that wishes to find suitable 3D structures for docking to a biological target.

Drugbank is a Canadian “Open Drug and Drug Target Database” (66) that contains (as of October 17, 2013) 6811 drug entries and 4294 non-redundant protein sequences linked to those drugs. The site is maintained (and partly funded) by David Wishart and his research group at the University of Alberta. Searching is only text-based, but does accept a range of inputs such as name, CAS Registry Number, IUPAC name, and it can accept fields such as “approved” (by the US FDA).

Other Sites of Interest

Worldwidescience.org is an aggregator site that allows a user to conduct text searches of scientific papers across many public collections from around the world (67). Search results include full text papers, abstracts and PubMed records, for academic papers as well as scientific studies (e.g., toxicological data). Structure searching is not available, but searches can be done on the full text of the resources.

The Crystallography Open Database (COD) (68) contains close to a quarter of a million crystal structures available via open access. The site is supported by the Research Council of Lithuania, and it can be searched using text or structure (69). Data are collated from various crystallography journals, as well as from direct deposit by scientists.

The Future: Towards the Semantic Web

Wikipedia defines the semantic web as follows: “By encouraging the inclusion of semantic content in web pages, the semantic web aims at converting the current web, dominated by unstructured and semi-structured documents into a “web of data”” (70). Although this idea is supported by Tim Berners-Lee and the Web’s governing body, the World Wide Web Consortium (W3C), the concept only took off slowly (71). However, the semantic web has been influential in the sciences, largely because of the importance of data for scientists (71). Murray-Rust and Rzepa laid the foundations with Chemical Markup Language (CML), and advocated the use of documents seamlessly integrated with data, which they refer to as “datuments” (72).

The semantic web community has laid down some foundations with common standards such as the RDF (Resource Description Framework) and OWL (Web Ontology Language) (73). Hastings has explained that RDF is used for data representation, and OWL is used for classification (74). OWL was used in building the ChEBI ontologies, allowing the database to provide important substance relationships (75).

In chemistry, the Crystallographic Information File (CIF) format for crystallographic data has supported data-rich documents since the 1990s. Indeed, the International Union of Crystallography has a long history of open data (76). Likewise Jmol is a popular open standard for representing 3D structures on the

Web (77), and JCAMP serves a similar role for the storing of spectra (78). The Blue Obelisk group of scientists aims to develop a complete range of open source chemistry software and to promote the semantic web (79).

The arrival of the InChI (2005) and the InChIKey (2007) (both described above) greatly accelerated the integration of chemical structures into the Web (80). The InChI allows one to connect a structure drawing on one's personal computer to a blog post, a journal publication, or a vendor catalog entry, and the InChIKey allows one to search the internet for a specific structure.

Much more than simple structure searching is possible, as may be seen in projects such as the new European OpenPHACTS (Open Pharmacological Concept Triple Store) initiative (81). This aims to catalyze drug development by creating a semantic web concept - an 'open pharmacological space' (OPS) that fosters collaboration between academia, publishers and industry (28). As new pharmaceuticals become ever more complex and harder to bring to market, OpenPHACTS allows data to be shared among the drug development community using standard ontologies and controlled vocabularies. RSC ChemSpider (see above) serves as a chemical substance resource for the project, providing drug researchers with easy access to information such as physical properties and literature references.

Taylor predicts (82) that eventually the Electronic Laboratory Notebook (ELN) will tend to merge with the Laboratory Information Management System (LIMS) often used to organize data in the laboratory and link with instrumentation. He envisages an electronic laboratory environment (ELE), where the ELN and all instruments "talk to each other" online (83). This could in time lead to a "dark laboratory" where a researcher simply specifies a task, and automated laboratory equipment delivers the result (84).

The openness aspect is crucial. Bird and Frey have noted that (chemists) "have been slow to recognise the value of sharing and have thus been reluctant to curate their data and information in preparation for exchanging it" (85). They conclude that "In our opinion, the issues that require community action are centred on chemical data. We believe that it is essential to increase the amount of chemical data available for open access, while ensuring that new mechanisms for validating the data are provided. The community should use this data to develop more efficient links between the worlds of cheminformatics and those of materials, environmental informatics, bioinformatics, and medical informatics." Likewise, in its 2012 report "Science as an open enterprise," the RSC highlights six key areas for action (86):

- Scientists need to be more open among themselves and with the public and media
- Greater recognition needs to be given to the value of data gathering, analysis and communication
- Common standards for sharing information are required to make it widely usable
- Publishing data in a reusable form to support findings must be mandatory
- More experts in managing and supporting the use of digital data are required

- New software tools need to be developed to analyse the growing amount of data being gathered

In the book *The Fourth Paradigm: Data-Intensive Scientific Discovery* (87), a vision is given of a new, data-driven approach that will revolutionize the way scientific discoveries are made. As explained by Goble and De Roure in their chapter on workflow tools: “We are in an era of data-centric scientific research, in which hypotheses are not only tested through directed data collection and — analysis but also generated by combining and mining the pool of data already available” (88). This new world will require scientists to work openly and collaboratively using shared standards and ontologies. At the heart of this will lie our open data, accessible through our public databases. We will consider these repositories of data as being vital infrastructure for science, in the same way that roads are for transportation.

It is clear, then, that data will need to be open and unfettered in order for chemists to reap their full benefits, and public databases will play a pivotal role in disseminating data via the semantic web.

Dobbs has explained (89) that during the Renaissance and the Enlightenment, the alchemist’s personal quest to reach heaven was transmuted into Bacon’s aim of serving society, “the relief of man’s estate.” She pointed out that “the information chemists had so painfully garnered through centuries of experiment obviously had to be made public, and so it was. The new openness of chemistry stands in sharp contrast to the secrecy of the older alchemy...”

Three to four hundred years ago, the alchemists learned that sharing knowledge greatly accelerated the pace of discovery, and modern chemistry was born. The chemistry community of today will likewise have to learn and adapt to the free sharing of data, but the benefits of this change will be far-reaching.

Conclusion

We in the chemistry community must learn to integrate traditional laboratory research with analysis and mining of external data if we are to improve rates of discovery and take full advantage of the power of computers and the semantic web. Open sharing of data will be essential, and public databases will play a central role in facilitating this switch to data-based discovery. Governments and professional organizations such as ACS and RSC should support and embrace this new reality, so that chemistry can continue to flourish in the coming century.

References

1. For a blog post by Skolnik award winner Peter Murray-Rust, see <http://blogs.ch.cam.ac.uk/pmr/2011/11/25/the-scandal-of-publisher-forbidden-textmining-the-vision-denied/> (accessed October 14, 2013).
2. For further comparison of CAS files and Reaxys, see chapters in this volume by Buntrock and Rusch.

3. Wagner, R. P. Information Wants to Be Free: Intellectual Property and the Mythologies of Control. *Columbia Law Review* **2003**, *103*, doi: 10.2139/ssrn.419560.
4. Baum, R. The Google Model. *Chem. Eng. News* **2007**, *85* (45), 3, <http://cen.acs.org/articles/85/i45/Google-Model.html>.
5. Alexa web traffic rankings as of October 14, 2013: *ChemSpider* (<http://www.alexasiteinfo.com/siteinfo/chemspider.com>), 119,283; *cas.org* (<http://www.alexasiteinfo.com/siteinfo/cas.org>): 118,134. Alexa provides approximate rankings of traffic on websites.
6. Williams, A. J. InChI: connecting and navigating chemistry. *J. Cheminf.* **2012**, *4*, 33, doi: 10.1186/1758-2946-4-33.
7. Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI - the worldwide chemical structure identifier standard. *J. Cheminf.* **2013**, *5*, 7, doi: 10.1186/1758-2946-5-7.
8. Southan, C. InChI in the wild: an assessment of InChIKey searching in Google. *J. Cheminf.* **2013**, *5*, 10, doi: 10.1186/1758-2946-5-10.
9. Heller, S. CHMINF-L Listserv posting. <https://list.indiana.edu/sympa/arc/chminf-l/2013-07/msg00149.html> (accessed October 16, 2013).
10. InChI Trust home page. <http://www.inchi-trust.org/>.
11. Lindberg, D. A. B. Internet Access to the National Library of Medicine. *Eff. Clin. Pract.* **2000**, *4*, 256–260.
12. NLOM Minutes of the Board of Regents, May 2013 Meeting. <http://www.nlm.nih.gov/od/bor/5-13BORMinutes.pdf> (accessed October 16, 2013).
13. Erickson, B. E. Open-Access Teamwork Takes Off. *Chem. Eng. News* **2013**, *91* (23), 24–26, <http://cen.acs.org/articles/91/i23/Open-Access-Teamwork-Takes-Off.html>.
14. Schwartz, M. Publishers Offer CHORUS as Solution to Federal Open Access Requirements. *Library J.* June 6, 2013. <http://lj.libraryjournal.com/2013/06/oa/publishers-offer-chorus-as-ostp-solution/> (accessed October 16, 2013).
15. Walport, M.; Kiley, R. Open access, UK PubMed Central and the Wellcome Trust. *J. R. Soc. Med.* **2006**, *99* (9), 438–439, <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1557892/>.
16. Canadian Institutes of Health Research CIHR Open Access Policy. <http://www.cihr-irsc.gc.ca/e/32005.html> (accessed October 15, 2013).
17. PubChem “About” page. <http://pubchem.ncbi.nlm.nih.gov/about.html> (accessed October 15, 2013).
18. Kaiser, J. Chemists Want NIH to Curtail Database. *Science* **2005**, *308* (5723), 774, doi: 10.1126/science.308.5723.774a.
19. PubChem news announcement, September 12, 2012. <http://pubchem.ncbi.nlm.nih.gov/pcnews.html> (accessed October 16, 2013).
20. Statistics may be found by using the URL [http://www.ncbi.nlm.nih.gov/pccsubstance/?term=all\[filter\]](http://www.ncbi.nlm.nih.gov/pccsubstance/?term=all[filter]).
21. PubChem compound home page. <http://www.ncbi.nlm.nih.gov/pcccompound> (accessed October 16, 2013).
22. Dr. Evan Bolton, PubChem Project. Personal communication, March 14, 2014.

23. Li, Q.; Cheng, T.; Wang, Y.; Bryant, S. H. PubChem as a public resource for drug discovery. *Drug Discovery Today* **2010**, *15* (23–24), 1052–1057, doi: 10.1016/j.drudis.2010.10.003.
24. Bin Chen, B.; Wild, D.; Guha, R. PubChem as a Source of Polypharmacology. *J. Chem. Inf. Model.* **2009**, *49*, 2044–2055, doi: 10.1021/ci9001876.
25. Van Deursen, R.; Blum, L. C.; Reymond, J.-L. A Searchable Map of PubChem. *J. Chem. Inf. Model.* **2010**, *50*, 1924–1934, doi: 10.1021/ci100237q.
26. ChemBank home page. <http://chembank.broadinstitute.org/>.
27. Pence, H. E.; Williams, A. J. ChemSpider: An Online Chemical Information Resource. *J. Chem. Educ.* **2010**, *87* (11), 1123–1124, doi: 10.1021/ed100697w.
28. Williams, A. J.; Harland, L.; Groth, P.; Pettifer, S.; Chichester, C.; Willighagen, E. L.; Evelo, C. T.; Blomberg, N.; Ecker, G.; Goble, C.; Mons, B. Open PHACTS: semantic interoperability for drug discovery. *Drug Discovery Today* **2012**, *17* (21–22), 1188–1198, doi: 10.1016/j.drudis.2012.05.016.
29. For example, see <http://www.rsc.org/learn-chemistry/wiki/Substance:Overview> for a set of over 2000 pages that use data from ChemSpider (accessed October 15, 2013).
30. Williams, A. Dedicating Christmas Time to the Cause of Curating Wikipedia (blog post). <http://www.chemconnector.com/2008/01/09/dedicating-christmas-time-to-the-cause-of-curating-wikipedia/> (accessed October 15, 2013).
31. NIST Chemistry WebBook home page. <http://webbook.nist.gov/chemistry/>.
32. NIST WebBook entry page. <http://webbook.nist.gov/>.
33. <https://www.ebi.ac.uk/chebi/>.
34. <https://www.ebi.ac.uk/chebi/aboutChebiForward.do>.
35. <https://www.ebi.ac.uk/chembl/>.
36. Wikipedia home page. <http://www.wikipedia.org/>.
37. Wikimedia Commons home page. https://commons.wikimedia.org/wiki/Main_Page.
38. Wikipedia contributors, Wikipedia, Wikipedia, The Free Encyclopedia. <http://en.wikipedia.org/w/index.php?title=Wikipedia&oldid=576981546> (accessed October 15, 2013).
39. Wikipedia article traffic statistics (Sulfuric acid). <http://stats.grok.se/en/201309/Sulfuricacid> (accessed October 15, 2013).
40. Wikipedia article traffic statistics (Gold(III) chloride). [http://stats.grok.se/en/201309/Gold\(III\)chloride](http://stats.grok.se/en/201309/Gold(III)chloride) (accessed October 15, 2013).
41. Keen, A. *The Cult of the Amateur*; Doubleday/Currency: New York, 2007; pp 39–46.
42. Lichter, S. R. Are chemicals killing us? http://www.stats.org/stories/2009/are_chemicals_killing_us.html (accessed October 17, 2013).
43. dBpedia home page. <http://dbpedia.org/About>.
44. Project summary table (Chemicals project). <http://tools.wmflabs.org/enwp10/cgi-bin/table.fcgi?project=Chemicals> (accessed October 16, 2013).

45. This can be seen by clicking through the pages found at: <https://en.wikipedia.org/w/index.php?title=Special:WhatLinksHere/Template:Chembox&hidelinks=1&limit=500> (accessed October 16, 2013).
46. Wikipedia editors, Identifying reliable sources. https://en.wikipedia.org/w/index.php?title=Wikipedia:Identifying_reliable_sources&oldid=575962745 (accessed October 15, 2013).
47. This is another example of the customizability of Wikipedia. This automated “bot,” CheMoBot, was designed specifically to protect validated chemistry content. See <https://en.wikipedia.org/wiki/User:CheMoBot>.
48. See [https://en.wikipedia.org/wiki/Methanol_\(data_page\)](https://en.wikipedia.org/wiki/Methanol_(data_page)) for a well-developed example data page.
49. Pharmacology WikiProject Statistics. <http://tools.wmflabs.org/enwp10/cgi-bin/table.fcgi?project=Pharmacology> (accessed October 21, 2013).
50. Molecular & Cellular Biology WikiProject Statistics. <http://tools.wmflabs.org/enwp10/cgi-bin/table.fcgi?project=MCB> (accessed October 21, 2013).
51. CAS Launches Free Web-Based Resource “Common Chemistry” for General Public <https://www.cas.org/news/media-releases/common-chemistry> (accessed October 16, 2013).
52. https://en.wikipedia.org/wiki/Wikipedia_talk:WikiProject_Polymers#Working_with_IUPAC_Polymer_group_on_cleaning_up_definitions (accessed October 16, 2013).
53. A recent webinar presentation by the author provides more detail on chemical information within Wikipedia: <http://www.acscinf.org/content/webinar-using-wikipedia-source-chemical-information>.
54. WikiTrust home page. <http://www.wikitrust.net/>.
55. To generate tables of chemical substance information, begin at <http://tools.wmflabs.org/enwp10/cgi-bin/list2.fcgi?projecta=Chemicals&limit=50>. The EI data are seen only if the external interest data box is checked.
56. EPA’s ACToR home page. <http://actor.epa.gov/actor/faces/ACToRHome.jsp> (accessed October 21, 2013).
57. EPA’s Toxicity Reference Database (ToxRefDB). <http://actor.epa.gov/toxrefdb/faces/Home.jsp> (accessed October 16, 2013).
58. EPA’s Toxicity Forecasting Database (ToxCastDB). <http://actor.epa.gov/actor/faces/ToxCastDB/Home.jsp> (accessed October 16, 2013).
59. EPA’s Distributed Structure-Searchable Toxicity (DSSTox) Database Network. <http://www.epa.gov/ncct/dsstox/index.html> (accessed October 16, 2013).
60. Licensing Agreement Signed for RTECS. <http://www.cdc.gov/niosh-rtecs/RTECSLicense.html> (accessed October 16, 2013).
61. RTECS: How do I access it? <http://www.cdc.gov/niosh/rtecs/RTECSaccess.html> (accessed October 16, 2013).
62. NIOSH Pocket Guide to Chemical Hazards. <http://www.cdc.gov/niosh/npg/> (accessed October 16, 2013).
63. SIRI home page. <http://siri.org/> (accessed October 16, 2013).
64. ZINC home page. <http://zinc.docking.org/> (accessed October 16, 2013).

65. Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. J. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757–1768, doi: 10.1021/ci3001277.
66. Drugbank home page. <http://www.drugbank.ca/>.
67. Wordlwidescience.org home page. <http://worldwidescience.org/index.html>.
68. Crystallography Open Database home page. <http://www.crystallography.net/>.
69. Gražulis, S.; Daškevič, A.; Merkys, A.; Chateigner, D.; Lutterotti, L.; Quirós, M.; Serebryanaya, N. R.; Moeck, P.; Downs, R. T.; Le Bail, A. Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucl. Acids Res.* **2012**, *40* (D1), D420–D427, doi: 10.1093/nar/gkr900.
70. Wikipedia contributors, Semantic Web, Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/w/index.php?title=Semantic_Web&oldid=576694095 (accessed October 17, 2013).
71. Shadbolt, N.; Hall, W.; Berners-Lee, T. The semantic web revisited. *Intelligent Systems*, **IEEE**. Available from http://eprints.soton.ac.uk/262614/1/real/OLD_Semantic_Web_Revisted.pdf.
72. Murray-Rust, P.; Rzepa, H. S. Scientific publications in XML - towards a global knowledge base. *Data Sci. J.* **2002**, *1*, 84–98, doi: 10.2481/dsj.1.84.
73. For a detailed example, see the chapter in this volume by Batchelor, describing ontologies, OWL and RDF, including ChEBI and resources developed for OpenPHACTS.
74. Hastings, J. Chemical Classification for the Semantic Web, presented at the Skolnik Symposium, August 2012, in Philadelphia, PA. <http://www.slideshare.net/jannahastings/chemical-classification-for-the-semantic-web>.
75. Hastings, J.; Magka, D.; Batchelor, C.; Duan, L.; Stevens, R.; Ennis, M.; Steinbeck, C. Structure-based classification and ontology in chemistry. *J. Cheminf.* **2012**, *4*, 8, doi: 10.1186/1758-2946-4-8.
76. McMahon, B. Applied and implied semantics in crystallographic publishing. *J. Cheminf.* **2012**, *4*, 19, doi: 10.1186/1758-2946-4-19.
77. http://wiki.jmol.org/index.php/Main_Page.
78. <http://www.jcamp-dx.org/>.
79. O’Boyle, N. M.; et al. Open Data, Open Source and Open Standards in chemistry: The Blue Obelisk five years on. *J. Cheminf.* **2011**, *3*, 37, doi: 10.1186/1758-2946-3-37.
80. Williams, A. J. Internet-based tools for communication and collaboration in chemistry. *Drug Discovery Today* **2008**, *13* (11/12), 502–506, doi: 10.1016/j.drudis.2008.03.015.
81. OpenPhacts home page. <http://www.openphacts.org/> (accessed October 16, 2013).
82. Taylor, K. T. The status of electronic laboratory notebooks for chemistry and biology. *Curr. Opin. Drug Discovery Dev.* **2006**, *9* (3), 348–353.
83. Taylor, K. T. Evolution of Electronic Laboratory Notebooks. In *Collaborative Computational Technologies for Biomedical Research*; Ekins,

- S., Hupcey, M. A. Z., Williams, A. J., Eds.; John Wiley & Sons: Hoboken, NJ, 2011; pp 303–320.
84. Frey, J. G. Dark Lab or Smart Lab: The Challenges for 21st Century Laboratory Software. *Org. Process Res. Dev.* **2004**, 8, 1024–1035, doi: 10.1021/op049895.
 85. Bird, C. L.; Frey, J. G. Chemical information matters: an e-Research perspective on information and data sharing in the chemical sciences. *Chem. Soc. Rev.* **2013**, 42, 6754–6776, doi: 10.1039/c3cs60050e.
 86. Science as an open enterprise, the Royal Society Science Policy Centre report, June 21, 2012. <http://royalsociety.org/policy/projects/science-public-enterprise/report/>.
 87. *The Fourth Paradigm: Data-Intensive Scientific Discovery*; Hey, A., Tansley, S., Tolle, K., Eds.; Microsoft Research: Redmond, WA, 2009; <http://research.microsoft.com/en-us/collaboration/fourthparadigm/>.
 88. See Goble, C.; de Roure, D. *The Impact of Workflow Tools on Data-centric Research*; in ref 75, p 137.
 89. Dobbs, B. J. T. From the Secret of Alchemy to the Openness of Chemistry. In *Solomon's House Revisited: The Organization and Institutionalization of Science*; Frängsmyr, T., Ed.; Science History Publications: Canton, MA, 1990; pp 75–94.

Chapter 13

Chemistry Ontologies

Colin Batchelor*

Royal Society of Chemistry, Thomas Graham House, Cambridge,
United Kingdom CB4 0WF

*E-mail: batchelorc@rsc.org

I provide an overview of ontologies in chemistry, what they are, how they are used at present, where they might be used in future and where they fall short of what you might hope for. In particular I describe their application in a large drug discovery infrastructure project and how the approach taken there might be applied to providing machine-readable descriptions of chemical experimentation in general.

Introduction

A growing area of research in recent years has been the application of ontologies to microarray data and integrating such data with other sources of information. Concomitantly there has also been a great deal of interest in the semantic web. As a discipline, chemistry is no exception in producing and disseminating large amounts of heterogeneous data, as exemplified by PubChem, ChemSpider and ChEMBL. A recent innovation has been the use of ontologies in order to classify, disseminate and link this information. An early example of this, linking together genes, proteins, genetic variations, chemical compounds, diseases and drugs is given by Chen et al. (1) in the form of the Chem2Bio2RDF project.

In this chapter I will explain what “ontology” means in this context and cover how ontologies are structured, means of representing ontologies, examples of chemical ontologies, including those produced and distributed by the Royal Society of Chemistry as part of its mission to advance the chemical sciences, and applications of ontologies. In terms of applications I will chiefly focus on the data produced by various sources as part of the Open PHACTS project (2). This project has been set up as part of the European Union’s Innovative Medicines initiative to support drug discovery programs in the public domain

and in pharmaceutical companies by delivering web interfaces and application programming interfaces (APIs) for providing chemical, pharmacological and biological data about small molecules and proteins. As part of this project at the Royal Society of Chemistry we generate data sets that describe synonyms and identifiers, calculated physicochemical properties for compounds and links between different data sources.

Background

What Is an Ontology and How Is It Structured?

First of all, what do I mean by “ontology”? In the Stanford Encyclopedia of Philosophy, Hofweber (3) offers the following four possibilities:

- (O1) *the study of ontological commitment, i.e. what we or others are committed to,*
- (O2) *the study of what there is,*
- (O3) *the study of the most general features of what there is, and how the things there are related to each other in the metaphysically most general ways, and*
- (O4) *the study of meta-ontology, i.e. saying what task it is that the discipline of ontology should aim to accomplish, if any, how the questions it aims to answer should be understood, and with what methodology they can be answered.*

For my purposes in this chapter I shall modify (O3) and take an ontology to be a machine-readable account of what there is in a given domain and how the things there relate to other things, not necessarily in the most metaphysically general way, but in a way that is consistent with how practicing scientists in a domain understand the relations.

What do I mean here by a machine-readable account? The word “proposition” in the philosophical literature has a number of meanings, but a useful one, and one that is compatible with machine reasoning, is as a bearer of truth-value: a statement that can be evaluated as being true or false. Any proposition must be “about” something, something that makes it true, and we call the things mentioned in a proposition “entities”. In this chapter I take a machine-readable account to mean a series of propositions about the entities within a given domain, for example chemistry or stamp-collecting. Why is machine-readability important? Simply because the datasets encountered for cheminformatics applications, particularly virtual screening in the context of drug discovery, often number millions of structures or tens of thousands of scientific articles and this is too many for an unassisted human being to deal with.

One’s first thought about machine-readability in cheminformatics might be something expressed in a file format, such as a V2000 mol file, or in a line notation. Line notations express a chemical structure or a reaction involving chemical structures as a linear sequence of characters and five examples of line notations in contemporary use are given in Table I. SMILES notation (4) represents molecules

in a human-readable way, for example cyclobutane is C1CCC1, indicating that there are four carbon atoms arranged so that the last is bonded to the first, with hydrogen atoms as needed to make up the numbers. The InChI representation (5) is InChI=1S/C4H8/c1-2-4-3-1/h1-4H2, which specifies all of the atoms present and their connectivity. This being largely composed of punctuation is ill-suited to indexing in a search engine and so the InChIKey was introduced to fill the gap. To specify parts of molecules, the SMARTS specification has been built upon SMILES, and the SMIRKS notation combines this with atom indexing in the reactants and products in order to provide atom–atom mapping. On their own, however, expressions in line notation are neither true nor false. For this reason, a string written in a line notation does not count as a proposition on its own. However, it might be part of a proposition, for example “the SMILES string for benzene is c1ccccc1”.

Table I. Line notations in cheminformatics

<i>Line notation</i>	<i>Example</i>	<i>Interpretation</i>
SMILES	C1CCC1	Cyclobutane molecule
InChI	InChI=1S/C4H8/c1-2-4-3-1/h1-4H2	Cyclobutane molecule
InChIKey	PMPVIKIVABFJJI-UHFFFAOYSA-N	Cyclobutane molecule
SMARTS	[CX1]#[NX2]	Nitrile group
SMIRKS	[c:1][C:2](=O)O>>[c:1][C:2]=C(=O)O	Perkin reaction

The sorts of proposition we find most often in chemistry ontologies include:

Subsumption: for example, every benzene is an aromatic molecule. (E1)

Parthood: for example, every benzene has part some benzene ring. (E2)

Representation: for example, this connection table represents such-and-such a molecule. (E3)

Participation: for example, every Diels–Alder reaction has participant some diene. (E4)

Within an ontology, the propositions are not, however, represented as sentences as in the above examples. They are represented as a string of textual identifiers for the domain entities and the relations between them. Turtle format (6) provides a human-and-machine-readable method for this, listing textual identifiers for the subject (benzene, connection table, Diels–Alder reaction), the predicate (is a, has part, represents, has participant) and the object (aromatic molecule, benzene ring, such-and-such a molecule, diene) separated by spaces and completed with a full stop. Example (E1) rewritten in this format would be:

obo:CHEBI_16716 rdfs:subClass obo:CHEBI_33655

where `rdfs:subClass` is the “is a” relation and the strings beginning with “obo:” refer to the classes “benzene molecule” and “aromatic molecule”. By “class” I mean that the proposition relates to benzene molecules in general, as opposed to a specific benzene molecule under the tip of a scanning–tunneling electron microscope. As you can see, the identifiers are relatively opaque so that they do not have to change if our knowledge about a subject changes or if someone has misspelt something or, for example, a taxonomic species is renamed in the light of new discoveries. This is a feature of biomedical ontologies; the ontologies that computer scientists develop, often to test code that draws inferences based on the propositions in an ontology, to teach people about how reasoning works or to explore the expressiveness of a given ontology language, will have non-opaque IDs because it is important that these propositions should be easily readable by a human being, and revision in the light of new scientific knowledge is less important. One principle of the Semantic Web is that identifiers should have a readily-accessible definition over the web; this implies that URLs should be used.

The prefixes `rdfs:` and `obo:` are shorthands for fragments of HTTP URLs. One underlying notion of the Semantic Web is that data should be in some sense self-describing; hence these identifiers should (and in the `rdfs:` and `obo:` cases do) resolve over the web to a machine-readable description. Later in this chapter (under Representing Ontologies) I will illustrate some of these machine-readable descriptions and the conventions behind them.

Taken together, these propositions constitute a system that can be checked for internal consistency, for example if the ontology defines somewhere that no protection reaction can also be a deprotection reaction, then deprotections that have been manually misclassified as protections can be identified. This is not a fanciful example; this is something I personally have done by mistake. This can be done programmatically, for example using a reasoner, that is to say a program that draws inferences, or within an ontology editor. A reasoner can also be used to infer things not made explicit in the system. For this sort of reasoning we need quantification, that is to say, what propositions are true of every x , some x or perhaps no x . I will discuss this in greater detail later on.

Even without the propositional structure, the mapping between identifiers and human-readable names, ideally the names that are found in the scientific literature, is in itself a useful artefact that can be used for indexing or more sophisticated forms of text mining, and I will discuss this in more detail in the Applications section.

There are clear similarities between an ontology and a database. One way of thinking about a database is that records contain propositions about entities, the role of the identifiers in an ontology being played by the primary keys in the database. We can think of a query with joins (one that combines data from different tables to extract information that may not have been explicitly put into the database) as being analogous to reasoning over an ontology. To this end, just as there exists SQL, a standard query language for relational databases, so there is SPARQL (7), a query language for ontologies and knowledge bases. SPARQL does not allow the underlying knowledgebase to be altered; this has led people, perhaps incautiously, to set up public SPARQL endpoints on the web allowing anyone to query a knowledge base, something which would be

extremely hazardous for relational databases as it would enable members of the public to modify and delete information within the database without an audit trail. In fact, “SQL injection”, sending an appropriately-formatted string to a website that alters the underlying relational database, is a well-known vulnerability. An analogous vulnerability for a SPARQL endpoint might be a query that returned all of the underlying data. A further important distinction is the contrast between the Closed World Assumption of databases – that anything unknown to the database is false, and the Open World Assumption of ontologies, that anything otherwise unspecified by the ontology we can draw no conclusion from.

In laboratory domains it is often useful to relate the entities to an upper-level ontology, which is a small ontology that typically distinguishes objects (for example molecules) from the processes (for example cyclization) they participate in. This has been used, for example, in the Gene Ontology (8) to find errors and inconsistencies. The distinction is less obvious and perhaps less useful when describing software artefacts as computer programs are themselves data. Examples of upper-level ontologies include the Descriptive Ontology for Linguistic and Cognitive Engineering, DOLCE (9) and the Basic Formal Ontology, BFO (10). In general the former is more popular among ontology researchers and the latter is more popular in the biomedical ontologies community.

Representing Ontologies

At the moment, ontologies are typically stored in an XML serialization of the Web Ontology Language, OWL (11). An example of this is given in Table II. Some of the XML elements are in the owl: namespace, but many others come from the Resource Data Format, RDF (12) namespace, as OWL has been built on top of RDF, and as such Table II provides an example of both. The best way of thinking about the two is that OWL is best suited to describing the relations between things in general (types, classes, universals), whereas RDF is better suited to things themselves (tokens, individuals, particulars). The conventional example is that when one talks about Socrates being a man, Socrates is the individual, and man is the type or class.

Table II. A sample of OWL serialized as XML.

```
<owl:Class rdf:about="http://purl.obolibrary.org/obo/CHEBI_15734">
  <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">primary
  alcohol</rdfs:label>
  <rdfs:subClassOf rdf:resource="http://purl.obolibrary.org/obo/CHEBI_24431"/>
  <oboInOwl:id rdf:datatype="http://www.w3.org/2001/
  XMLSchema#string">CHEBI:15734</oboInOwl:id>
</owl:Class>
```

A key feature of OWL, which is an evolving standard, is that it is a subset of first-order logic which deliberately hobbles what you can say in order to ensure that any inferential process will actually terminate. This is important for web

applications in order to avoid denial-of-service attacks that would involve a website being given a request that was known not to terminate. OWL comes from the Description Logic (DL) tradition. A key difference between conventional first-order logic and DLs is that first-order logic allows definitions that contain variables. Hence an epoxy molecule can be defined as one where an oxygen atom o is bonded to exactly two carbon atoms c_1 and c_2 , and those are themselves bonded to o . However, DL disallows this. One can only say that there is an oxygen atom that is bonded to exactly two carbon atoms which are themselves bonded to exactly one oxygen atom and one other carbon atom. While there may be workarounds for small systems, in general the problem is intractable and I will come to some potential solutions in the next section.

DL also has some peculiar terminology – this does not limit its power or scope but adds a layer of perhaps unnecessary opacity to papers describing how it works. What would normally be called a predicate is called a role (or an object property in OWL), what would normally be a class is called a concept and what would normally be called a proposition is called an axiom.

OWL divides up the universe of discourse as follows: there are classes (in the language of DL, the TBox or terminology box) for example “man”, and there are individuals that instantiate those classes (the ABox or assertion box), for example “Socrates”. We write propositions about those classes and individuals in terms of properties, the division of which is threefold. There are object properties, which relate classes to classes and individuals to individuals (hence “all men are mortal”), data properties, which relate individuals to strings and other instances of data types, such as floating-point numbers, integers and dates (Socrates was born on the 21st of January), and there are annotation properties, which describe the classes and individuals themselves (Socrates is called “Socrate” in French), rather than, for example, their referents.

Particularly popular in the biomedical domain is the Open Biomedical Ontologies (OBO) format (13), which has two practical advantages over unadorned OWL. Firstly, the basic format makes detailed provision for synonyms – as such OBO is well-suited to handling terminologies if not disambiguating them. Secondly, the format is readily human-writable using a simple text editor; see Table III for an example. OWL format, on the other hand, is typically best handled using the OWL API (14) or a tool such as Protégé.

Table III. An example class definition in OBO format.

<p>[Term] id: RXNO:0000036 name: Reformatsky reaction def: "A carbon-carbon coupling reaction where an aldehyde or amine reacts with a alpha-halo ester and zinc to form a beta-hydroxy ester." [RSC:cb] synonym: "Reformatskii reaction" EXACT [] is_a: RXNO:0000002 ! carbon-carbon coupling reaction relationship: has_part MOP:0000580 ! ketone reduction relationship: has_part MOP:0001550 ! dehalogenation</p>
--

The bridge between the two has come from both the logical end and the human-interface end. Firstly Golbreich et al. (15) have hardened up the previously informal semantics of OBO by providing a mapping to OWL. I will give an example of why this is necessary later on in my discussion of quantification. Secondly, OWL Manchester Syntax has been developed (16) to serve two ends: firstly to provide a human-writable way of typing propositions in to an editor, and secondly to provide a more user-friendly way of showing these propositions than the symbolic “German” syntax that had been used previously. Table IV shows an example of this. A particularly exciting use is for providing explanations of inconsistencies (17), as in my protection/deprotection example earlier. and a human-readable notation is also particularly useful for explaining the output of a reasoning process. A trivial example might inferring that a cat is a mammal because all cats are felids and all felids are members of Carnivora and all members of Carnivora are mammals.

Table IV. An example of OWL Manchester syntax taken from (15)

Class: VegetarianPizza EquivalentTo: Pizza and not (hasTopping some FishTopping) and not (hasTopping some MeatTopping) DisjointWith: NonVegetarianPizza

An important feature of most relations in scientific ontologies, for example parthood and participation, is that they express an ontological dependence, that is to say that it is impossible for something to be a benzene molecule without having as part some benzene ring. However, names and identifiers are not like this. It is neither necessary to the word “benzene” nor to a benzene molecule itself that the other exist. The same is true of words like “wyvern” and “polywater”. For many applications we do not want to reason over these inessential properties so OWL provides annotation properties which can be used for indicating synonyms and identifiers, particularly line notation in the case of molecules and reactions.

