Herman Skolnik Award Symposium 2014

Honoring Engelbert Zass

A report by Wendy Warr (wendy@warr.com) for the ACS CINF *Chemical Information Bulletin*

Introduction

Throughout his career Dr. Engelbert Zass ("Bert"), head of the Chemistry Biology Information Centre at ETH Zürich (retired), has been a bridge builder and mediator between database producers, vendors, publishers, librarians, and end users in chemistry, contributing to advancing chemical information as a whole. Specializing in chemical information after receiving his Ph.D. in organic chemistry, Dr. Zass has more than 30 years of experience in searching, operating and designing chemistry databases, as well as in the support, training, and education of users of chemical information. He has given numerous lectures and courses in Europe and the United States, is author of more than 60 papers on chemical information, and served on several publisher advisory boards. From 1999 till 2004, he was a partner in the German Federal Ministry of Education and Research's project "Vernetztes Studium – Chemie", where he was engaged in the design of multimedia educational material for chemical information. Through his leadership, vision, and collaborative efforts with his staff, ETH Zürich developed a model 21st century library. Dr. Zass did his undergraduate studies in chemistry at Universität zu Köln, followed by a Master's degree (Diplom) in Chemistry with Prof. E. Vogel. He went on to complete his Ph.D. (Dr. sc. nat.) studies with Prof. A. Eschenmoser at ETH Zürich. He then became a lecturer and senior scientist at ETH, later serving as Head of the expanded ETH Chemistry Biology Pharmacy Information Center until his retirement in 2012.

Evolution and transformation of journals in a digital environment

Grace Baysinger of Stanford University opened the proceedings with a talk about electronic journals. The number of electronic journals continues to increase: CrossRef (http://www.crossref.org/01company/crossref_indicators.html) covers over 35,000 and the Directory of Open Access Journals (http://doaj.org/) lists about 9,700. Chemists make extensive use of journal articles. Most publishers now offer them electronic manuscript submission systems and authoring tools. The Royal Society of Chemistry (http://www.rsc.org/Publishing/Journals/guidelines/AuthorGuidelines/AuthoringTools/), for example, offers author templates, an experimental data checker, and a Crystallographic Information File data importer.

We need more automated data checking tools, not just to aid authors, but also to help prevent fraud. It is important that readers should be able to reproduce the work reported in an article; reproducibility depends partly on the availability of supporting information. Open source software and lower computing costs make it easier nowadays to re-use data. NISO and NFAIS have published *Recommended Practices for Online Supplemental Journal Article Materials* (http://www.niso.org/workrooms/supplemental). CrossCheck (http://www.crossref.org/crosscheck/index.html) prevents plagiarism; CrossMark

(http://www.crossref.org/crossmark/) provides a standard way for readers to locate the authoritative version of a piece of content; FundRef (http://www.crossref.org/fundref/) provides a standard way to report funding sources, and ORCID (http://orcid.org/) provides a persistent digital identifier that distinguishes an author from every other researcher.

Many people think that traditional peer review is "broken" (http://www.nature.com/nature/peerreview/debate/) because it causes delays, and reviewers are overloaded. Alternatives are pre-publication review as carried out by arXiv.org (http://arxiv.org); post-publication review as in Faculty of 1000 (http://f1000.com/); and open two-stage peer review as used by *Atmospheric Chemistry & Physics* (http://www.atmospheric-chemistry-and-physics.net/review/review_process_and_interactive_public_discussion.html).

Copyright may be transferred to publishers, or held by the author, who grants a distribution license to a publisher, or held by an institution or employer. There are also creative commons licenses, and some works are in the public domain. The Copyright Clearance Center operates RightsLink (http://www.rightslink.com) to enable publishers to give permission for an item to be reproduced. SHERPA/RoMEO (http://www.sherpa.ac.uk/romeo/) summarizes permissions that are normally given as part of a publisher's copyright transfer agreement.

Collection management is different in the era of the electronic library. Expenditure reports have to be produced; electronic resource management systems are needed for storing licenses; and authentication and security must be handled correctly. Catalog records for bibliographic data and knowledge bases for online holdings must be maintained. COUNTER (http://www.projectcounter.org/about.html) and SFX reports (http://www.exlibrisgroup.com/category/SFXOverview) measure electronic usage. Print collections get sent to storage and archiving of print may be shared. Online access may be perpetual or there may be archival access to an online version. Repositories for data have been established. Pricing is a major issue. Chemistry journals reportedly have the highest average cost (http://lj.libraryjournal.com/2014/04/publishing/steps-down-the-evolutionary-road-periodicals-price-survey-2014/#) of all subject areas: $4,215. A recent article has drawn attention to variations in pricing across institutions and a lack of transparency.[1] All sorts of metadata issues may arise. There can be multiple titles in one catalog record for the print version of a journal. Only the latest title may be on the publisher site, or in the open URL knowledge base, but the link and content may include older titles. Digitization and metadata for a journal may be incomplete. Multiple versions or copies of the same article may be on different sites. Problem solving is more difficult now that print copies from libraries are in storage or withdrawn. NISO has published *Recommended Practices for the Presentation and Identification of E-Journals* (http://www.niso.org/workrooms/piej).