How would we deal with references to wyverns and polywater in text? We certainly shouldn’t define polywater as a kind of polymer or a wyvern as a kind of animal as this would lead to nonsensical entailments, for example any papers that showed that there was no such thing as polywater would contain existential statements along the lines of “there is some p such that p is polywater and there is no p ”. It is better to define them in terms of the polywater hypothesis, or, in heraldry, the wyvern shape. There are as yet no good automated ways of handling these in text-mining as seldom-referenced dead-end hypotheses such as polywater are not generally amenable to high-throughput analysis.

Examples

Example Ontologies in Chemistry: Small Molecules

One might ask, given the power of chemoinformatics methods such as 2D fingerprinting, substructure search and scaffold hopping, why one might need a hierarchical hand-built system for managing knowledge about small molecules.

Hastings et al. (18) argue, *inter alia*, that there is a wealth of information in textual descriptions of classes of molecule, listing examples like “1-alkenoylcyclopropane carboxamides”, which better describe the focus of an article or indeed the research agenda of an entire group than any scaffolds one might be able to infer from looking at full structures. Aside from the systematic names there are also natural-product-based names such as “polyketide” or “spongistatin” that reflect the origin of the molecules in question.

As discussed by Richter (19), the basics of chemical classification in its modern form date back to around 1840, in terms of parent nuclei, homologous series and functional groups. These have subsequently been refined and enlarged and codified by bodies such as IUPAC as detailed in the Red (20) and Blue Books (21).

A long-established flagship ontology project in the biomedical domain consists of the Gene Ontology (8), and the Gene Ontology annotation (GOA) database (22), which together provide a wealth of information about biology from the molecular up to the organismal level. The Gene Ontology provides vocabularies for cellular components, molecular functions and biological processes, while GOA is an abstracting service for the biological literature that annotates gene products, that is to say proteins and messenger RNAs, with their molecular function, where in the cell they do their work, and what broader biological processes these are implicated in.

Chemical Entities of Biological Interest, ChEBI (23), started initially as a dedicated chemical classification for those classes in the Gene Ontology that reference small molecules. A detailed description of how the two ontologies interact is given by Hill et al. (24). It was subsequently developed as a reference implementation of IUPAC guidelines as specified in the Red and to some extent the Blue Books (20, 21) for chemical nomenclature in the sense that the entries for a given name are to be taken as definitive, although it does not provide a resolution service for unknown names. It has subsequently gained links to patent databases and the natural product literature.

An important development in the transition of ChEBI from being merely a controlled vocabulary to an ontology *per se* has been its treatment of quantification. An important feature of ChEBI is that its underlying storage medium is a relational database on which humans make queries and the OBO and OWL versions are merely serializations. The interpretation of an entry in the database, for example, that the relation table has a record that connects the oxygen atom with a parthood relation and the water molecule, depends on the chemically-aware reader. Informally one might say that an oxygen atom is part of a water molecule. Initially, based on the Gene Ontology, such relations were expressed in terms of a *part_of* relation. However, the quantification of the OWL translation of the OBO-style relation “tetracyanonickelate(2-) *part_of*

potassium tetracyanonickelate(2-)” is that “all tetracyanonickelate(2-) ions part of some potassium tetracyanonickelate(2-) complex”, which is patently absurd. Better to say, as ChEBI now does, that “all potassium tetracyanonickelate(2-) complex has part some tetracyanonickelate(2-) ion” – any complex lacking a tetracyanonickelate(2-) ion cannot be a potassium tetracyanonickelate(2-) complex. This is now codified in the OWL files that are available for download from the ChEBI website.

ChEBI is, however curated by hand. As of January 2014 it contains 37271 classes that represent molecules, families of molecules with a detailed structural classification or roles played by those molecules organized into a hierarchy, which is rather too big a number to systematically maintain without automated assistance. There are also many classes within the ontology with only a rudimentary classification which the curators will come to. One example of how its curation might be automated is given by Bobach et al. (25) who build a ChEBI-like ontology with a hybrid approach that relies on well-established cheminformatics methods to automatically classify molecular structures. To be precise, they use SMARTS expressions as seen in Table I to specify the connectivity of atoms within a molecular structure. Note that only the *a priori*, structural component can be automated reliably; the process of curating what molecules do inside organisms is a posteriori and is necessarily based on experiment.

A wholly formal-logical approach is demonstrated by Magka (26), who expresses chemical structures in terms of propositions in first-order logic, for example, two-place predicates to express bonding between atoms ($\text{single}(f_{12}(x), f_i(x))$ expressing the notion that there is a single bond between atom 12 and atom i), and one-place predicates to express the properties of those atoms, for example $c(f_i(x))$ indicating that atom i is a carbon atom, as shown in Fig. 1, and then shows rules that may be used to determine subclass relations between them. Unlike previous work, for example by Hastings et al. (27), the classification code runs in a reasonable amount of time, though it is still slower than implementations based on matching SMARTS expressions.

$$\begin{aligned} \text{ascorbicAcid}(x) \rightarrow \bigwedge_{i=1}^{13} \text{hasAtom}(x, f_i(x)) \wedge \text{molecule}(x) \wedge \bigwedge_{i=1}^6 o(f_i(x)) \wedge \bigwedge_{i=7}^{12} c(f_i(x)) \wedge \\ h(f_{13}(x)) \wedge \text{single}(f_8(x), f_3(x)) \wedge \text{single}(f_9(x), f_4(x)) \\ \wedge_{i=1,9,11,13} \text{single}(f_{10}(x), f_i(x)) \wedge_{i=5,11} \text{single}(f_{12}(x), f_i(x)) \\ \wedge_{i=1,8} \text{single}(f_7(x), f_i(x)) \wedge \text{single}(f_{11}(x), f_6(x)) \wedge \\ \text{double}(f_2(x), f_7(x)) \wedge \text{double}(f_8(x), f_9(x)) \end{aligned}$$

Figure 1. Formal-logical representation of ascorbic acid according to Magka (26). (Reproduced with permission from reference (26). Copyright 2012)

There has been comparatively little work on larger systems; however the NanoParticle Ontology (28) is a promising piece of work for the nanosciences which contains not only a set of classes of nanoparticles, but also their properties (in the familiar sense of the word), for example the surface area, chemical composition, surface charge and zeta potential of a nanoparticle surface. It also provides relations that can be used to specify the composition of a nanoparticle

(has_entrapped_component_part, has_encapsulated_component_part and so forth). It is being used by the United States National Cancer Institute's Nanotechnology Working Group in their work on the rational design of nanomaterials and in finding nanomaterial structure–activity relationships. As of December 2013 it has 1904 classes.

Example Ontologies in Chemistry: Processes

The domain of the CHEMINF ontology (29) is chemoinformatics. To this end it represents both chemoinformatics algorithms and the data that they process and output. The algorithms mentioned include those to calculate the polar surface area of a molecule, to calculate partition coefficients and to standardize chemical structures according to some set of rules. The data items include molecular connection tables, molecular formulae and numbers of freely-rotating bonds. Importantly the ontology includes references to software packages, which provides a means to give the provenance of a given calculation. It is being used in the Open PHACTS project (2) as I will describe below. As of January 2014, it has 652 classes.

The Chemical Methods Ontology (CHMO) (30) was initially based on the IUPAC Orange Book (31) and intended to cover the methods described therein for collecting analytical data, such as mass spectrometry and electron microscopy. Subsequently it has been extended to cover the methods to prepare and separate material for further analysis, such as sample ionization, chromatography and electrophoresis, to synthesize materials, such as epitaxy and continuous vapour deposition, the instruments used in these experiments, like mass spectrometers and chromatography columns, and their outputs. It now (December 2013) has 2745 classes. It was initially developed for text mining as part of the RSC's Project Prospect, this text-mining being ongoing, but should be usable for describing all aspects of an experiment. The Golm Metabolome Database, a reference library of GC-MS experiments (32) uses CHMO to describe some of the parameters in gas chromatography and mass spectrometry experiments

As for small-molecule reactions, Ingold's nomenclature for reaction mechanisms goes back to before the Second World War. Carey et al. (33) offer a categorization of small molecule reactions and classify reactions from the databases of AstraZeneca, GlaxoSmithKline and Pfizer against them. This categorization, excluding "miscellaneous", has 11 categories, which are focussed on the chemical transformations from a synthetic point of view rather than the precise mechanism. These are heteroatom organylation, acylation, carbon–carbon bond forming, aromatic heterocycle formation, deprotection, protection, reduction, oxidation, functional group interconversion, functional group addition and resolution. This has not, however, been formalized into an ontology that can be reasoned over. That falls to RXNO, the name reactions ontology (34), which has 511 classes as of January 2014. The top levels of the "intentional" classification in slight contrast to Carey et al.'s classification are cleaving, condensation, functional modification, joining, rearrangement, ring breaking, ring contraction, ring expansion, ring formation and ring rearrangement. The "intentional" classification is based on two principles: firstly comparing the

unbroken carbon chains in the reactants and products and secondly considering whether a ring system is created, destroyed or altered. These are all worked out from the perspective of an organic chemist

Here, as in the case of small molecules, we come up against the limits of the DL approach. To take the Diels–Alder reaction, it is necessary but not sufficient for a reaction to be a Diels–Alder reaction if it involves the reaction of a diene with a double-bonded system producing a cyclohexadiene. The ring itself must consist of those atoms that previously constituted the diene part of one reactant and the double-bonded system of the other. We can express this using SMIRKS notation because SMIRKS notation allows us to number atoms and hence provide a mapping from the reactants to the products, but not with the resources available to us within OWL as it is impossible within the definitions allowed in a DL framework to talk about a given atom as we saw in the epoxy example previously. Any approach would have to, like in Magka’s work on chemical structures, go outside the DL framework and this is not currently supported by well-established web standards.

Applications

Text Mining

The most straightforward application of an ontology, and one that does not require any of the logical apparatus, is in named-entity extraction to provide a controlled vocabulary of terms found in text and identifiers for those. The generic named-entity extraction process works roughly as follows: a document is segmented into sentences, then those sentences are tokenized (split on spaces and relevant punctuation) and those tokens or token sequences assessed for their likelihood of being named entities relevant to the domain. The simplest way of doing this is to compare them to a pre-existing dictionary, for which ontologies are pre-eminently suitable. It is worth mentioning in passing that in chemical documents most of the compounds of interest will be brand new and hence in no dictionary, so in general a name-to-structure approach will be needed.

An example explicitly using chemical ontologies is provided by Batchelor and Corbett (35) who describe in detail how named entity recognition based on ontology identification can be applied to annotate a journal article stored as XML by adding more XML elements to it, but more general examples abound outside chemistry, particularly one hand-built example by Shotton *et al.* (36). Ontologies do not inherently help with the task of word-sense disambiguation beyond providing different identifiers for the same name, which is useful to distinguish the senses of a word like ‘cell’, which could refer to a biological cell, an electrochemical cell, a solar cell or possibly in an environmental monitoring journal a room that accommodates prisoners, or “plant” in a botanical context as opposed to a manufacturing context. Ontologies also provide textual definitions which could be used in examining the immediate textual context of a name to provide clues as to its referent. In (37) Corbett *et al.* use the word ‘pyridine’ to exemplify the more tractable case where chemical names may have more than one reading and how they may be disambiguated.

The first distinction is between lab-scale and molecular-scale. “Pyridine” may refer to the substance in a bottle or to a given molecule. In (37) the authors leave this unresolved as it is in practice a less important distinction than the second, which is between “pyridine” the cyclic molecule with formula C_5H_5N and “a pyridine”, any molecule containing an unfused aromatic C_5H_5 ring. They call these the EXACT and CLASS readings respectively. A third, practically-driven sense is that of “pyridine” in “pyridine ring”, which they call the PART reading. This is related to the CLASS sense in that it refers to the aromatic C_5H_5 ring *per se* rather than merely a molecule that contains one. This threefold distinction is honoured in ChEBI where “pyridine”, “pyridines” and “pyridine ring” are different classes and have different identifiers.

The dissemination of these annotations is not restricted to the “Rich HTML” view on the RSC Publishing Platform as described in (38). As most readers still prefer to read PDFs rather than HTML, Pettifer et al.’s Utopia Documents PDF reader (39) uses information from the RSC’s web services to show the annotations found by text mining within the PDF on screen.

Open PHACTS and Other Datasets of Pharmacological Interest

As part of the Open PHACTS project (2), the Royal Society of Chemistry pulls together molecular structures from a variety of databases, chiefly ChEBI, ChEMBL and DrugBank, validates them and produces linksets between them. We use the Vocabulary of Interlinked Datasets (VoID) (40) to specify what predicates are used and what sorts of subject and objects are being interlinked. These are particularly valuable because RDF documents can be huge, containing millions of triples, and the VoID provides a concise summary of many triples using each predicate there are. We also use the Open PHACTS dataset specification (41) to specify what the justifications are for each connection – for example the structure–structure mapping is based on the InChIKey (see Table I) and a class from the CHEMINF ontology is used to indicate this. The SKOS vocabulary (42) is also used to distinguish those links that hold in all cases and those links that only hold under certain circumstances, such as those produced by disregarding stereochemistry or isotopic substitution.

Another dataset we produce is sets of validated and unvalidated synonyms for free-text querying. Given that these synonyms, especially the “unvalidated” ones (synonyms that have come in to ChemSpider from a chemical vendor and will not have been curated by a human being) are inessential properties of the molecule, we take the chemical identifier classes from the CHEMINF ontology and treat them as OWL annotation properties. This enables us to provide more detail in the RDF about what the identifiers are while keeping the RDF relatively simple and easy to query over.

The EBI provides pharmacological data from the ChEMBL database as RDF, as described by Willighagen et al. (43). ChEMBL contains a very heterogeneous set of pharmacological data with over 5000 different kinds of activity being reported, so in order to ensure 100% coverage of the data within ChEMBL, the authors took a non-Semantic-Web approach to the RDF, using textual strings in

the RDF to specify the activities instead of making the considerable effort of adding nearly 5000 classes to a pre-existing ontology.

We take a different approach to the ChEMBL RDF for the physicochemical properties. Because we have a relatively small (about two dozen) set of physicochemical properties as listed in Table V, we have minted classes in the CHEMINF ontology for each of them. Then we take a Davidsonian event semantics (44) approach, similar to that taken by some groups using the Ontology for Biomedical Investigations (OBI) (45). By “Davidsonian” I mean that we base everything around an event, call it *e*, which in our case is a particular execution of an algorithm to calculate a physicochemical property that takes input in the form of a connection table and produces output in the form of a calculated property, both of which we label using classes taken from CHEMINF. We relate the inputs and outputs to the event *e* with relations from OBI, specifically *has_specified_input* and *has_specified_output*, which capture those inputs and outputs that are necessary and characteristic of the process. A cheminformatics calculation, for example, is likely to need a molecular connection table and to output some calculated value. It may also take in as input the start time, or the username of the account under which it is running, and output debugging information, heat and a peculiar whirring sound from the hard drive, but those latter are not “specified” inputs or outputs in the sense that OBI uses them.

Table V. Properties calculated by ACD/Labs software for the Open PHACTS project.

<i>Kind</i>	<i>Property</i>
Bulk	log <i>P</i> , log <i>D</i> at pH 5.5, bioconcentration factor, <i>K</i> _{oc} at pH 5.5, molar refractivity, molar volume, surface tension, density at STP, flash point at 1 atm, boiling point at 1 atm, enthalpy of vaporization at STP, vapour pressure at STP
Molecular	polar surface area, polarizability, index of refraction, number of hydrogen bond acceptors, number of hydrogen bond donors, number of freely rotating bonds

Summary and Outlook

In this chapter I have given a perhaps necessarily personal overview of the current state of play for ontologies in chemistry, what ontologies are in this context, how they are being used and whom by, giving an inside view focusing on the vast datasets produced as part of the Open PHACTS project. As a scientist by training and as someone working for a learned society and often collaborating for these purposes with other scientists at, for example, the European Bioinformatics Institute, my viewpoint will be somewhat different from that of a computer scientist working on an ontology research project.

These are still early days in the field of chemistry ontologies and there is as yet much untapped potential. The approach detailed in the previous section for handling cheminformatics calculations, based partly on OBI, could, taken together

with the process ontologies (CHMO and RXNO) described above, be extended to much of the rest of chemical experimentation. Typically an experimental process will take a physical sample as input, process it in some way, and then make a measurement. In the Davidsonian approach the single event e is replaced with a chain of events e_i , the specified inputs of each event being an output or outputs of a previous event. Frey *et al.* at the University of Southampton have for a long time espoused a vision whereby a similar sort of machine-readable account of experiments is generated automatically by electronic lab notebooks (46) which are integrated with the experimental apparatus. A simple way in which this approach could benefit from ontologies is if different research groups shared the same identifiers for the different stages in their experiments. One clear opportunity is in computational chemistry, where the “experiments” are necessarily born digital. However, ontologies have historically had most traction in fields such as biocuration where the practitioners are skilled enough to take advantage of them but not having so much skill, for example at programming or data processing as to have a large collection of home-grown Perl or Python scripts to satisfy their data needs beforehand. Perhaps computational chemists are closer to the latter category than to the first.

As far as the IUPAC colour books are concerned, the Red Book (inorganic chemistry) and Blue Book (organic chemistry) have been at least partly codified by the ChEBI team, much as the Gold Book (chemistry in general) (47) and the Orange Book (analytical chemistry) have been codified in the Chemical Methods Ontology. The Green Book (48), however, is pristine and untouched. It is a varied book, acting partly as an *aide memoire* for practicing chemists, partly to instruct new chemists in the correct use of notation and the Greek alphabet, and partly to define terms used in particular fields. In the last role it could well provide a useful addendum to the Chemical Methods Ontology. Aside from careful hand curation of ontologies on inspection of the literature, which has been the route hitherto, it could well be possible to leverage the vast amount of experimental information about machines and operating conditions that is stored in CIFs in a semi-systematic way.

As far as chemists in general are concerned, the impact of ontologies has been limited. The pervasiveness of ontologies in the biomedical realm is largely due to the existence of large databases, and as the chemical sciences move more towards large databases, repositories and data-sharing mandates, and as the boundaries between journal articles, supplementary data and raw data become more fuzzy, it could well be that ontologies take a more central role in organizing chemical data than they have hitherto.

Acknowledgments

This chapter is based on a talk given at the University of Exeter in April 2013, for which invitation the author thanks Zena Wood. Thanks also to Leah McEwen for the invitation to write this chapter and to Leah and an anonymous reviewer for extensive feedback on how the first draft could be improved.

References

1. Chen, B.; Dong, Z.; Jiao, D.; Wang, H.; Zhu, Q.; Ding, Y.; Wild, D. *BMC Bioinf.* **2010**, *11*, 255.
2. Williams, A. J.; Harland, L.; Groth, P.; Pettifer, S.; Chichester, C.; Willighagen, E. L.; Evelo, C. T.; Blomberg, N.; Ecker, G.; Goble, C.; Mons, B. *Drug Discovery Today* **2012**, *17*, 1188–1198.
3. The Stanford Encyclopedia of Philosophy. <http://plato.stanford.edu/archives/spr2013/entries/logic-ontology/>, accessed on 2014-05-09.
4. SMILES – A Simplified Chemical Language. <http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html>, accessed on 2014-05-09.
5. About the InChI Standard. <http://www.inchi-trust.org/about-the-inchi-standard/>, accessed on 2014-05-09.
6. RDF 1.1 Turtle. <http://www.w3.org/TR/2014/REC-turtle-20140225/>, accessed on 2014-05-09.
7. SPARQL 1.1 Query Language. <http://www.w3.org/TR/2013/REC-sparql11-query-20130321/>, accessed on 2014-05-09.
8. Ashburner, M.; et al. *Nucleic Acids Res.* **2000**, *25*, 25.
9. WonderWeb Deliverable D17; <http://www.loa.istc.cnr.it/old/Papers/DOLCE2.1-FOL.pdf>, accessed on 2014-05-09.
10. Grenon, P.; Smith, B.; Goldberg, L. In *Ontologies in Medicine*; Pisanelli, D. M., Ed.; IOS Press: Amsterdam, 2004; pp 20–38.
11. OWL 2 Web Ontology Language Document Overview (Second Edition). <http://www.w3.org/TR/2012/REC-owl2-overview-20121211/>, accessed on 2014-05-09.
12. RDF 1.1 Concepts and Abstract Syntax. <http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/>, accessed on 2014-05-09.
13. The OBO Flat File Format Specification, version 1.2. http://www.geneontology.org/GO.format.obo-1_2.shtml, accessed on 2014-05-09.
14. OWL API Documentation. <https://github.com/owlcs/owlapi/wiki/Documentation>, accessed on 2014-05-09.
15. Golbreich, C.; Horrocks, I. In *Proceedings of the OWLED 2007 Workshop on OWL: Experiences and Directions*; CEUR Workshop Proceedings: 2007.
16. OWL 2 Web Ontology Language Manchester Syntax. <http://www.w3.org/TR/2012/NOTE-owl2-manchester-syntax-20121211/>, accessed on 2014-05-09.
17. Horridge, M.; Parsia, B.; Sattler, U. In *Proceedings of the 16th Automated Reasoning Workshop (ARW 2009)*; University of Liverpool: 2009.
18. Hastings, J.; Magka, D.; Batchelor, C.; Duan, L.; Stevens, R.; Ennis, M.; Steinbeck, C. *J. Cheminf.* **2012**, *4*, 8.
19. Richter, F. *J. Chem. Educ.* **1938**, *15*, 310.
20. *Nomenclature of Inorganic Chemistry, IUPAC Recommendations 2005*; Royal Society of Chemistry: Cambridge, 2005.
21. *Nomenclature of Organic Chemistry, IUPAC Recommendations and Preferred names 2013*; Royal Society of Chemistry: Cambridge, 2013.
22. Camon, E.; et al. *Nucleic Acids Res.* **2004**, *32*, D262–D266.

23. Hastings, J.; de Matos, P.; Dekker, A.; Ennis, M.; Harsha, B.; Hale, N.; Muthukrishnan, V.; Owen, G.; Turner, S.; Williams, M.; Steinbeck, C. *Nucleic Acids Res.* **2013**, *41* (D1), D456–D463.
24. Hill, D. P.; Adams, N.; Bada, M.; et al. *BMC Genomics* **2013**, *14*, 513.
25. Bobach, C.; Boehme, T.; Laube, U.; Pueschel, A.; Weber, L. *J. Cheminf.* **2012**, *4*, 40.
26. Magka, D. In *Proceedings of the 5th International Workshop on Semantic Web Applications and Tools for Life Sciences (SWAT4LS 2012)*, Paris, 2012; CEUR Workshop Proceedings: 2012.
27. Hastings, J.; Dumontier, M.; Hull, D.; Horridge, M.; Steinbeck, C.; Sattler, U.; Stevens, R.; Hörne, T.; Britz, K. In *Proceedings of the 7th International Workshop on OWL: Experiences and Directions (OWLED 2010)*, San Francisco, CA, 2010; CEUR Workshop Proceedings: 2010.
28. Thomas, D. G.; Papu, R. V.; Baker, N. A. *J. Biomed. Inf.* **2011**, *44*, 59–74.
29. Hastings, J.; Chepelev, L.; Willighagen, E.; Adams, N.; Steinbeck, C.; Dumontier, M. *PLOS One* **2011**, doi: 10.1371/journal.pone.0025513.
30. Chemical Methods Ontology project home. <https://code.google.com/p/rsc-cmo/>, accessed on 2014-05-09.
31. *IUPAC Compendium of Analytical Nomenclature*, 2nd ed.; Blackwell Scientific Publications: Oxford, 1987.
32. Kopka, J.; Schauer, N.; Krueger, S.; Birkemeyer, C.; Usadel, B.; Bergmüller, E.; Dörmann, P.; Weckwerth, W.; Gibon, Y.; Stitt, M.; Willmitzer, L.; Fernie, A. R.; Steinhauser, D. *Bioinformatics* **2005**, *21*, 1635–1638.
33. Carey, J. S.; Laffan, D.; Thomson, C.; Williams, M. T. *Org. Biomol. Chem.* **2006**, *4*, 2337–2347.
34. The RSC Name Reaction Ontology project home. <https://code.google.com/p/rxno/>, accessed on 2014-05-09.
35. Batchelor, C. R.; Corbett P. T. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume: Proceedings of the Demo and Poster Sessions*; Association for Computational Linguistics: 2007; pp 45–48.
36. Shotton, D.; Portwin, K.; Klyne, G.; Miles, A. *PLoS Comput. Biol.* **2009**, e1000361.
37. Corbett, P.; Batchelor, C.; Copestake A. In *Proceedings of Building and Evaluating Resources for Biomedical Text Mining at LREC 2008, Marrakech, Morocco*; 2008.
38. Kidd, R. *Integr. Biol.* **2009**, *1*, 293.
39. Pettifer, S.; McDermott, P.; Marsh, J.; Thorne, A.; Villeger, A.; Atwood, T. K. *Learned Publ.* **2011**, *24*, 207–220.
40. Describing Linked Datasets with the VoID Vocabulary. <http://www.w3.org/TR/2011/NOTE-void-20110303/>, accessed on 2014-05-09.
41. Dataset Descriptions for the Open Pharmacological Space. <http://www.openphacts.org/specs/2012/WD-datadesc-20121019/>, accessed on 2014-05-09.
42. SKOS Simple Knowledge Organization System Reference. <http://www.w3.org/TR/2009/REC-skos-reference-20090818/>, accessed on 2014-05-09.

43. Willighagen, E. L.; Waagmeester, A.; Spjuth, O.; Ansell, P.; Williams, A. J. *J. Cheminf.* **2013**, *5*, 23.
44. Donaldson, D. In *Essays on Actions and Events*, 2nd ed.; Oxford University Press: Oxford, 2001.
45. Brinkman, R. R.; et al. *J. Biomed. Semant.* **2010**, *1* (Suppl. 1), S7.
46. Bird, C. L; Frey, J. G. *Chem. Soc. Rev.* **2013**, *42*, 6754–6776.
47. IUPAC Gold Book. <http://goldbook.iupac.org/>, accessed on 2014-05-09.
48. *Quantities, Units and Symbols in Physical Chemistry*, 3rd ed.; Royal Society of Chemistry: Cambridge, 2007.

Chapter 14

Cheminformatics: Mobile Workflows and Data Sources

Alex M. Clark*

Molecular Materials Informatics, Inc., 1900 St. Jacques #302,
Montreal, Quebec, Canada H3J2S1
*E-mail: aclark@molmatinf.com

This chapter explores some of the ways that cheminformatics software is adapting to the overall industry transition toward consumer oriented mobile devices and cloud computing. While scientific software in general lags the trend due to high complexity and narrowly defined market segments, a significant amount of technical progress has been made. Mobile apps have solved the difficult challenge of providing a touchscreen user interface for drawing chemical structures, and have been demonstrated as effective ways to visualize structures (2D and 3D) and accompanying data. Effective strategies have been developed for accessing large databases using Internet protocols, as well as delegating intensive calculations to cloud-hosted servers. Mobile apps typically have a strong focus on data communication, due to their modular design, which makes it possible to execute diverse and heterogeneous workflows by concatenating the functionality of a series of apps to accomplish a given task.

Introduction

The study of cheminformatics, like any other computing discipline, is subject to the evolutionary pressure of the devices and operating systems that its practitioners have access to. Just like the personal computer ushered in a new era in the 1980s, the 2010s are witness to a platform shift of similarly tectonic scope. The first important point regarding the personal computer revolution is

that essentially none of the software from the previous mainframe era was capable of running within the constrained resources of the new devices, necessitating that all software be redesigned, rewritten and in most cases reimaged in order to fit the new platform. More importantly, the new breed of computing devices served a market that was orders of magnitude larger, which introduced an exponentially increasing variety of use cases that made no sense in the previous era when a computer was at least the size of a refrigerator and the cost had to be split over a whole department.

The shift to mobile devices is every bit as radical as the shift to personal computers. The tablets and phones that make up the device space have a fraction of the resources that their heavier predecessors have, and the modes of user interaction have very different properties. This means that migrating the user experience of a contemporary product requires a significant redesign, often resulting in a new creation that has little resemblance to its predecessor. But perhaps more importantly is that the market penetration of mobile devices is far greater than for personal computers. While it may be true for the moment that the majority of owners of a smartphone or tablet also own either a laptop or a desktop computer, the number of waking hours that a modern *digital native* spends in the presence of an always-connected mobile device often approaches saturation. Phones and tablets live in our pockets, in our travel bags, on our coffee tables, and anywhere that a few square centimeters of space can be found. We use them to look up information, communicate with friends, organize our schedules, work from wherever we are, and increasingly to pass around complex data structures, which we can visualize, modify and update.

Chemistry in particular, and science in general, has been relatively slow to embrace this new platform (1–4). Cheminformatics software has a history of being very specialized, and since it became commercially important in the late 1970s, has been for all practical purposes a tool of the pharmaceutical industry. Drug discovery chemists were the first industry segment to establish a need to manage large collections of chemical structures, and so provided the market incentive for a flourishing cottage industry of several dozen small to medium companies. This cozy relationship of software creators and consumers worked well for 30 years, but there are many disruptive trends working to change it.

The economic woes of the pharmaceutical industry and the demise of the traditional large research group in favor of a shifting landscape of startups, contract research organizations, academic researchers, and intellectual property with frequently changing owners is not well suited to the vertically integrated market that most cheminformatics software is designed for. The potential domain area for cheminformatics is also expanding beyond the narrow realm of small water-stable organic molecules, as other disciplines of chemistry working with inorganic compounds, exotic reagents, molecular materials, polymers and macromolecules continue to expand their knowledge base such that the same large data techniques are needed for new realms. Similarly, consumers of the original cheminformatics institution - libraries - increasingly need access to powerful and accurate retrieval techniques, which do not require mastering esoteric technology.

In addition the cheminformatics field has matured significantly over the last few decades, and has clearly demonstrated its value particularly with regard

to capturing, organizing and presenting data. For the most part, contemporary cheminformatics software is comparable to the early days of automobiles: owning and operating a car involved a steep learning curve for each vehicle, as well as becoming a competent mechanic. In this day and age, standardization and simplification allows us to be able to drive almost anything with 4 wheels, and depend on a supporting infrastructure of experts to perform routine maintenance and solve our problems when something goes wrong. The same evolution is needed for cheminformatics, as the technical problems migrate into the background. Like any other successful technology, those who need it will have access to software that is affordable, straightforward to learn, and for the most part *just works*. The democratization of computing that has been proceeding steadily throughout the PC era has taken a huge leap forward with the ubiquitous presence of mobile devices and web services. The consumerization of many formerly technical software categories has had an incredible effect on the expectations of the end users, who have become more demanding and less tolerant. The days when software creators could charge high prices for difficult user interfaces to calculation engines that regularly failed and needed to be coaxed into functioning correctly are almost over.

From the point of view of keeping up with the latest disruptive trends in information technology, the complexity and relatively small user base of cheminformatics software has made the field one of the later entries into the realm of mobile user interfaces. Nonetheless, the migration process is well underway, and the technological barriers are rapidly being eroded. A growing number of chemistry apps have established proof of concept functionality to demonstrate that they are suitable for a wide range of cheminformatics tasks. As an indicator of what is to come, mobile cheminformatics apps are far more important than merely adding a more portable form-factor, rather they are a key motivator for a new generation of software: *cheminformatics 2.0*.

Challenges

One of the biggest difficulties for cheminformatics software to be realized on a mobile device is the need to provide a user interface for drawing chemical structures. There are numerous software applications that have been created for drawing 2D structures using a desktop computer with a mouse (5), or the laptop equivalent. Without exception, these packages draw their style from a paradigm that was pioneered for general purpose *painting* tools from the early era of graphics workstations: most of the screen is used as a canvas, with a set of selectable tools that can be selected prior to drawing an object by clicking and dragging with the mouse.

Unfortunately this paradigm breaks down when it is applied to a tiny screen, which in the case of smartphones is often about the size of a hand. The method of interaction is by placing a finger onto the screen, which, unlike a mouse, is highly inaccurate. While a mouse pointer is accurate to the nearest pixel, a normal sized

finger can cover more than a thousand pixels, and it is not always easy to tell what object is underneath, given that the finger (and the rest of the hand) is obscuring the screen. While some users own a stylus, which could partially alleviate this problem, creating a user interface that is dependent on nonstandard peripherals would likely slow the pace of adoption by an unreasonable extent. For this reason, the best user experiences on a mobile device are had with interfaces that can be designed by offering a small set of menu choices at each step, and making all of the functionality available without the need for either accuracy or precision. Designing an interactive editor for drawing manuscript quality chemical structure diagrams under this constraint is no simple task, but nonetheless it can be done, by capturing all of the necessary *user intentions* and describing them as a small set of gestures, primitives and templates (6). By implementing a comprehensive set of algorithms for placement of chemical objects, it is possible to design an editor that allows the user to draw perfect geometries, using a miniscule touchscreen. Most importantly the time needed to draw a complex structure is quite competitive with traditional drawing interfaces.

Once the drawing interface is established, contemporary mobile devices are very capable when it comes to many kinds of visualization. Mobile apps that have been built using the native development tools for the corresponding platform have access to the raw power of the device, which provides sufficient resources for animating graphically intensive representations of data, both using 2D objects such as tables of structures and graphs, or 3D representations such as structural models and surfaces.

Nonetheless, mobile devices are generally not appropriate for the kinds of long, grinding calculations that are often carried out routinely by cheminformatics software designed for desktop computers. They are even less well suited for handling any task that involves manipulating large data collections. Ideally, mobile devices should operate on locally stored datafiles that should be considered as being cached resources of modest size, checked out from a centralized network storage resource. Software designed for mobile devices often builds on extensive integration with so-called *cloud servers*, which are virtualized servers running useful software for storing data and performing calculations. Integration with cloud computing resources is a way to make up for the limited performance and capacity of mobile devices, but it also imposes limitations of its own, such as necessitating access to a network connection and a reliable server, as well as the additional engineering complexity of designing a client-server workflow.

Early History

The earliest mobile apps for handling chemical structures were experimental attempts to migrate functionality to the new platform, which, by 2010, was gaining significant traction in almost every major segment of the software industry. Early products typically had limited focus on accomplishing important workflow tasks, and many were hobby projects or marketeering efforts to draw attention to existing desktop-based products. Important success stories include

visualization apps, especially for 3D structures (e.g. Molecules (7), PyMol (8)) and apps for searching online catalogs (e.g. Mobile Reagents (9), ChemSpider Mobile (10), SPRESImobile (11)). Some of the first apps to include the ability to draw chemical structures did so for the purpose of sketching search queries, which does not require the ability to create manuscript quality diagrams, so a simple interface can provide useful functionality. Also, a large number of simple reference apps were created, often for educational purposes. These apps by and large skirted some of the more difficult technical barriers to implementation on mobile devices, and so were successful at accomplishing their goals.

The introduction of the Mobile Molecular DataSheet (MMDS) (12) app brought the first manuscript-quality chemical diagram sketcher to the mobile platform, and enabled flexible workflows that revolved around managing collections of structures with a table-like schema, referred to as *datasheets*. By providing the ability to create, edit, view, import, export, share and utilize datasheets for a variety of different purposes, the technical proof of concept established that many cheminformatics tasks were viable on small touchscreen devices.