Archival material can be accessed through the Wayback machine (http://archive.org/web/), Portico (http://www.portico.org/digital-preservation/), Hathi Trust (http://www.hathitrust.org/help_general), and Lots of Copies Keep Stuff Safe (LOCKSS, http://www.lockss.org/) and Controlled LOCKSS (CLOCKSS, http://www.clockss.org/clockss/Home), but there are still problems with "bit rot" and multimedia content. LOCKSS and CLOCKSS are the only services that check for bit rot. Data are being stored in repositories and databases, and on Amazon cloud.

Mobile access is another big theme. Enhanced journal article services are now being added by publishers: see, for example, the tools supplied by ACS (http://pubs.acs.org/doi/abs/10.1021/cb500271c) and Wiley (http://onlinelibrary.wiley.com/doi/10.1002/ajoc.201402054/full). The University of California Irvine has an online guide to research impacts using metrics (http://libguides.lib.uci.edu/researchimpact-metrics). Article level metrics are increasingly becoming an alternative method of measuring the impact of scholarly and other output: Altmetric (http://www.altmetric.com/) is an example.

There are many resources for discovery and delivery including open URL knowledge bases, Portico (http://www.portico.org/digital-preservation/), databases, federated sites and tools (e.g., xSearch at Stanford, https://xsearch.stanford.edu/search/), alerts and RSS feeds, data mining and visualization, XML parsing of content, linked data and taxonomies, and machine-to-machine retrieval *via* APIs and Web Services (e.g., Stanford Profiles (https://profiles.stanford.edu/), a LinkedIn-type service for Stanford faculty).

End user behavior

Research publications are ultimately intended to be read by scientists. What do they want and need in a publication? How do they gather their information and decide what to read? When and how much do they read? User studies of various kinds have been done to try and find answers to these questions. Andrea Twiss-Brooks of the University of Chicago Library reported on recent studies in which she has been involved.

User studies carried out in around 2005, many conducted by Tenopir and King, showed that users browse for current articles and current awareness, but search databases or follow citations for older articles. They rely on the library copy for all but a few core titles, and authoritative and trusted sources are preferred. They need an efficient means of accessing literature: time management is a priority. Reading supports primary research, background research, teaching, and writing.[2]

User behavior research may be basic or applied, and methods may be quantitative or qualitative. Quantitative methods include surveys (using the Likert Scale, for example, a psychometric scale often used with questionnaires), return on investment metrics, citation metrics, altmetrics, and server log analysis. These methods answer the “how much?” “what?” and “when?” type of question. They are easier to manage and design (sometimes), empirical, and perhaps extensible. Standardized statistical approaches can be used.

Qualitative methods include focus groups, interviews, open-ended comments on surveys, and applied ethnographic techniques which may include research diaries, mapping the diaries with interviews, and observation. Qualitative methods often require more oversight by Institutional Review Boards (IRBs). They produce information only on the cases studied: generalization is more difficult. They give insight into “how?”, “why?” and “who?” These questions can often not be answered by quantitative methods, but analysis of qualitative results can be more challenging and is usually not statistical.

Nancy Fried Foster and Susan Gibbons have used anthropological and ethnographic methods to examine how undergraduate students at the University of Rochester write their research papers. Students (with

informed consent) were watched, and they kept diaries. The results were published in *Studying Students: the Undergraduate Research Project at the University of Rochester* (http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/booksanddigitalresources/digital/Foster-Gibbons_cmpd.pdf).

Early in 2012, David Bietila and Gina Petersen at the University of Chicago conducted a qualitative study of the research process of eight graduate students in the humanities and social sciences. This study aimed to model the research process, enhance the understanding of the relationships between information tools and services, and identify gaps between participant behavior and librarian expectations. The study's data collection comprised three components: research logs, semi-structured interviews, and commentary by subject librarians on student research practices. The findings indicated that source discovery was not a primary concern for participants, but that developing a feasible, clear, and relevant topic was a much greater obstacle for most participants. Bietila and Petersen spoke about this in "Guiding Interface Design with Ethnographic Methods" at the American Library Association Annual Meeting in 2013.