Chemical Structures

The fundamental datastructure of cheminformatics is the representation of chemical structures, using formats that are often referred to as *connection tables*, which describe the molecular species as a labeled graph. There are many different variations, but all share some fundamental properties: nodes represent atoms and edges represent bonds (13). Atoms are assigned properties such as element, formal charge, unpaired electrons, isotope, etc. Bonds are typically labeled by bond order. Most representations provide 2D or 3D coordinates for atoms (for diagram-style representations and conformations, respectively), or in some cases coordinates are omitted. Properties such as stereochemistry can be assigned using atom or bond labels, depending on the coordinate style.

Most 2D cheminformatics algorithms operate in a way that does not require atom coordinates (as long as stereochemistry is properly encoded). Nonetheless, most structure collections maintain 2D diagram-style coordinates, because the ability to be able to render the structures using the stylistic conventions of chemistry is almost always a vital part of any workflow. Many cheminformatics applications produce chemical diagrams as part of the deliverable (e.g. reports showing structure-activity correlations), while techniques such as model building typically require a certain amount of inspection during the validation process. In order to present a large number of structures to a chemist, 2D diagrams are the most ergonomic viewing method (14). One problematic misunderstanding that has plagued cheminformatics since the beginning is that there is a difference between the software user interfaces and datastructures used to represent a diagram for presentation purposes (15), and those used to represent a machine-readable structure. These two goals are sufficiently similar that it is quite possible to use the same software to accomplish both tasks, but when representing structures for

cheminformatics purposes, it is important to avoid using annotations that have only graphical meaning, at the expense of representing the chemistry in a way that can be interpreted by an algorithm. Common problems include resonance bonds, stereochemistry, implicit hydrogen counts and deceptive choice of bond order or charge localization (16).

With the establishment of a datastructure and file format suitable for representing individual molecular entities, it is useful to define higher layers of abstraction, for associating molecules with properties such as name, physical properties, references, etc., as well as grouping together collections. Molecules can be combined into sets to represent composite entities such as chemical reactions, groups of molecules related by tautomer shifts, partitioning between scaffold templates and substituent fragments, etc.

Software for working with collections of structures and associated data (and metadata) has traditionally been provided for use on workstation class computing devices, a category that grew to include personal computers and then laptops (17). Locally stored data is often supplemented by compound registration databases running on a centralized database server, and resource-intensive tasks may be offloaded onto a cluster or computing grid. More recently, web services have become a popular way of delivering functionality such as property calculations or structure lookups (18–21). These services are often supported by web-based interfaces, many of which provide the ability to draw chemical structures, for use in constructing queries.

The migration of important cheminformatics functionality onto Internet-accessible servers via a well-defined API (so-called *cloud computing*) is a welcome development for a number of reasons, one of them being that the user interface and the functionality become disconnected. Rather than having a single vendor-provided interface tightly bound to the feature set, any number of applications can make use of the subset of functionality that they need, and provide a user experience to match the task at hand. This paradigm is particularly beneficial to mobile apps, which can generally muster enough resources to provide a compelling user experience, but need to offload any kinds of calculations that require heavy computational capacity, or need to access large centralized data sources.

Practical applications of cheminformatics can be thought of as workflows, whereby the user begins with an objective, and some notion of how to get there: at the beginning of workflow, some data may be already available in digital form, while other data may need to be entered as part of the process. The goal may be some combination of gaining insight into a chemical problem, producing a more focused dataset, or preparing presentation quality graphics for communication purposes. A workflow may be a single step (e.g. looking up a chemical in a database) or it may be many steps, in some cases using diverse methodologies and heterogeneous platforms. The use of mobile apps and cloud-hosted web services are quite a natural fit for composing workflows based on best-of-breed technologies, since data communication is such an integral capability. The remainder of this chapter will describe some of the ways in which workflows can be accomplished using mobile apps as the primary user interface, and cloud-hosted services for heavy duty calculations and large data storage.

Creating Content

Creation of new data for cheminformatics purposes is to a large extent predicated on an effective way to allow the users to draw structures. Ideally the interface should allow the user to quickly draw complicated structures, at publication quality, using a palm-sized touchscreen device. The classic interface paradigms that were developed during the desktop era are unable to achieve all of these goals, and so difficult decisions need to be made: an effective sketcher requires a major redesign, which in turn introduces a learning curve to new users. A number of app creators have opted to port the traditional interface design to the mobile form factor, including *JSDraw* (22), *ChemJuice* (23), *Chirys Draw* (24), *ChemDoodle* (25), *ChemWriter* (26), *Elemental* (27) and more recently *ChemDraw* (28). These interfaces are significantly more viable on the larger tablet form factor, and some products, such as *ChemDraw*, have opted to target tablets exclusively, having reached the conclusion that phone-sized devices are not viable.

Heavy-duty cheminformatics apps, such as the *Mobile Molecular DataSheet* and *SAR Table* (29) make use of a different interface that has been redesigned to operate as touchscreen-optimized gestures and context-specific menu actions, which are capable of performing the necessary geometry placements and template fusion calculations necessary to draw difficult structures, to a standard suitable for publication. The implementation is described in detail in the literature (6). The interface is fast and efficient on small devices such as iPhones, but the caveat is the additional learning curve that is necessary to take advantage of these capabilities.

Other products, such as *MolPrime* (30, 31) and *ChemSpider Mobile*, offer the user the choice between the expert mode and a simplified, traditional sketcher interface, which is easy to learn and effective for straightforward tasks such as drawing small molecules for search queries, as shown in Figure 1.

Reaction editors fall into two categories, one of which is the free canvas approach, wherein the user draws each molecular species onto a single ensemble and links them together using arrows and the plus symbol. *ChemDraw* and *Chirys Draw* are two examples of apps that take this approach. The alternative is to represent each reaction component separately, which has two advantages: the sketcher is not overburdened by the need to represent additional molecular entities, and the more rigidly defined format makes features such as reaction balancing assistance more effective. Figure 2 shows the *Reaction101* (32) app rendering a reaction that has a total of 3 components: two reactants and one product, as well as the *Yield101* (33) app, which provides additional markup by assigning stoichiometry-normalized quantity information. The use component wise representation is inherently more suitable for electronic lab notebooks.

Besides apps that allow creation of single molecules or reaction schemes, there are also higher end tools that allow management of collections. As shown in Figure 3, the *Mobile Molecular DataSheet* (MMDS) provides an interface to creating and organizing datasheets, which are collections of molecules, reactions, scalar data and higher order markup. The *SAR Table* app specializes in datasheets, which organize the structural makeup of each molecule into a scaffold and some number of substituent fragments, which can be browsed, edited and manipulated.

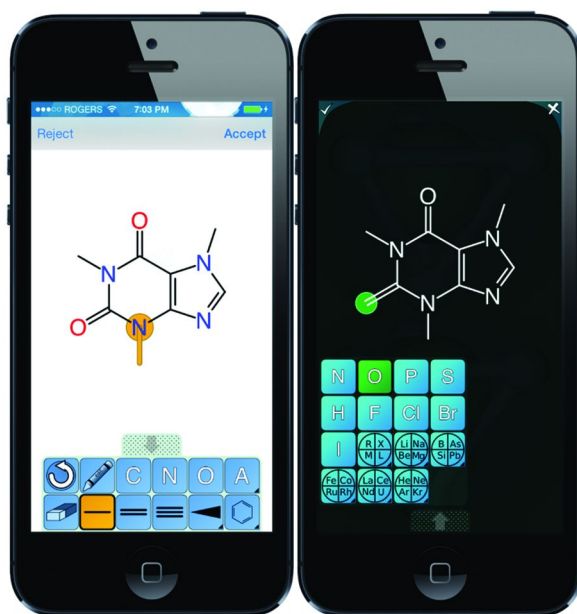


Figure 1. Simplified casual drawing interface (left), advanced gesture-based primitives (right).



Figure 2. Component wise reaction editor used by Reaction101 (top), and the addition of quantity metrics using Yield101 (bottom).



Figure 3. Collections of datasheets, MMDS (left) and SAR Table (right).

Importing Content

While a number of useful apps can function well as standalone products (e.g. database searching apps like *Mobile Reagents* or *SPRESImobile*), the ability to import data from elsewhere is essential to being able to use an app as part of a workflow. The more useful chemistry apps typically have the ability to recognize at least one standard file format and import it using algorithms that are incorporated into the app (e.g. SketchEl and MDL Molfile). Some apps, such as the Mobile Molecular DataSheet (MMDS) can make use of a web service to import formats such as the Chemical Markup Language (CML) (34) and ChemDraw (binary and XML) (35).

The system clipboard shares the same lowest common denominator functionality on mobile and desktop platforms: an app can place arbitrary text into a global repository, and any other app can read the text. By placing text representations of molecules using standard formats, e.g. MDL Molfile or SketchEl, apps can provide a simple user interface for transferring structures to other parts of the same app, or to other apps.

Several importing mechanisms are made available on the iOS platform by having each app register some number of file types, which are recognized by either MIME type or extension. This opens up three main ways of bring data into an app: download files from within the web browser, unpacking mail attachments and launching files from other apps. The last form is a kind of interprocess communication.

Because mobile devices provide convenient class libraries for making HTTP calls, use of web service APIs is a flexible way to locate data and download it to the device. Common examples include services for searching molecules (transforming a simple query into a list of structures and data) and authenticated access to remote file systems (such as Dropbox (36)).

Exporting Content

For many cheminformatics tools, the software is only as good as its ability to export its data. An app that can import, process and export can be treated as a node within a workflow, and can be designed as a modular component that focuses on a limited domain of functionality.

When exporting, chemists generally have one of two objectives: to hand off machine-readable cheminformatics data, or to create graphics for a publication of some kind.

If the app is exporting its native format, or a data or graphics format that it natively capable of interconverting to, the content can be prepared on the device. Data formats that are essentially one-to-one conversions (e.g. SketchEl to MDL Molfile) are typically done locally, while conversion to formats that require significant processing (e.g. extended MDL Molfile (16), SMILES (37), InChI (38, 39)) may require support from a web service. Similarly for graphics, the generation of bitmaps is done by the same process that is used to render a structure onscreen, so adding the additional functionality to the app is a trivial matter. Creating various kinds of vector graphics may be done on the device, or by making use of a web service. Vector graphics formats have significant advantages for manuscript preparation, and these include Portable Document Format (PDF), Encapsulated PostScript (EPS), Scalable Vector Graphics (SVG) and Microsoft Word & Excel with embedded DrawingML vector diagrams (40).

Exporting of data can be done locally by using the system clipboard. On the iOS platform, this can be in the form of text or a bitmapped image. The former is useful for pasting structures into other input forms, while the latter is particularly useful when creating general-purpose presentations using the mobile device, e.g. presentations using Apple's Keynote app (41).

For sharing and data transfer, one of the most versatile and robust methods is to initiate an outgoing email and incorporate one or more attachments containing data and/or graphics. Sending an email to oneself is a very effective way to transfer data in standard formats between heterogeneous computing platforms operating over a network, and it can also be useful as a way to backup data, since many email services can store the content indefinitely. Sending an email to colleagues is a simple but practical way to communicate data for collaboration purposes, since the recipient can open the attachment and make use of it with any software that understands the format.

A more sophisticated way to collaborate with chemical data is to use a remote file system such as Dropbox. By using the native Dropbox app as a file browser, or a chemically aware client such as the *MolSync* (42) app, the online content can be organized, managed and synchronized. By leveraging the intrinsic features of a platform like Dropbox, multiple users can share files and folders, and actively

work on chemical documents simultaneously, and with a high degree of confidence regarding security.

Publicly sharing data on the open internet is also an option, though it requires a service that can persistently retain the data. Various kinds of photo-sharing sites are insufficient, since the browsable result needs to be able to present the data in graphical form *and* allow access to the data itself. A number of apps are capable of uploading their data (molecules, reactions, collections) to a service hosted by *molsync.com*, such as shown in Figure 4. Once the upload is complete, the service generates a unique URL, which can be shared. The browser page renders the data in a viewable form, and allows users to access the data, either in its own native format, or one of the formats that the underlying libraries are able to create. It also allows the user to custom-generate graphics in a variety of different bitmap and vector formats, and select color schemes and sizing metrics. In addition to creating the sharable link, in-app integration with social networks such as Twitter allow the user to emit the content and publicize it within their network of friends and associates.



Figure 4. Sharing a molecule from MolPrime⁺ (left), viewing the resulting web page (right).

Structure-Based Calculations

A number of apps provide basic structure-based property calculation features (e.g. *Elemental*, *MolPrime*). There are a number of simple properties that should always be calculated on the device itself, such as molecular weight and formula. In addition to simple scalar properties, there are those which require access to more sophisticated functionality, and make practical sense to operate as a web service some combination of reasons, e.g. computational intensity, the need to have access to large datasets, or because of a complex or legacy codebase built with incompatible development tools.

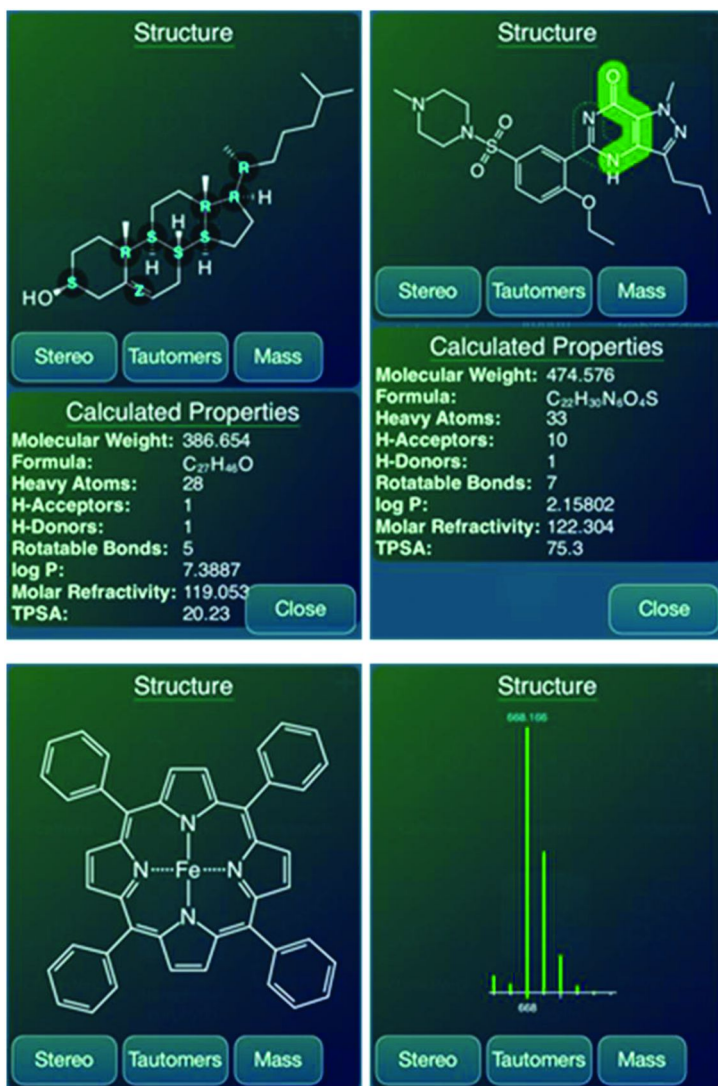


Figure 5. MolPrime⁺ property calculation, including scalar properties, stereochemistry labels, tautomer transforms and mass distribution.

The MolPrime⁺ app provides a number of different types of calculations based on a single molecular structure, as shown in Figure 5. Trivial scalar properties are calculated within the app itself, but more difficult properties such as octanol partitioning coefficient (log P), molar refractivity and topological polar surface area (TPSA) are obtained by transmitting the structure to a dedicated web service. Non-scalar properties can also be calculated: chirality (R,S) and double-bond stereochemistry (Z,E) labels can be obtained and displayed as an overlay. For tautomer transformations, the web service returns a *graph* of

transforms, which allows the app to present the information interactively: the tautomer shift pathways behave like buttons. Pressing any of them causes the transformation to be applied, and the new structure displayed in its place, with its own set of available transforms. The app can also display a mass distribution, which is a useful tool for interpreting the results from a mass spectroscopy experiment.

The *Mobile Molecular DataSheet* (MMDS) app includes the ability to calculate scalar properties for a collection of molecules, rather than just one at a time. It also interfaces with the Open Notebook Science melting point calculation, via a documented web service protocol (43). Other types of calculations include assistance with reaction balancing, which is one of the primary features of the *Reaction101* app: the always-on display of leftover element counts for unbalanced reactions is a useful guide, especially for students who are new to chemistry, and the app also has an auto-balance feature.

Representing Structures

Besides providing editing and visualization of structures, reactions and collections, some efforts have been made to design apps that go beyond the core functionality of cheminformatics. One example is the *SAR Table* app, which was originally designed to solve the twin problems of creating manuscript figures involving a series of related compounds, and of data entry of these structures after publication.

Figure 6 shows the redundant structure representation used by the *SAR Table* app: the fragment columns denoted by **Scaffold**, **R1**, **R2** and **R3** are sufficient to provide the composition of the **Molecule** column. The whole molecule is color-coded to show the common scaffold portion, and it is automatically rebuilt each time the user modifies one of the fragment fields. This allows the app to present a user interface that involves the absolute minimum amount of redrawing: because the molecules are part of a series, there are numerous possibilities for reuse, since there are usually just a handful of scaffolds, and there are often common substituents (e.g. hydrogen, methyl, ethyl, etc.). There are two main advantages of this approach: (1) by providing an efficient interface to copy fragments from one cell to another, the data entry time can be greatly reduced; and (2) by synchronizing the composite molecule so that it is always consistent with the fragment definitions, it is straightforward for the operator to verify the actual structure, and identify mistakes.