Some examples of research log questions were:

- What were you trying to accomplish during this research session?
- At what time did you conduct this research session?
- Where were you when conducting this research session?
- What tools did you use to help conduct your research?

Examples of interview questions were:

- What do you anticipate will be the most difficult part of the research?
- How do you evaluate sources?
- How have you converted your topic into a searchable phrase or keywords?

Graduate students tended to use the same resources, for example, JSTOR. They did a broad search and read a few articles. They used a non-linear, non-structured process; they did not have a set process for doing things. The librarian would always have done the job differently. A report on the study has been published (http://www.lib.uchicago.edu/gradstudy).

A second project, "A Day in the Life", is ongoing. This mapping project is applying ethnographic methods to clinical health information research. It is a low-cost, six-institution study investigating how third-year medical students seek and use information in the course of their daily activities. The students mark their movements on a map for one full day. Each participant is then interviewed (and rewarded with a $100 gift card). Interviews are audio recorded and transcribed. The transcripts are coded and analyzed in order to identify possible service, facility, resource and other improvements. Nancy Fried Foster, who was involved in the University of Rochester work, is the analyst and consultant. The medical students have a much more focused life than the University of Rochester students did: they went to the clinic and stayed there. The interviewers carried out "interested, neutral listening". They were taught how to code the transcripts in a two-day workshop.

Preliminary results suggest that “putting on the white coat” is significant, and time management matters. When selecting the best information tool, print books are more important than predicted. Andrea has learned some other lessons concerning timeframes, the challenges of multi-institutional studies, the funding and IRB process, and achieving consistency in methods in study design. The importance of the team leader for each institution, and the value of the consultant are significant success factors.

Panel discussion on information literacy

Two speakers were unable to attend at the last minute so an impromptu panel discussion was arranged. While impromptu, this discussion sparked a lively dialogue between panel members and attendees. The panelists were Engelbert Zass, Grace Baysinger, Andrea Twiss-Brooks and Donna Wrublewski (of Caltech).

Bert has been running courses at ETH Zürich and at the Universities of Bern and Innsbruck, teaching chemical information since 1981. Nowadays, he said, the providers do a much better job of training, but it is source-oriented; Bert does problem-oriented training. He reported that registration for SciFinder is not liked; at Innsbruck students found it even difficult. In order to register as SciFinder users, students have to find a link on their University’s website, and from there connect to the specific CAS connection page; for SciFinder, there is no direct registration from the general database start page as in Web of Knowledge, Scopus, or Reaxys. Trainees are better at searching nowadays (because of better user interfaces), but they find it harder to find the full text and original source if the link is broken: they often do not know how to use the library’s online public access catalog (OPAC). Many also have problems in defining a structure query.

Grace offers workshops, gives presentations to classes, and does a lot of one-on-one consultation to users. All of the major database vendors also provide short online tutorials that users can consult. The spectrum of expertise levels varies considerably. While some users are very “tech savvy”, they may have limited experience searching chemical information. Because Google provides a couple of highly relevant citations quickly and easily, multi-tasking users have developed “short attention span theater” and now expect to locate information without any training. Unfortunately, they do not know what they do not know. One colleague reported that some students hired to work in a corporate environment did not have the information skills needed for the job because they had relied too much on Google while in school. One area that Grace has been concentrating on is compiling information about laboratory safety resources as users may not be as familiar with them as they are for materials in other areas of chemistry. With so many materials now being purchased only in a digital format, it is critical for users to learn how to navigate in the OPAC rather than to rely on browsing print copies in the stacks. With student populations in universities now being more diverse, English may be a second language. Hands-on practice is essential in chemistry, not just a demo.

Donna was taught by a member of the audience as an undergraduate student; her academic advisers were not much help. Some people just go to the library to hide. When Donna arrived at the University of Florida she was able to think what would have made her thesis project less painful. She modularizes her information literacy courses. Controlled vocabulary is very important in chemistry. She teaches one

useful strategy for each of Google, Wikipedia, the OPAC, and Web of Science. She uses a topic with which the user is familiar to start formulating a strategy. If you can find it on Amazon you can use an OPAC.

Andrea outlined some challenges faced at the University of Chicago. Two thirds of students are graduate students and one third are undergraduates. There are non-bibliographic information sources and tools used in science and medicine research that librarians may not be familiar with, nor have the training or knowledge to use. Students are being asked to use all sorts of new IT and analysis tools, for example, OpenBLAST, FlyBase, and other bioinformatics tools. The University's Computer Science Instructional Laboratory tutors teach the use of computer programming tools. Librarians collaborate with the tutors to see what resources the students need and how best to refer the students needing expert training. The library tries to keep aware of where the expertise lies at the University. For example, 3D visualization is available at the Research Computing Center. In joint courses, the library provides the space, and the partners provide the expertise. The library also works with the IT group: the IT group can do problem-solving for mobile devices setup, and they will do things such as MatLab training, for example. The library also works with vendors for on-site and Webinar training, providing the room with computers, or the needed audio and video setups.

Questions were invited from the audience. One person said that the four panelists were all from big libraries; what about smaller schools? Bert replied that Bern and Innsbruck are much smaller than ETH Zürich. Even with a limited number of databases you can teach people to use what they have, but this demands creativity. The questioner felt that Europe is different. Grace said that at undergraduate institutions librarians usually spend more time teaching than doing collection development, but the number of resources being taught is smaller. Grace helped revise a document on information literacy skills that undergraduate chemistry students should achieve by the time they graduate (e.g., search by topic, author, and physical properties). This document included resources that could be used to learn that skill. Due to small budgets, at some skills an effort was made to include high-quality, free resources and to put a dollar sign by resources that need to be purchased. Often, instruction that is integrated into a course and tied to an assignment that a student must complete is the most effective type of training at the undergraduate level. Starting her academic career at a junior college, Grace benefited by getting personalized help from her instructors and the librarians. Because so much stuff in chemistry can be related to ordinary life, it provides an opportunity to introduce undergraduates to chemical information tools and to increase their interest in chemistry as a discipline.

Another person asked if it is appropriate to use Google. Yes, said Donna, with some hesitation. Bert recommends using *all* sources. Yes, said Grace, if the risk is low. It should not be used for explosives, for example. The questioner pointed out that Google points you to a source. Could you run a course on Google? Andrea thought you probably could; searching is a general skill that everyone should have. Grace would prefer to teach Google Scholar rather than Google for finding information. When Grace attended the Biennial Conference on Chemical Education (BCCE) in 2012, she heard a presentation by a chemistry professor who is using Wikipedia to help teach students better writing skills. First the students evaluate an article present in Wikipedia and then they have to write an article for Wikipedia. With Google you can get 1 million answers in under a minute, but it can take hours to evaluate the hits if you

are trying to do a more comprehensive search. Donna pointed out that Google has offered courses, but not in chemistry. A librarian in the audience has taught a laboratory class for non-chemistry majors. She used Wikipedia and got the message across about the pros and cons. Faculty and students do not think they need teaching. There are cultural differences too: students need to get something useful that they can use in their own projects.

Donna reported the same problem in teaching ethics and communication and safety. Faculty do not have time to care. Grace said that the Stanford Chemistry Department has laboratory safety coordinators for each laboratory group, who are semi-expert. She has resisted this model for chemical information as *everyone* should learn how to use information, but maybe an information coordinator in the laboratory would be helpful. Bert said that in "the STN days" ETH had this concept in order to save money. When Beilstein CrossFire came along, the information coordinators were known and could be useful contact persons.

Another librarian in the audience said that the answer is in the kind of question you ask on Google. In the print days you asked the librarian if you could not find something in a book. In the electronic era you *do* find something so you may not ask the librarian. Google is superlative at the "good enough" answer, but it fails on an exhaustive search. Yet another librarian said that he introduces Web of Science. Students do not use the right fields. He uses queries such a "XYZ was published in *Tetrahedron*; how often has it been cited".

A German attendee pointed out that courses of this sort are not usually taught in Germany. Andrea said that, in theory, information literacy has to be taught in the United States. The German attendee said that the course could take two hours out of a laboratory class. Grace noted that if it becomes a library class rather than a chemistry class, the faculty member may object. For example, faculty may want the students to analyze peaks in their spectra manually, rather than do a spectral peak search in a database Doing this may be harder and take longer when trying to identify an unknown, but the students understand chemistry better if they have done the analysis manually. Donna said that learning how to use SciFinder should not be made difficult.

Grace added that in the days of print, people used to be able to browse the stacks; they *have* to use the OPAC now. Bert feels that it is important to show people how to find a good book or review. You have to show them an example. Grace said that some people know how to use RSS feeds, but others do not know what tools are available for keeping current.

Chemical publications revisited

Guido F. Herrmann of Georg Thieme Verlag also addressed the topic of chemical publications, but concentrated on the connections between full-text information and information embedded in the chemical structures and reactions. Thieme (http://www.thieme.com/) is a medium-sized publisher that has had an internationally strong position in chemistry (https://www.thieme.de/en/thieme-chemistry/home-51399.htm) since 1886. It produces journals, text books, monographs, reference works, dictionaries, databases, continuous education products and interactive online libraries, in multiple formats. These basic categories have been very robust and stable for more than a hundred years. In

contrast the published formats (digital *versus* print), the user expectations, the production processes, and the distribution channels have seen significant change over the last decade.