The app provides strong export capabilities, e.g. creation of a Microsoft Word (.docx) file with an embedded table containing vector diagrams of the structures and fragments, which can then be edited and incorporated into a manuscript. Data can also be exported in common formats like MDL SDfile, which is useful for reentering literature data and incorporating it into efforts such as QSAR studies. It is also possible to import data from molecule collections, such as MDL SDfiles, then use semi-automated scaffold matching algorithms to decompose the incoming structures into the scaffold-substituent representation. As well as data creation, the app can also be used to search-and-match public databases based on scaffold templates, use a web service to build a structure-activity model with known data

and apply it to missing fields, and plot degrees of freedom against each other in a matrix form, e.g. **R1** vs. **R2**.

These kinds of higher order, domain-specific interfaces are still relatively unusual in the ecosystem of mobile chemistry apps, but they are growing in scope, and represent a competent alternative to moderately sophisticated desktop-based products.

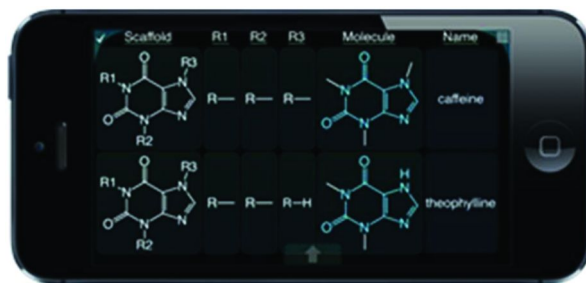


Figure 6. Representation of molecules in terms of scaffolds and substituents.

Conclusion

The introduction of mobile apps for cheminformatics is a recent development, but already there is a significant ecosystem in place for a variety of cheminformatics workflow tasks, molecular modeling, access to reference data and information-rich communication and collaboration methods. The scope and functionality of mobile apps grows by the month, and the list of tasks that no longer require a desktop or laptop computer grows with it. The roll call of companies that now have a presence by way of at least one native app now includes most of the larger companies from the industry, as well as dozens of startups.

It is easy to mistake the importance of this progression as adding value primarily because of the improved mobility of the latest generation of devices. The migration of cheminformatics to mobile platforms presents a timely opportunity to reevaluate more than 30 years of incremental development during the personal computing era, and selectively choose from the methodologies found to be most successful. Since the ways in which software workflows are carried out need to be redesigned anyway, the pressure to conform to suboptimal legacy design decisions is greatly reduced.

The influx of non-expert users who expect mobile apps to provide flawless functionality with a minimal learning curve and a delightful user experience is also in stark contrast to the basement-dwelling troglodyte stereotype of expert cheminformatics software practitioners, and is already exerting its influence. Software creators are being held to the same high standards as mass market lifestyle apps, and this is forcing a difficult but ultimately welcome readjustment.

With regard to the development of supporting algorithms, the need to provide web services to bring mobile apps up to par with desktop equivalents is also driving higher standards. Computational algorithms that are decoupled from their user

interface need to be designed with greater rigor, and their functionality clearly defined and bridged via a clean and sensible API. The performance profile and domain applicability of each service has to be well studied, and there is much less tolerance for mismatch than would be the case with a desktop application or a command line tool. The demands for reliability are much higher, since a non-terminating loop or fatal crash can bring down an entire server, which affects more than just the operator.

The inherent modularity of apps means that passing data around is a significantly more frequent operation, and often involves a diverse selection of tools, platforms and collaborators with different specialties. The increased need for communication and interoperability means that more attention needs to be paid to the use of standard file formats, and for ensuring that chemical concepts can be losslessly transmitted.

Having powerful and well-crafted cheminformatics workflow tools on mobile devices opens up this relatively niche field to a far wider audience of non-experts. Making this work involves a combination of increased simplification of the tools, but also needs to be met in the middle by an improvement in the level of informatic literacy of chemists. It is important to keep in mind that science will never be simple, no matter how user friendly the software becomes.

Mobile apps also bring with them an expectation of affordability, which is counterbalanced by having access to a larger customer base, which is in turn highly disruptive to a number of established vendors. Adapting a business model from a vertical enterprise niche to a horizontal consumer mass market will continue to be a challenge, for both established and upstart vendors alike.

Despite the potential of mobile apps and the rapidity of their adoption, for the foreseeable future large segments of cheminformatics will continue to be done on computing platforms that resemble contemporary desktop or laptop computers. The strong trend toward modularity and simplification that mobile app users demand is effective for well-established workflows, which can be implemented one at a time, but scientific research always conspires to provide unanticipated scenarios. It should be expected that porting a unit of scientific functionality to a mobile app is a part of the maturation process: once a method becomes routine, there will be an app for that.

References

1. Williams, A. J.; Ekins, S.; Clark, A. M.; Jack, J. J.; Apodaca, R. L. Mobile apps for chemistry in the world of drug discovery. *Drug Discovery Today* **2011**, *16*, 928–939.
2. Clark, A. M.; Ekins, S.; Williams, A. J. Redefining cheminformatics with intuitive collaborative mobile apps. *Mol. Inf.* **2012**, *31*, 569–584.
3. Ekins, S.; Clark, A. M.; Williams, A. J. Cheminformatics workflows using mobile apps. *Chem-Bio Inf. J.* **2013**, *13*, 1–18.
4. Ekins, S.; Waller, C. L.; Bradley, M. P.; Clark, A. M.; Williams, A. J. Four disruptive strategies for removing drug discovery bottlenecks. *Drug Discovery Today* **2013**, *18*, 265–271.

5. Ertl, P. Molecular structure input on the web. *J. Cheminf.* **2010**, *2*, 1.
6. Clark, A. M. Basic primitives for molecular diagram sketching. *J. Cheminf.* **2010**, *2*, 8.
7. Molecules. iTunes Preview. <http://itunes.apple.com/app/molecules/id284943090> (accessed March 7, 2014).
8. PyMOL. iTunes Preview. <http://itunes.apple.com/app/pymol/id548668638> (accessed March 7, 2014).
9. iTunes. <http://itunes.apple.com/ca/app/mobile-reagents-universal/id417616789> (accessed March 7, 2014).
10. ChemSpider. iTunes Preview. <http://itunes.apple.com/app/chemspider/id458878661> (accessed March 7, 2014).
11. SPRESImobile by InfoChem. iTunes Preview. <http://itunes.apple.com/app/spresimobile-by-infochem/id505308290> (accessed March 7, 2014).
12. Mobile Molecular DataSheet. iTunes Preview. <http://itunes.apple.com/app/mobile-molecular-datasheet/id383661863> (accessed March 7, 2014).
13. Warr, W. A. *WIREs Comput. Mol. Sci.* **2011**, *1*, 557–579.
14. 2D Structure Depiction. Clark, A. M.; Labute, P.; Santavy, M. *J. Chem. Inf. Model.* **2006**, *46*, 1107–1123.
15. Brecher, J. Graphical representation standards for chemical structure diagrams. *Pure Appl. Chem.* **2008**, *80*, 277–410.
16. Clark, A. M. Accurate specification of molecular structures: The case for zero-order bonds and explicit hydrogen counting. *J. Chem. Inf. Model.* **2011**, *52*, 3149–3157.
17. The contemporary definition typically refers to any computer with a mouse-like pointing device, a relatively large screen, and a high performance CPU, running an operating system not originally designed for running on battery power.
18. Canny, S. A.; Cruz, Y.; Southern, M. R.; Griffin, P. R. PubChem promiscuity: A web resource for gathering compound promiscuity data from PubChem. *Bioinformatics* **2012**, *28*, 140–141.
19. Willighagen, E. L.; Waagmeester, A.; Spjuth, O.; Ansell, P.; Williams, A. J.; Tkachenko, V.; Hastings, J.; Chen, B.; Wild, D. J. The ChEMBL database as linked open data. *J. Cheminf.* **2013**, *5*, 23.
20. Bolton, E. E.; Chen, J.; Kim, S.; Han, L.; He, S.; Shi, W.; Simonyan, V.; Sun, Y.; Thiessen, P. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem3D: A new resource for scientists. *J. Cheminf.* **2011**, *3*, 32.
21. Jeliaskova, N.; Jeliaskov, V. AMBIT RESTful web services: An implementation of the OpenTox application programming interface. *J. Cheminf.* **2011**, *3*, 18.
22. Scilligence. <http://www.scilligence.com/web/jsdraw.aspx> (accessed March 7, 2014).
23. ChemJuice. iTunes Preview. <http://itunes.apple.com/app/chemjuice/id342895394> (accessed March 7, 2014).
24. <http://itunes.apple.com/app/chirys-draw/id455125162> (accessed March 7, 2014).
25. Chirys Draw. iTunes Preview. <http://itunes.apple.com/app/chemdoodle-mobile/id435468742> (accessed March 7, 2014).

26. ChemWriter. <http://chemwriter.com> (accessed March 7, 2014).
27. Elemental. iTunes Preview. <http://itunes.apple.com/app/elemental/id518655328> (accessed March 7, 2014).
28. ChemDraw. iTunes Preview. <http://itunes.apple.com/app/chemdraw/id631620841> (accessed March 7, 2014).
29. SAR Table. iTunes Preview. <http://itunes.apple.com/app/sar-table/id477451419> (accessed March 7, 2014).
30. MolPrime. iTunes Preview. <http://itunes.apple.com/app/molprime/id437087077> (accessed March 7, 2014).
31. MolPrime+. iTunes Preview. <http://itunes.apple.com/app/molprime+/id497295446> (accessed March 7, 2014).
32. Reaction101. iTunes Preview. <http://itunes.apple.com/app/reaction101/id423115765> (accessed March 7, 2014).
33. Yield101. iTunes Preview. <http://itunes.apple.com/app/yield101/id433416999> (accessed March 7, 2014).
34. Phadungsukanan, W.; Kraft, M.; Townsend, J. A.; Murray-Rust, P. The semantics of Chemical Markup Language (CML) for computational chemistry: CompChem. *J. Cheminf.* **2012**, *4*, 15.
35. CDX. <http://www.cambridgesoft.com/services/documentation/sdk/chemdraw/cdx/General.htm> (accessed March 7, 2014).
36. Dropbox. <http://www.dropbox.com> (accessed March 7, 2014).
37. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
38. McNaught, A. The IUPAC International Chemical Identifier: InChI. *Chem. Int.* **2006**, *28*, 12–14.
39. Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI - the worldwide chemical structure identifier standard. *J. Cheminf.* **2013**, *5*, 7.
40. Clark, A. M. Rendering molecular sketches for publication quality output. *Mol. Inf.* **2013**, *32*, 291–301.
41. Keynote. Apple. <http://www.apple.com/mac/keynote> (accessed March 7, 2014).
42. MolSync. iTunes Preview. <http://itunes.apple.com/app/molsync/id461044999> (accessed March 7, 2014).
43. WebServices Protocol. <http://molmatinf.com/websvcproto.html> (accessed March 7, 2014).

Chapter 15

Tying It All Together: Information Management for Practicing Chemists

Steven M. Bachrach^{*,1} and Carmen I. Nitsche^{*,2}

¹Department of Chemistry, Trinity University, 1 Trinity Place, San Antonio, Texas 78212

²CINforma Consulting, 254 Rockhill, San Antonio, Texas 78209

*E-mail: sbachrach@trinity.edu; cnitsche@swbell.net

As the world-wide web enters into its third decade, the tools for a significant reimagining of the practice whereby chemists handle and process information are largely in place. In this article, we discuss how the Internet is changing chemical publication in both form and function, how chemical information is obtained (especially on mobile devices through the cloud), how electronic laboratory notebooks enhance the research enterprise, and how social media is bringing about new means for collaboration. The potential now exists for revolutionary change to the practices of chemists.

While the Internet dates back to the 1960s, most people probably became aware of the Internet subsequent to the development of the first modern web browser, *Mosaic*, in 1994 (1). Within a few years, the Internet became ingrained within popular culture, rapidly becoming the way people shopped (think *Amazon*) and read newspapers (think *Huffington Post*) and acquired music (think *iTunes*) and watched videos (think *YouTube*) and kept up with their friends and family (think *Facebook*).

The web altered the landscape of sciences as well, with many common practices of chemists substantively changed by the accessibility of information and the opportunity to collaborate now available through the magic of the web browser and the web site. Some of the first baby steps in this direction were presented in the 1996 book one of us edited *The Internet: A Guide for Chemists* (2). Much of this book now seems quaint and our vision of the future was a bit

under-optimistic in parts, while some ideals we presented have yet to come to fruition. Recognizing this assessment places a sense of constraint on what we will comment upon in this essay – a fear of not being sufficiently visionary along with a dread that some goals may be unattainable even over the next 15 years. What we will attempt to do here is to discuss a few areas in which the Internet might affect practicing chemists, with a view towards what areas might see some real dramatic changes in the not-too-distant future.

Scientific Publication

Perhaps the most obvious change to daily behavior of most chemists is that we no longer read hardcopy (printed) versions of journals. Rather, journals are delivered electronically and we read them at our desks or on our mobile devices. We no longer need to physically walk over to the library or receive a journal in the mail.

That being said, what has really changed is *only* the delivery method. The form of most journal articles remains unchanged. The majority of us read articles in their PDF form, a format designed to reproduce the look and feel of the traditional article printed on a piece of physical paper. Most people likely print the PDF to paper and read it in hardcopy, rather than off of a screen. We will come back to the implications of this situation and the opportunities that electronic media provide for enhanced publication.

Open Access

The most publicly visible new development to scientific publication that occurred with the advent of the web is the Open Access (OA) movement. The major driver behind the OA movement is to make scientific publications available to everyone around the world. This ultimately means that articles need to be made available at no cost, removing any financial barriers to access. OA was inspired in part by the “serials crisis” (3, 4), the decades-old trend of increasing subscription prices for STM journals coupled with flat, if not declining, library budgets, leading to fewer subscriptions and ever more limited access to scientific publications.

While OA has grown as a movement, and virtually all STM publishers now offer some form of OA, many problems exist. First is a lack of uniformity as to just what the term “Open Access” means. While the Budapest Open Access Initiative (5) and the Bethesda Statement on Open Access (6) led the charge on OA, the more widely used definitions are so-called “Green Open Access” and “Gold Open Access” (7). With “Green” OA, a publisher produces a copy-edited, typeset version of the article and makes it available through a journal with some fee assessed (via subscription or pay-per-view). In addition, the author retains the right to make the version he sent to the publisher available through a personal web

site or through an institutional repository at no cost to any reader. In “Gold” OA, the publisher charges no fee for access to any article in the journal.

The key element of OA is a shift of the revenue generation from the consumer (the reader, principally through the library) to the author. OA journals assess a fee on the author to cover the publication cost. Hybrid OA journals are ones which publish some articles that can be read only through subscription or pay-per-view while some articles are made available to all at no cost by the author paying a fee.

An important consideration here is that most authors and most publishers have focused on the *cost* issue, but the *rights* issue is critical as well. The distinction here is analogous to the difference between “free beer” and “free speech”, now referred to as *gratis* and *libre*. In *gratis* OA, the reader pays nothing to access an article, but they have no other rights to the article. The full force of copyright is in effect. In *libre* publication, the reader is granted some rights, if not unlimited rights, to the article. This can mean allowing a reader to mine the journal contents, reproduce figures and table, or even reproduce the article in its entirety. Many publishers use the *Creative Commons* (8) licenses to control rights.

At the current time (late 2013) the resolution of who will pay for STM publication is very unclear. We are faced with a mixture of publication models with no clear winner and no clear direction. It seems the marketplace is very much undecided as to which model it prefers. The situation is made murkier by a rash of new OA publishers, some of which appear to practice in a predatory fashion: offering journals intended to be confused with established journals, offering phony editorial boards, sponsoring conferences with the sole goal of garnering articles, etc. (9). The role of government and government agencies is also unclear and evolving. In the United States, the NIH has a mandate upon authors to make their NIH-sponsored work available for free after an embargo period (10). The Office of Science and Technology has issued a memorandum regarding making all-government sponsored research available to the public (11). In response, the Association of Research Libraries (ARL), the Association of American Universities (AAU) and the Association of Public and Land-grant Universities (APLU) have put forward the SHARE (Shared Access Research Ecosystem) proposal, which would establish a federated collection of institutional repositories (12). Domestic and international publishers have proposed the Clearinghouse for Open Research for the United States (CHORUS) (13). In the United Kingdom, the Research Councils UK (RCUK) has its own open access policy (14).

To us, the serious question, and one that has not been widely acknowledged, is how the economics of STM publishing is improved under OA (15). There is a cost to publishing: the servers need to be acquired and maintained, articles should be reviewed and copy-edited, meta-data established, DOIs registered, etc. And somewhere a profit must be made too; many of the societies who act as non-profit publishers need to make money to support the societies’ other functions. In the pre-OA, subscription-only world, subscription rates were steady climbing at a rate greater than inflation. If we move to a fully OA-world, where authors foot the bill entirely, where do the cost savings come from? How will authors be able to pay for, presumably, ever-increasing article publication charges? We anticipate the next decade to be a rather bumpy ride for publishers of all stripes as we shake out the winning and losing publication models.

Data Publication and Open Data

Until the advent of the Internet, chemistry publication was greatly restricted by the medium itself. Information was disseminated solely on paper. There were limits as to how long an article could be, and, therefore, much data was often omitted simply to save space. Supplementary materials were sometimes made available, typically through unwieldy media like microfiche.

All that changed with the web. Disk space is cheap. Large data sets can now be widely disseminated at little cost. Page limits evaporate since one is not tied to producing a physical “page”. This has manifested in more published articles, the potential for longer articles, and greater deposition of data into supporting materials.