According to the November 2012 STM report *An overview of scientific and scholarly journal publishing* (http://www.stm-assoc.org/2012_12_11_STM_Report_2012.pdf) "the number of articles published each year, and the number of journals, have both grown steadily for over two centuries, by about 3% and 3.5% per year respectively. The reason is the equally persistent growth in the number of researchers, which has also grown at about 3% per year". A major change, a veritable revolution, has been the move from print to electronic distribution, and Thieme is still managing an ongoing change process. All Thieme's chemistry publications now have digital versions, including backfiles from 1909 onwards.

The basic role of a publisher has remained, but the actual operations and production processes have changed drastically. The STM Tech Trends 2013 poster (http://www.stm-assoc.org/future-lab-trend-watch-2013/) illustrates many things that are starting to happen: where should a medium-sized publisher focus? Can a publisher help in converting scientific information into applied knowledge? One goal is not just to produce information, but to create value. In May 2008, a Research Information Network report (http://www.rin.ac.uk/system/files/attachments/Activites-costs-flows-report.pdf) estimated that "the global cost each year of undertaking and communicating the results of research reported in journal articles is £175bn, made up of £116bn for the costs of the research itself; £25bn for publication, distribution and access to the articles; and £34bn for reading them." Publishers could add value by reducing the cost of reading, that is, by providing researchers with relevant information more effectively.

Guido gave two examples. The first concerned primary data. Guido estimates that there are 500,000 to 1 million datasets a year in organic chemistry. To preserve them, and make them discoverable and re-usable, requires servers and data centers, metadata, and digital object identifiers (DOIs). FIZ Karlsruhe (http://www.fiz-karlsruhe.de/home.html?&no_cache=1&L=1) houses the Thieme data, and Technische Informationsbibliothek (TIB, the German National Library of Science and Technology (http://www.tib-hannover.de/en/) assigns DOIs to them, stores the metadata and keeps them searchable. TIB is the managing agent of the DataCite organization (http://www.datacite.org/). At the same time as an article is published, the primary data are published as an independent entity: the article quotes the research data as reference items with the assigned DOI.

Authors of articles in the Thieme journals *SYNLETT* and SYNTHESIS are now being invited to submit their datasets for publication alongside their articles. The primary data have their own DOI, different from the one of the paper, and can thus be cited independently. Spectra, for example, are published not as PDFs or JPEGs, but as raw, interactive data, which can be downloaded and analyzed. Benefits are citability and high visibility of research data, easy re-use and verification of the datasets, avoidance of duplication, and motivation for new research. Unfortunately, authors are, thus far, not enthusiastic about supplying their data, and reviewers claim they have no time to check the data.

Guido's second example concerned full text, structures and reactions. *Science of Synthesis* (https://science-of-synthesis.thieme.com/app/home) is the successor to *Houben-Weyl*, the archive of which contains approximately 146,000 experimental procedures, 580,000 structures and 700,000 references, in 160 volumes. *Science of Synthesis* (from year 2000 onwards) contains approximately 50,000 experimental procedures, 270,000 reactions, and 1,250,000 structures in 48 volumes. *Science of Synthesis Updates* (since 2010) contains a further 18,000 experimental procedures and 40,000 reactions in 17 Volumes. *Science of Synthesis Reference Library* (since 2010) contains 15,000 experimental procedures and 40,000 reactions in 13 Volumes. New material is uploaded several times a year.

Science of Synthesis 4.0 has a new production system this year, developed in collaboration with InfoChem (http://www.infochem.de). Previously, all reaction schemes were completely redrawn and indexing was done manually. Now authors' schemes are used (with modification), the schemes are checked and modified by a scientific editor, and the indexing is mostly automated. In the days of manual indexing, each structure was taken from a scheme and loaded into a database; starting materials, products, reagents, solvents, temperature and yield were defined by an indexer; each individual structure and reaction was extracted from tables and scheme tables; and it took about three months in all to index a *Science of Synthesis* volume. Now Thieme, in a continuing collaboration with InfoChem, has developed an automated indexing system: structures and reactions are automatically indexed using InfoChem's SchemeAnalyzer software, which is 85% successful in extracting structures and single-step reactions directly from complex schemes in ChemDraw files. Another new development is the implementation of a MarkLogic NoSQL database (http://www.marklogic.com/what-is-marklogic/) for a new graphical user interface, and full-text and data search. Work is ongoing further to improve the link between the InfoChem system (i.e., chemical information) and the MarkLogic system (i.e., full-text information). The final system will give even better value to the user.

CAS keeps pace with the worldwide growth in disclosed chemistry

Chemical Abstracts Service (CAS) has always been a leader in providing scientists with access to chemical information. Matt Toussant of CAS described how CAS has adapted to the phenomenal growth in chemical information being published today. The ACS is committed to "improving people's lives through the transforming power of chemistry". Its mission is "to advance the broader chemistry enterprise and its practitioners for the benefit of Earth and its people". The mission of CAS is "to provide the world's best digital research environment to search, retrieve, analyze, and link chemical information". Chemistry is the central science.

Matt showed a timeline of some influential events in publishing. Johannes Gutenberg invented the printing press in about 1450. The Internet started with the time-sharing of computers in the early 1960s at U.S. universities and with the Advanced Research Projects Agency Network (ARPANET), developed after the launch of Sputnik in 1957. Ebooks are now an alternative to printed books. Galileo's heliocentric dialogue was published in Latin by Elzevir.[3] Robert Boyle's five-person dialogue *The Sceptical Chymist*[4] was published in 1661. Books were printed in small numbers in those days; nowadays books are widely available.

The history of scientific journals dates from 1665, when the French *Journal des sçavans* and the English *Philosophical Transactions of the Royal Society* first began systematically publishing research results. The number of serials (http://www.stm-assoc.org/2012_12_11_STM_Report_2012.pdf) has grown every year since then. Eventually abstracting and indexing services such as Chemisches Zentralblatt (born in 1830) were needed to help readers keep pace with the literature. *Chemical Abstracts* (*CA*) began in 1907. Later, CAS extended its reach into the patent world. In 1641, Samuel Winslow was granted the first patent in North America for a new process for making salt. The first U.S. patent was granted in 1790. The first patent in *Chemical Abstracts* dates back to 1808 and concerns an alcohol still.

CAS has covered several serials for more than 100 years, for example, *Annalen der Chemie und Pharmacie* (later *Justus Liebigs Annalen der Chemie und Pharmacie*, now part of the *European Journal of Organic Chemistry*) which began in 1840. CAS has covered 50,000 journal titles over the years; it now covers 10,000. It has more than 100 years' experience of analyzing and organizing disclosed chemistry from around the world. The work is no longer done manually in a library at Ohio State University; nowadays computerized data entry and sophisticated tools are used. Chemist labor around the world, and "postal support" to analyze chemical publications, ended in 1994 because it was too slow. At its peak in 1967, this process involved nearly 3,500 people. A newsletter called *The Little CA*, published up to four times a year, kept the "volunteers" informed. E. J. Crane (editor of *CA* from 1915 to 1958) was the main author. The volunteers worked all over the world: Czechoslovakia, Poland and the United States were well-represented but the largest number was that of Japanese chemists. Neutral parties helped during the war years. Much analysis is now outsourced to India, Japan and China.

The history of CAS REGISTRY goes back to a concept of Malcom Dyson's in the 1950s. The original database was a file of fluorine compounds using the Dyson-IUPAC notation on edge-notched cards. In the 1960s, Harry Morgan of CAS, building on the work of Donald Gluck at DuPont, published an algorithm that converted structure diagrams into unique tabular forms.[5] This handled aromatic and tautomer bond representations and established the basis for REGISTRY. The building of REGISTRY began in 1964.

The CAS indexer analyses the whole document, creates a CAS REGISTRY record and interprets when compounds are described in terms other than singular structures or names. A typical chemistry patent (a PCT application for "A new antibacterial", with 250 pages and 24 claims) took 15 days to index completely, with 917 compounds, 576 new compounds, 613 single-step reactions, 5,394 multistep reactions, 1029 reaction participants, and one MARPAT Markush structure with 2,119 substituent definitions. CAS specialists in many fields of chemistry interpret author terminology to register compounds. Spectra, numeric properties, tags, and published sources are recorded.

CAS databases continue to show strong growth. In particular the number of patents has greatly increased recently (9.2% growth in 2012); much of the increase is related to China, Korea and Japan. About half of small molecule registrations are compounds from patents. Sixty-three patent authorities are now covered, and it usually takes less than 27 days from receipt for a patent to appear in *CA*. In non-patent information, Matt drew attention to growth in Asia. Nowadays, 69% of items indexed by CAS are in English, 13% in Chinese. More than 89 million CAS Registry Numbers have been issued; 29 million

registered substances have been indexed in the last 5 years alone. The number of prophetic substances and Markush structures is up more than 10%.