Since the supporting materials are delivered electronically, these materials can be made up of any file type whatsoever: text, images, movies, etc. More interesting is that data could be deposited in native form, so x-ray structures can be deposited as CIF files, computational chemistry program output can be deposited, spectral data can be deposited as JCAMP-DX, etc. The distinct advantage here is that these data files can then be reused without loss by the reader piping the files into their favorite software tools. This is what we and others have termed “enhanced publication” (16–19). By utilizing the technology afforded by the Internet, we can allow for a much richer publication stream, where the reader can actually manipulate the data that the authors have generated.

Unfortunately, adoption of this enhanced publication system has been very slow. Most data deposited into supporting materials today is in the form of pdf files, a means of actually stripping out most of the data content and format, and making it very difficult for the reader to reuse. For example, instead of depositing a JCAMP file, which contains the full spectral data that can be reused in a variety of programs, authors are typically depositing just an image of the spectrum.

We are optimistic that a greater percentage of re-useable data deposition will take place over the upcoming decade. Reviewers and editors are becoming more aware of the need to get supporting data into the hands of other scientists. Many editors have already dictated that some data must be deposited as a condition of publication: for example, many journals demand deposition of the results of x-ray structure analyses as CIF files.

One key question is where data should be deposited. Most data currently is being deposited as supporting (supplementary) materials associated with an article and held by the publisher. A few organizations have emerged as data warehouses, such as the *Cambridge Crystallographic Data Centre (CCDC)* and the *RSCB Protein Data Bank (PDB)*. This is, we believe, a very good way to house data, as data experts in a specific field are best capable of selecting appropriate formats, curating the data, and planning for migration to future formats and media. An alternative approach is represented by *Figshare*, a venture project to store, index, and distribute data. An interesting aspect of the *Figshare* approach is not just the broad range of supported data, but the inherent ability to cite the data and provide credit to the depositor. This last point, making sure that the creators of data be properly credited, is a crucial element to the scientific process. We believe that

what really matters to scientists when it comes to their scientific communication is not the rights associated with the publication, nor any monetary reward, but rather proper recognition for the work they have accomplished, for being known as the first to create an idea, or synthesize a molecule, or generate the spectrum. Whatever data deposition system(s) are developed in the future, they must have this as a key element of the depository. A very good system for identifying the author is ORCID (20), which associates an author with a unique identification number. The metadata for a data deposition could thus include the ORCID to unambiguously identify the author.

Another essential element of data deposition must be its discoverability, meaning the ability for someone to search and find that data. This will require careful crafting of metadata, which again advocates for discipline-specific design and control of data depositories like *CCDC* and *PDB*.

For chemists, an essential metadata component for any chemical dataset must be the compound or compounds included in the data. The *InChI Project* (21) provides a clear means for providing this metadata. The InChI string is an alphanumeric representation of the chemical structure (22, 23). It is both Open Source and free to use, so there are no restrictions on its use as metadata. The InChI provides a unique identifier and is applicable to a vast majority of chemical compounds. We anticipate that the current broad use of InChI will continue to expand and it will become the metadata of choice for indicating chemical structure within most data formats.

The last major issue regarding data distribution is the notion of rights or restrictions associated with any data set. We advocate for Open Data, a complete sharing of data amongst scientists with no restrictions on use and reuse, other than attribution. Scientists should be free to gather any data and use it in whatever form they wish. This might mean mining across large data sets. It might mean comparing data generated by a colleague with data generated in her own lab. It might mean collecting data from a variety of open sources and creating a new data set to be shared with others. At the end of 2013, many major STM publishers signed on to a roadmap delineating non-commercial data-mining rights in the EU (24).

Open Data will facilitate collaboration and corroboration. It should be a required element of all scientific publication; when an author submits a manuscript for consideration, the editor should require that all data generated for that work be made available to reviewers, and upon publication, all that data should be deposited as Open Data in the appropriate repositories. The National Science Foundation recognizes the value in data sharing and a component of all grant proposals must include a plan for disseminating the data generated by the project (25).

The Amsterdam Manifesto on Data Citation Principles (26) articulates many of the themes presented here, offering strong guidance for future developments. A few examples of chemistry articles using these deposition principles have appeared (27, 28). Most of the technological needs for a broad system of Open Data sharing are in place today. What is needed is a groundswell of grass roots support for creating and funding these repositories and applying pressure on authors and editors and publishers to mandate such a system.

Open Chemical Information Sources

The culture within the chemical community used to be significantly different from the biology realm. While many of the key biology data sets started off as free, chemists have been conditioned to pay for data collections since the late 1800 and early 1900 with resources like *Beilstein* and *Chemical Abstracts*. With the explosion of the Internet and computerized tools, however, we have seen a drop in the barrier to entry for those trying to deliver a broad range of chemical information data types. Experimentation with delivery of such data at no fee has exploded, resulting in some highly regarded and useful data collections and compound repositories like *ChemSpider*, *ZINC* and *PubChem*.

We believe all of these experiments are extremely useful, and some of these projects have become mainstays in the chemistry community. But there remain significant challenges. The first example is quality. At a time when a kid in the basement can generate a public presence on the web that can rival a well-funded corporate entity or credible educational institution, evaluating what is good and what isn't has become a daily challenge for the individual chemist. To be clear, this is not a new problem. One of the most critical roles libraries and librarians have always played in our discipline is to wade through the abundance of scientific information and literature, and curate and edit down to a core of high-quality resources. But what has changed is the scale and reach. Keeping up with what is credible and what is not can be a daunting task, and guarantees job security for trained information professional into the future. Libraries and information centers need to continue to embrace new resources and provide the same expert filtering they used to deliver for hard-copy materials.

The second issue is resource-intensiveness. Many of the experiments in open databases have been run on a shoestring budget. Early experiments revolved around small molecule databases, but some types of information are objectively more difficult to handle than others, scientifically as well as technically. We believe that this is why reaction databases, while of significant interest to chemists, remain under-represented in the open database realm.

Hand-in-hand with resource-intensiveness is upkeep. Curation is tedious and time-consuming and requires expertise and money. At some point the fun of the hobby turns into the grind of work and it is at this point where these new databases are at risk of disappearing or being taken over by commercial fee-based entities. While in theory the entire community could pitch together to keep these new "free" resources going, the reality is that only a tiny minority of chemists actually participate to supporting these efforts. There are some new paths to longevity, but again these involve finding longer-term resourcing. For example, the commercial databases *StARlite* and *DrugStore* struggled in the marketplace but were scientifically interesting. Each would have represented a real loss to the community had they disappeared. They are now run by the non-commercial entity European Bioinformatics Institute as *ChEMBL*. There are various government funded efforts, such as *PubChem* and the *UK National Chemical Database Service*, that support critical data collections and will continue as long as appropriations are made. Since a common sentiment right now is that publicly-funded research be made accessible to the public, one might expect that

such efforts will continue to grow. On the other hand, the equally strong desire to cut government expenditures creates uncertainty in just how far such efforts can go.

The sweeping “culture of free” has been breaking traditional business models left and right. Commercial entities have so far been largely unable to develop new approaches that added sufficient new value-add to justify the previous high rates they used to be able to charge. So we see continuing consolidation in the area, with small database vendors going out of business, exiting the market, or being bought up by the large information vendors.

We believe that the Internet has allowed many scientists to more frequently opt for lower quality in the name of speed and cost. Sometimes one just needs an answer, not necessarily the absolutely best answer. In such cases, ease of access and speed trump depth and even quality; many of the free databases deliver to this level. This is why even chemists with domain-specific resources at hand still seem to first check Google, and who can blame them. Yet, there still is a compelling role for commercial vendors. Just as in the OA publishing debate, there is an underlying reality that the creation of information incurs tangible costs. Some of the traditional costs have dropped dramatically, and if all a vendor adds to the equation is no longer that resource intensive or no longer that valued by the scientist, then it is no wonder that traditional players are having a tough time. To survive, existing or new commercial players must figure out how to add the kind of new value that will compel a user to opt for a pay service over a free model: be it superior discoverability, significantly improved speed in delivery, enhanced utility to manipulate the underlying data, or new juxtaposition of information that provides rewards that go beyond those offered by access to the isolated information sets.

Mobile Chemistry and the Cloud

With the explosion of the smartphone and tablet, every industry, business and group is coming to the conclusion that “we need an app too.” The chemical community is no exception. Experimentation began with the very first phones, and new players are entering the fray constantly, either with mobile-friendly websites or stand-alone apps. The number of chemistry apps is starting to get unwieldy, leading many librarians to prepare guides, spurring the establishment of the *SciMobile Apps wiki*, and commercial ventures, such as *Nanostuffs Technologies*, to offer various subject-based compilations and suggestions. Pharma industry group *Pistoia Alliance* has launched a life-sciences specific *Pistoia App Catalog* to help scientists find apps that are particularly relevant to the pharmaceutical R&D community.

Mobile chemistry applications today offer read-access to the literature, news and abstracting/indexing services (*Nature.com*, *C&EN Mobile*, *RSC Mobile*, *SPRESImobile*), molecule drawing (*ChemDraw for iPad*), calculators (*LC Calculator*) and health and safety information (*HazMatPocket Guide*, *Chemical Compatibility Database*). The economics of this explosion are not yet clear. Free apps seem to garner more attention than priced apps, and it is difficult to imagine

any chemistry-centric company surviving on app fees alone. In many cases what is being offered is mobile access to already paid-for online services, and in these instances, the additional access point helps solidify the business relationship by making the subscribed-to service potentially more valuable.

Perhaps the most interesting development in chemistry is mobile access to carry out actual chemistry research. This takes advantage of the fact that actual research data and research activities are starting to move into the cloud, with various Software-as-a-Service (SaaS) offerings making data and research activities readily accessible at remote locations. No longer will the researcher necessarily be tied to the lab in order to monitor an experiment's progress, trigger a new run, or write up their experimental plans and conclusions. In a demonstration of what is now feasible, British Telecom Global Services President Bas Burger kicked off a computational experiment and pulled back results, all facilitated on the Apple *iPhone*, during a presentation at the 2012 BioIT conference (29). To the question "will chemists really want to do chemistry on the go?" the answer is becoming more and more clearly "yes". The trick in the mobile development community will be to find useful workflows that expand a chemist's ability to do the job from anywhere.

Social Networks and Social Media

Social networks and social media made their first breakthroughs with individuals' private lives; they too are advancing into chemists' professional lives. New experiments in networking, collaboration, peer review, annotation, crowd-sourced curation and annotation, blogging, and altmetrics are appearing regularly.

The Internet has been a tremendous boon to networking around the globe. One can now easily find, contact, and communicate with people across continents. It is not unheard of to find scientific collaborations involving individuals who have never actually met each other in person.

Collaborations are often born from some type of social network. Nonetheless, one faces a quandary: how many networks does one need to belong to? How many of the existing networks are actually successful platforms for exchange? General sites like *LinkedIn* address a broad range of domains while platforms like *ACS Network* focus just on the chemistry community. *ResearchGate* is trying to bring scientists across disciplines together. *Mendeley* started as reference depository but grew into a scholarly community. It is unclear whether chemists want to join a broad science community or affiliate with a more domain- or subdomain-specific online community.

Social media offers a host of new opportunities to gather information from and share information with the community at large. Today, many conference participants will be live-tweeting the events, and occasionally even broadcasting key presentations, which helps involve those who were unable to travel to the physical location. Some freely available databases are looking to crowdsourcing for deposition and curation of data. The most well-known is *ChemSpider*, but despite their stature they have found limited participation (30). Social

media platforms allow readers to add their commentary to published articles, opening experimentation on how peer review might actually be reconceived or at least expanded to include post-publication critique, as is planned at the new *ScienceOpen* open access publishing platform. These are all very exciting and worthy endeavors, and one looks forward to many more such experiments. However, one needs to be cognizant of some key challenges that face the long-term dependence on broad altruistic community input. As noted in the open chemical information sources section, getting people engaged remains a challenge, and even among long-time dedicated individuals it can be difficult to continue curating and commenting in the face of practical everyday demands of earning a living and advancing a professional career.

Not wanting to be left out, chemical corporations too have been trying to figure out how to engage their customer-bases more actively through social media. Companies feel compelled to maintain a presence on *LinkedIn*, *Facebook*, *Twitter*, and *YouTube*. Many companies have started their own online communities and blogs, such as *Perkin-Elmer*, *ChemAxon*, and *3E*. These vehicles can challenge the corporate marketers to more elegantly hone their messages, going beyond blunt straightforward advertising, and affording them the opportunity to cast their company in an advisory, expert role.

In addition, social media offers a new way to judge scholarly work. Tools for evaluating the value and impact of a single article, a journal, or a scientist are woefully lacking. A well-known tool that has been used in this way is the *impact factor*, though it was not designed with this particular use in mind (31). The *impact factor* rests on citations as the means for judging the impact of a journal. The notion is that numbers of citations reflect the scope of the use or “impact” of the average article. Of course citation data today is also typically behind a pay wall, restricting research on the citation-impact model. David Shotton, Director of the Open Citation Corpus, advocates for an open repository of scholarly citations, to allow for more such research (32).

Since we can now discuss the value of scholarly work in different media, perhaps we can track the mentions of a work on *Twitter* and in blog posts, etc. as a means for judging “impact”. This is the idea behind some new ventures that use so-called alternative metrics to assess the value of scholarly work. Ventures like *AltMetrics*, *PlumX* and *Impact Story* use, among other items, social media mentions to create an assessment of the impact of scholarly activity. With the scarcity of chemical bloggers and tweeters, these alternative metrics have only minimal value in chemistry right now. However, as journal annotation facilities get built out and as the next generation of chemists emerge as young professors, it will be interesting to see if alternative metrics provide useful data for assessing scholarly value.

Lab Notebooks

Unlike publishing, the chemist’s notebook has truly been transformed in dramatic ways with the introduction of electronic laboratory notebooks. While the early products in this arena attempted to duplicate the paper workflow and

behavior, the latest slew of offerings offer a broad range of new functionality and utility that far exceeds what the paper notebook was ever able to achieve. These systems can also be tightly integrated with instrumentation, data repositories and knowledge management systems, creating a whole new information infrastructure for the scientific endeavor, often referred to as the “paperless lab”.

What can e-notebooks offer (33)? Perhaps the most distinguishing unique benefit is vastly improved discoverability. How many hours have researchers pored over old notebooks, or found them outright missing. Many corporate R&D organizations have brought together years of results from a multitude of merged companies, making it virtually impossible to truly assess the scope of knowledge held within the piles of physical notebooks. This challenge is removed when the notebook becomes fully searchable. E-notebooks also offer an improved accuracy of data entry. This can stem from sheer improved legibility to removal of transcription errors by direct feeding of experimental results from the instrumentation software. In addition, collaboration is enhanced because all interested parties can review real-time results. All of these functional benefits lead to clear-cut time and consequently cost savings, and they improve the integrity and quality of the recording of science.

While many of the technical barriers have been broken, the social and economic aspects of lab e-notebooks can still be a challenge. It will be of no surprise that many of the adopters are large multinational corporations, who have carried out analysis of return on investment, and are keenly aware of the cost of loss of intellectual property, lack of discoverability, and inefficient laboratory operations. For them, the economics are compelling and easily justified; but individual resistance can still be found, even in organizations that have adopted e-notebooks. Software tools cannot single-handedly solve organizational and cultural challenges.

Slow adoption can also be seen in academia. Many of the commercial vendors have focused on the high-value, multifunctional solutions that are too heavy and expensive for academics. A second barrier in academia revolves around a hesitation in sharing of results, even amongst members of one’s own group. The enterprise of academia, however, has been changing and universities are becoming much more sensitive to questions of intellectual property protection, technology transfer, and patent licensing (34). It seems reasonable to expect that universities will soon start making the same type of calculations that the corporations have, and will at least consider purchasing electronic lab notebooks at the university level rather than the current typical lab level, shifting the economic imperatives and opening up new markets for the commercial vendors.

A revolutionary experiment is the Open Notebook Science movement (35). Advocates of Open Notebook Science place ongoing research results into the world in real time, arguing that sharing data will lead to better science. As with the Open Data efforts, this area is also impacted heavily by cultural and economic barriers, from current methods of research funding and patent protection, to ways individual faculty are recognized for the contributions to the scientific endeavor. Tremendous sociological changes will be needed for Open Notebook Science to be widely adopted by chemists.

Regardless of whether the science is closed or open, it does seem imperative that chemistry teaching programs start requiring some use of e-notebooks for students. Today's students, whose lives revolve around technology, must find the still standard use of paper notebooks confounding and arcane. This will not be the environment they will find when they move into the workforce.

Collaboration

In many of the previous sections we have touched on collaboration, but this underpinning of science is of tremendous importance and value and warrants some direct commentary on its future as well. Collaboration has been a mainstay of science for centuries. What is new today is the extent to which technology can foster and drive collaboration. The advent of the social media, cloud computing and hosting, SaaS offerings, online conferencing and meeting services, electronic laboratory notebooks, and the paperless lab have generated a critical mass of tools needed for reconceiving the framework of collaboration. These tools allow for real time data sharing and conversation.

Seismic changes in industries like pharmaceuticals have altered how competitive, commercial entities approach the R&D phases of their business. Work is distributed within companies across broad geographies, work is outsourced to lower-cost regions of the world, and risk is mitigated by partnering with other companies (both small and large) who pursue specific lines of investigation. In these instances collaboration can turn on a dime, and the new technologies allow for easy turning on and off of the spigot of R&D information flow. Even in the not-for-profit charity endeavors, we see more groups tackling large global problems like malaria (36) and tuberculosis (37) in a highly distributed and parsed manner, with scientists from around the world able to contribute even a single data point to advance the cause. The OpenPHACTS consortium seeks to break down barriers of entry into pharmaceutical development (38). It is in this humanitarian scientific pursuit, where intellectual property concerns are removed and the altruistic goal mitigates the personal ambition, that we may find the true flourishing of Open Science.

Conclusions

The Internet has undeniably altered the practice of chemists. Chemical information is no longer found in physical libraries housing shelf after shelf of old musty journals and handbooks and monographs. Rather, the chemist largely never needs to leave his or her office or lab to access chemical information. The desktop computer, and now even the smart phone, provides instant access to virtually the entire corpus of chemical journals, along with important databases and even raw data. Collaborations with colleagues around the world are facilitated by email, video chat utilities, and electronic notebooks.

Nonetheless, these changes are more evolutionary than revolutionary. The information technologies available today are ripe for exploitation that truly could bring revolutionary change: large data stores available for immediate re-use, novel means for assessing quality and value, truly collaborative toolsets that facilitate lossless exchange, etc. The social communities that are the engines of such change are lagging behind. We strongly encourage those experimenting with the new communication and collaboration possibilities to continue their sometimes lonely endeavors, because we believe that we are getting close to the critical mass required to create an avalanche of changes to better the overall practice of chemistry.

References

1. Bachrach, S. M. In *Proceedings of the 1994 International Chemical Information Conference*; Collier, H., Ed.; Infonortics: Calne, England, 1994, pp 25–34.
2. *The Internet: A Guide for Chemists*; Bachrach, S. M., Ed.; American Chemical Society: Washington, DC, 1996.
3. Panitch, J. M.; Michalak, S. *The Serials Crisis: A White Paper for the UNC-Chapel Hill Scholarly Communications Convocation*, 2005. <http://www.unc.edu/scholcomdig/whitepapers/panitch-michalak.html> (accessed on April 1st, 2014).
4. Bachrach, S. M. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 264–268.
5. Budapest Open Access Initiative, 2002. <http://www.budapestopenaccessinitiative.org/> (accessed on April 1st, 2014)
6. Bethesda Statement on Open Access Publishing, 2003. <http://legacy.earlham.edu/~peters/fos/bethesda.htm> (accessed on April 1st, 2014).
7. Harnad, S.; Brody, T.; Vallières, F.; Carr, L.; Hitchcock, S.; Gingras, Y.; Oppenheim, C.; Stamerjohanns, H.; Hilf, E. R. *Serials Rev.* **2004**, *30*, 310–314.
8. Creative Commons, 2013. <http://creativecommons.org/> (accessed on April 1st, 2014).
9. Beall, J. Beal's List: Potential, possible, or probable predatory scholarly open-access publishers, 2013. <http://scholarlyoa.com/publishers/> (accessed on April 1st, 2014).
10. NIH Public Access Policy Details, 2008. <http://publicaccess.nih.gov/policy.htm> (accessed on April 1st, 2014).
11. Holdren, J. P. Increasing Access to the Results of Federally Funded Scientific Research, 2013. http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf (accessed on April 1st, 2014).
12. Shared Access Research Ecosystem (SHARE), 2013. <http://www.arl.org/storage/documents/publications/share-proposal-07june13.pdf>
13. CHORUS: Clearinghouse for the Open Research of the United States, 2014. <http://chorusaccess.org/> (accessed on April 1st, 2014).

14. RCUK Policy on Open Access and Supporting Guidance, 2013. <http://www.rcuk.ac.uk/documents/documents/RCUKOpenAccessPolicy.pdf> (accessed on April 1st, 2014).
15. Bachrach, S. M. *J. Cheminf.* **2009**, *1*, 2.
16. Casher, O.; Chandarmohan, G. K.; Hargreaves, M. J.; Leach, C.; Murray-Rust, P.; Rzepa, H. S.; Sayle, R.; Whitaker, B. J. *J. Chem. Soc., Perkin Trans. 2* **1995**, 7–11.
17. Bachrach, S. M.; Murray-Rust, P.; Rzepa, H. S.; Whitaker, B. J. *Network Science* **1996**, *2*, <http://www.netsci.org/Science/Special/feature07.html> (accessed on April 1st, 2014).
18. Murray-Rust, P.; Rzepa, H. S.; Whitaker, B. J. *Chem. Soc. Rev.* **1997**, 1–10.
19. Murray-Rust, P. M.; Rzepa, H. S. *J. Cheminf.* **2012**, *4*, 14.
20. ORCID, 2013. <http://www.orcid.org> (accessed on April 1st, 2014).
21. InChI Trust, 2013. <http://http://www.inchi-trust.org/> (accessed on April 1st, 2014).
22. Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. *J. Cheminform.* **2013**, *5*, 7.
23. Bachrach, S. M. *J. Cheminform.* **2012**, *4*, 34.
24. A statement of commitment by STM publishers to a roadmap to enable text and data mining (TDM) for non-commercial scientific research in the European Union, 2013. http://www.stm-assoc.org/2013_11_11_Text_and_Data_Mining_Declaration.pdf (accessed on April 1st, 2014).
25. NSF. Dissemination and Sharing of Research Results, 2010. <http://www.nsf.gov/bfa/dias/policy/dmp.jsp> (accessed on April 1st, 2014).
26. The Amsterdam Manifesto on Data Citation Principles, 2013. <http://www.force11.org/AmsterdamManifesto> (accessed on April 1st, 2014).
27. Cowley, M. J.; Huch, V.; Rzepa, H. S.; Scheschkewitz, D. *Nat. Chem.* **2013**, *5*, 876–879.
28. Rzepa, H. S. *J. Cheminf.* **2013**, *5*, 6.
29. Davies, K.; Proffitt, A. Hello Siri, Please Start My Experiment Now, 2012. <http://www.bio-itworld.com/news/04/25/12/Hello-Siri-please-start-my-experiment-now.html>.
30. Williams, A. J. Presentation at NFAIS 2012 on Five Years of Experience of Crowdsourcing Chemistry for the Community, 2012. <http://www.chemconnector.com/2012/02/28/presentation-at-nfa-is-2012-on-five-years-of-experience-of-crowdsourcing-chemistry-for-the-community/> (accessed on April 1st, 2014).
31. Garfield, E. *JAMA, J. Am. Med. Assoc.* **2006**, *295*, 90–93.
32. Shotton, D. *Nature* **2013**, *502*, 295–297.
33. Atrium Research, 2013. <http://www.atriumresearch.com/html/infocenter.htm#articles> (accessed on April 1st, 2014).
34. McSherry, C. *Who Owns Academic Work? Battling for Control of Intellectual Property*; Harvard University Press: Cambridge, MA, 2001.
35. Bradley, J.-C. *Nature Precedings*; 2007. <http://dx.doi.org/10.1038/npre.2007.39.1>

36. Open Source Malaria Project, 2013. <http://opensourcemalaria.org/> (accessed on April 1st, 2014).
37. Stop TB Partnership, 2013. <http://www.stoptb.org/> (accessed on April 1st, 2014).
38. OpenPHACTS, 2013. <http://www.openphacts.org/> (accessed on April 1st, 2014).

Subject Index

A

Advent of online services, 26

B

Biomedical ontologies, 221

C

21st Century chemical information
stewardship

information eras and continuum of
transition, 6
content documentation, 8
digital information era, 8
frame the transition, 7
human-readable information, 7
internet, 8
machine documentation, 7

introduction, 1

professionalization of chemical
information, 3

science and poetry of documentation, 9

“black-box” information systems, 11

cheminformatics techniques, 10

data compilation, 11

measurement-method agnostic, 10

methods-based science, 11

organizational processing work, 11

tensions or opportunities, 12

methodology of discipline, 13

power of modularity, 13

rule-driven computational systems, 13

Chemical abstracts service (CAS), 149

addressing information needs of
scientists, 151

CASREACT, 152

CHEMCATS, 152

Toxic Substances Control Act
(TSCA), 152

bioactivity and target indicators, 156

conclusion, 157

content innovation and technology, 155f

experimental procedures and reaction
transformations, 156

future of chemistry research, 154

overview, 150

relevancy ranking, 157

trusting for current and comprehensive
information, 153

Chemical information, from print to
internet

advent of online services, 26

business of information services and
searching, 27

company document, 28

pre-search interviews, 28

research information specialists, 28

searching practices, 28

chemical compound searching, 31

Beilstein System Number process, 32

CA abstracts, 33

CA file online, 33

CHEMLINE/TOXLINE training, 32

MEDLINE, 31

online transition period, 32

chemical reaction information, 34

classical searching, 20

current awareness, 21

information resources, 21

summer job, 21

collegial interaction with vendors, 30

computerization, 22

current awareness, 25

education and training, 37

education business, 38

eye to the future, 39

further collegial interactions, 36

information resources complementary to
chemistry, 35

introduction, 19

need for subject expertise, 23

information services, 24

Naperville labs, 24

research campus, 24

research groups, 24

online searching, 30

patents, 29

physicochemical data, 30

polymer information, 36

quality control, 39

SciFinder, 38

Chemical information transfer, 2

Cheminformatics, 237

Chemistry ontologies

applications

open PHACTS and other datasets of

pharmacological interest, 230

open PHACTS project, properties
 calculated, 231*t*
 text mining, 229
 definition and structure, 220
 example class definition in OBO format,
 224*t*
 example of OWL Manchester syntax,
 225*t*
 examples
 formal-logical representation of
 ascorbic acid, 227*f*
 processes, 228
 small molecules, 226
 introduction, 219
 line notations in cheminformatics, 221*t*
 OWL serialized as XML, 223*t*
 propositional structure, 222
 representing ontologies, 223
 summary and outlook, 231
 Computer-aided synthesis design, 101
 empirical approach, 102
 formal approach, 102
 numerical approach, 102
 Computer-based chemical information
 early online systems, 48
 early uses, 46
 introduction, 43
 inventions, 52
 CAS Registry Nomenclature File
 (RNF), 53
 CAS Registry Number, 53
 Chemical Substance Index, 53
 heterocycle terms, 54
 search term, 54
 large chemical information databases, 44
 Lockheed information systems, 51
 the move to online, 47
 patent information, 45
 summary, 55
 system development corporation, 50
 the transition, 49

D

Demonstrating general principles and
 pedagogical techniques
 ask Star Trek computer, 191
 chemists read, write and believe, 189
 information retrieval, identifying
 substances, 193
 Monty Python and search for information
 aim, 190
 quest, 190
 tools that exist to help, 191

use of library equation to transform, 192

E

Electronic information industry, 1
 Essential information skills all chemists
 should learn, 184
 deconstructing and reconstructing
 textual query
 brainstorm list of terms, 186
 break query into constituent concepts,
 185
 employ list of terms to generate sound
 query, 186
 state information need, 185
 deconstructing molecules to build
 substructures
 ask questions about atoms, 187
 ask questions about bonds, 188
 construct appropriate query using
 tools, 188
 draw core of molecule, 187

G

General principles of teaching chemical
 information, 174
 principle 1, basics of information
 retrieval, 176
 principle 2, information systems, 177
 principle 3, use source effectively, 177
 principle 4, all information sources not
 created equal, 178
 principle 5, to choose source effectively,
 179
 technology, 175

H

History of chemical information, 57
 chemical literature, 58
 categories, 59*f*
 landmarks, 58*t*
 high-quality information, 77
 outlook, 75
 present-day students, 76
 from print to online
 CAS Online, 63
 chemical databases, 63
 compound searches in CAS Registry,
 63

- early developments, 62
 - landmarks in history of electronic
 - chemical information, 64*f*
 - later enhancements, 64
 - problematic legacies, 76
 - protocol (part) of compound search, 61*f*
 - searching online, 65
 - access to Chemical Information at ETH Zurich around 1998, 67*f*
 - linking technologies, 66
 - search database, specialists, 66
 - searching printed chemical literature, 60
 - state of the art, 71
 - preparation and other reactions, 73
 - reaction substructure search, 74*f*
 - search terms, 72
 - sequence or subsequence searches, 72
 - TIS (tabular inorganic substances), 72
 - support, training, and education, 67
 - experiences at ETH Zurich, 69
 - searching for substructures with
 - properties, 68
 - traditional secondary sources, 76
 - History of chemical reactions information
 - the future, 104
 - introduction, 95
 - the past
 - abstracting services, handbooks and journals, 97
 - Alchemy, 96
 - beginnings of modern chemistry, 96
 - the present
 - computer-aided management of reactions, 103
 - computer-aided synthesis design, 101
 - electronic publications, 100
 - reaction databases, 101
 - structure representation, 99
 - social networks and social media, 262
 - Institute for scientific information, 109
 - the concepts, 111
 - Garfield, 110, 114
 - the information environment, 112
 - ISI – the conceptual years (1954 – 1960), 113
 - ISI – the digital years and the end of an era (1980 – 1992), 119
 - Atlas, 120
 - career at ISI, 122
 - CD-ROM products, 121
 - Chembase* and *ChemSmart*, 119
 - e-versions of core products, 120
 - ISI, crazy place to work, 122
 - Theodore Lamont Cross, 121
 - ISI – the early years 1960 - 1970, 115
 - Current Contents*, 116
 - Genetics Citation Index (GCI)*, 116
 - Index Chemicus*, 116
 - ISI – the middle years 1970 – 1980, 117
 - chemical information products, expansion, 118
 - citation index product line, 118
 - Current Contents/Chemical Sciences*, 118
 - ISI building, 123
 - ISI's chemical information services, 117
 - medical librarian, 114
- I**
- Information management for practicing chemists, 255
- collaboration, 265
- conclusions, 265
- lab notebooks, 263
- mobile chemistry and the cloud, 261
- open chemical information sources, 260
- scientific publication
 - data publication and open data, 258
 - models, 257
 - open access, 256
 - STM publication, 257
- L**
- Lockheed information systems, 51
- M**
- Mobile workflows and data sources
 - challenges, 239
 - chemical structures, 241
 - creating content, 243
 - exporting content, 246
 - importing content, 245
 - MolPrime+ property calculation, 248*f*
 - representing structures, 249
 - simplified casual drawing interface, advanced gesture-based primitives, 244*f*
 - structure-based calculations, 247
 - conclusion, 250
 - early history, 240
 - introduction, 237

O

Open chemical information sources, 260

P

Patent citation searching, 91
 citation searching, 92
 conclusions, 93
 sharing search results, 92
Patents, 29
Patents and patent citation searching
 first U.S. chemical patent, 82*f*
 Patent Cooperation Treaty, 86
 patents and journal articles, differences, 83
 patents as chemical literature, 81
 retrieving chemical information from patents
 chemical structure searching, 88
 classification systems, 86
 Derwent fragmentation coding, 89
 INPADOC database, 89
 methods for extracting searchable chemical structures, 90
 subscriptions, 89
 21st century patent specification, 84
Public chemical databases, overview, 200
Public chemical databases and semantic web
 ChEBI and ChEMBL, 206
 chemical data, brief history, 198
 ChemSpider, 204
 compound record in PubChem, 204*f*
 conclusion, 212
 drug-related databases, 209
 environmental and safety databases, 209
 international chemical identifier (InChI)
 as open standard, 202
 introduction, 237
 NIST, 205
 openness aspect, key areas, 211
 OpenPHACTS (Open Pharmacological Concept Triple Store), 211
 other sites of interest, 210
 part of substance record in ChemSpider, 205*f*
 part of Wikipedia ChemBox, 208*f*
 PubChem and PubMed, 203
 reason to be public, 199
 sample ontology in ChEBI, 206*f*
 selected public chemical databases, 200*t*

towards semantic web, 210
using public data, 202
Wikipedia, 206

R

Reaxys
 building on strengths in data linking and access, 134
 building on strengths in structure and reaction data, 135
 evolution of user interface, 140*f*
 expanding coverage scope, 137
 interactive Excerpton Interface (iEI), 138*f*
 synthesis generated in Synthesis Planner, 136*f*
 user-centered, rather than technology-centered development, 139

S

Spectra and searching from punch cards to digital data
 blank McBee card with holes punched, 163*f*
 double-beam instrument, 162
 the early years, 160
 FT-IR spectrum, example, 166*f*
 the future, 167
 IBM cards, 162
 infrared grating spectrum, 165*f*
 instrument, 161
 introduction, 159
 McBee cards, 162
 new instrumentation, 161
 optical instruments, 161
 the present, 99
 punch-card systems, 162
 Sadler collection of spectra, 164*f*
 spectral collections, 166
 using infrared, 161
Stewardship & long view of chemical information
 chemical research, 14
 chemists' documentation, 14
 documenting provenance, 14
 provenance tracking, 14
System development corporation, 50

T

- Teaching chemical information for the future. *See* Demonstrating general principles and pedagogical techniques
- chemical information instruction, past and present
- course integrated instruction, 173
 - dedicated courses, 172
 - need to read, 169
 - surveys in colleges and universities, 170
 - teaching chemical information, methods and models, 171
- conclusion, 194
- Teaching for retention and lifetime learning, pedagogical techniques
- appropriate reuse of material, students confused, 183
 - avoid "ooh! shiny!" syndrome, 183
 - keep small, keep active, and minimize redundancy, 181
 - teach relevant, transferrable skills, and use resources to demonstrate skills, 180

U

- Unobstructed access to relevant chemistry information, 127
- CrossFire revolution
- brings information to the chemist, 132
 - introduces data export and linking, 133

- introduction, 128
- Reaxys, 134
- steps toward unobstructed information access, 141
- envisioned fluid user experience, 143
- indexing and taxonomy beyond equivalence and hierarchical relationships, 142
- interoperability, 145
- search in Reaxys Medicinal Chemistry, 146f
- structurally translatable nomenclature terms, 143
- user's environment, integral component, 144
- visionary organizational heritage, 129

V

- Visionary organizational heritage
- Gmelin and Beilstein handbooks, 129
 - digitization, 131
 - structure-based organization, power, 130

W

- Whither future of chemical information, 4
- challenges, 6
 - implicit vs. explicit, 5
 - variability, 5