Apart from journals and patents, CAS also covers dissertations, meeting abstracts, and conference proceedings, more than 1,000 ahead of print journals, valuable Web sources, commercial chemical suppliers and regulatory inventories. Attendance at national meetings by larger societies does not really seem to be decreasing. "Ahead of print" appearance of journal articles, rapid publication, and "letters" journals are other trends. Speed of publication is critical. The proportion of ACS articles with supporting information rose from 60% in 2012 to 70% in 2014; supporting information is mostly in the form of PDF files, but there are also Crystallographic Information Files.

Matt ended with some predictions. Numbers of all forms of publications will continue to increase annually by 3% to 5%; Web-based "As Soon As Publishable" will dominate. Supporting information will broaden. Open access will play a significant role. Asia will be the origin of much new science. Disclosures originating from commercial sources will increase as new substances are being created in laboratories around the world. The pace of growth in patent applications will slow, but patents will remain a main vehicle for the monetization of science. Finally, meetings with a geography requirement will lessen further, as technology makes global connections easier and more efficient.

InChI and the information chain

Steve Heller of the National Institute of Standards and Technology (NIST) gave an overview of the IUPAC International Chemical Identifier (InChI, http://www.iupac.org/inchi). InChI is a non-proprietary, freely available, machine-readable string of symbols that enables a computer to represent a compound in a completely unequivocal manner. InChIs are produced by computer from structures drawn on screen with existing structure drawing software, and the original structure can be regenerated from an InChI with the same software.

Like bar codes, and QR codes, InChIs are not designed to be read by humans. InChI should be thought of as "plumbing": a modern enabling technology. It is not something the average chemist needs to know about: researchers merely *use* it to find and link information on the Web. There is too much information on the Web and it lacks integration and connection; InChI is an infrastructure foundation that allows for linking, and hence for higher productivity. It is not a replacement for any existing internal structure representations; it is in *addition* to what is used internally. There are four amusing videos on the Web that explain InChI simply: *What on Earth is InChI?* (http://www.youtube.com/watch?v=rAnJ5toz26c), *The Birth of the InChI* (http://www.youtube.com/watch?v=X9c0PHXPfso), *The Googlable InChIKey* (http://www.youtube.com/watch?v=UxSNOtv8Rjw), and *InChI and the Islands* (http://www.youtube.com/watch?v=qrCqJ0o4jGs).

Some people question why InChI should be used instead of SMILES. SMILES is a popular line notation but it is not a published standard. Each vendor has a different implementation of SMILES, so strings cannot reliably be compared. SMILES has no structure normalization, so different structural representations yield different SMILES strings: a subscriber to chminf-l has reported finding 172 different SMILES representations for caffeine on the Web. InChI is easy to generate using existing software, expressive of

structural information, unique and unambiguous, and amenable to searching for structures using Internet search engines (using a hash key).

The InChI standard was developed by consensus, by a technically competent team, with political and technical cooperation. The work involved precompetitive collaboration among publishers, database producers, and software vendors. InChI is not in competition with commercial products, it has no "mission creep", and it is endorsed by IUPAC. The standard has been widely adopted: there are, for example, tens of millions of InChIs in each of PubChem, ChemSpider and Reaxys, and an InChI can be input to SciFinder to search the 89 million compounds in CAS REGISTRY.

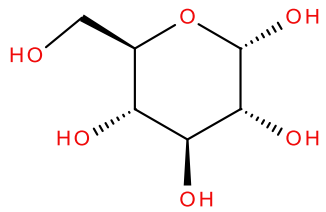
If the work of the InChI project were to endure, it needed to be turned over to an entity that would ensure its ongoing activities, and be acceptable to the community. A not-for-profit organization was best; hence the decision to create and incorporate the InChI Trust (http://www.inchi-trust.org) as a UK charity. The Trust has about 60 Members, Associate Members, and (non-paying) Supporters. The InChI project has experienced remarkable cooperation and support. It is a truly international project with programming in Moscow, computers in the cloud, incorporation in the United Kingdom, and a project director in the United States. Collaborators from over a dozen countries, from academia, the pharmaceutical industry, publishers, and the chemical information industry, have all offered senior scientific staff to develop the InChI standard.

Organizations need a structure representation for their content (databases, journals, chemicals for sale, etc.), so that it can be linked to and combined with other content on the Internet. InChI provides an excellent return on investment and increases productivity. It is a freely available, open source algorithm that anyone, anywhere can freely use, and it is certainly widely used: its success is proved by un-coerced adoption. InChI's combination of the Internet, open source software, crowdsourcing, graph theory, existing representation algorithms, digitized data available on the Web, and search engines has created a very valuable tool. This is taking advantage of "the second machine age",[6] which includes "recombinant innovation", or mashups.

InChI is a "layered" line notation, currently with the following layers:

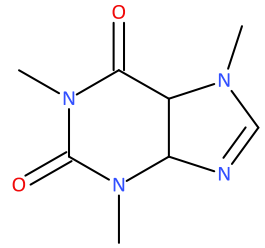
- Formula
- Connectivity (no formal bond orders)
    - disconnected metals
    - connected metals
- Isotopes
- Stereochemistry
    - double bond (*Z/E)*
    - tetrahedral (sp3)
- Tautomers (on or off).

Charges are added to the end of the string. Layers are separated by slash marks. Opening characters before the formula denote the version of the algorithm used. An example is alpha-D-glucose:



InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6?/m1/s1

The InChI algorithm normalizes chemical representation, and includes a “standardized” InChI, and a “hashed” form called the InChIKey. The key facilitates Web searching, previously complicated by unpredictable breaking of InChI character strings by search engines. The “standard InChI” and InChIKey for caffeine are shown below.



InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3
InChIKey=RYYVLZVUVIJVGH-UHFFFAOYSA-N

The first block of 14 letters of the InChIKey (RYYVLZVUVIJVGH) encodes the molecular skeleton (the connectivity). The first eight letters of the second block (UHFFFAOY) encode stereochemistry and isotopes. After that, “S” indicates that the key was produced from standard InChI and “A” indicates that version 1 of InChI was used. The final character, “N”, means “neutral”. The first 14 characters of an InChIKey can be used to search for structures with the same skeleton (e.g., to find all stereoisomers).

The InChI certification suite is a software package designed to check that an installation of the InChI program has been performed correctly; it ensures that InChIs have been generated properly and consistently. Currently, InChI handles straightforward organic molecules; it is being extended to handle more complex entities such as organometallics, Markush structures, macromolecules, and reactions.

Virtual communities and beyond

Wendy Warr, of Wendy Warr & Associates (the author of this *CIB* article), looked at the evolution of virtual communities and publishing platforms. As research has become increasingly collaborative, possibilities for communication and collaboration on the Web have also increased. Virtual communities in science such as EiVillage and BioMedNet began to spring up in the 1990s. The earliest virtual community in chemistry, ChemWeb.com,[7] was announced in August 1996 by MDL and Current Science Group, and launched in April 1997. It was acquired by Elsevier in October 1997. Elsevier closed all its “portals” in the middle of 2003, and ChemWeb.com was sold to ChemIndustry in April 2004.

ChemWeb.com was a pioneer in virtual conferences: the first was held in December 1997. The technology (interactive chat alongside PowerPoint slides and audio) was hardly ready to support such innovations at that time. Many chemists were unwilling to register themselves into a virtual community in the 1990s, but by 2002, ChemWeb.com had 300,000 members, and it offered 350 journals, 25 databases, structure searching, a careers center, a conference center, a bookstore, a magazine (*The Alchemist*), 11 forums, and a preprint server, the Chemistry Preprint Server (CPS, http://www.sciencedirect.com/preprintarchive).

The CPS was launched as an experiment in 2000. By Elsevier's own criteria it was a partial success: the number of readers and their geographic spread were excellent; the number of preprints (466) was encouraging, but rather less than had been hoped for; and, unfortunately, it was difficult to ascertain the number of preprints going on to traditional publication. The CPS was terminated in 2002. An evaluation has been published.[8]

Chemistry as a discipline has been slow in adopting open access, but some journals merit a mention: *Chemistry Central Journal*, *Journal of Cheminformatics*, the *Beilstein Journal of Organic Chemistry*, *Frontiers in Chemistry*, *Chemical Science,* and *ACS Central Science.* ChemSpider (http://www.chemspider.com/), in particular, is worthy of note. Most of the publishing services discussed in the rest of this talk concern biology, medicine and biomedical sciences rather than chemistry.

In 2004, the year that ChemWeb.com changed hands, the term "Web 2.0" was first coined by O'Reilly. Not everyone would agree on the definition, or usefulness, of the term, but the era of wikis, blogs, feeds, podcasts, webinars, social networks, social bookmarking, and virtual worlds had begun.[9] Facebook had 66 million users in March 2008: compare that with ChemWeb's 300,000 in 2004. People are no longer afraid of signing up to virtual communities. The traditional peer review process can now be challenged.

Peerage of Science (https://www.peerageofscience.org/), for example, promotes one peer review process for multiple journals, in ecology, and evolutionary and conservation biology. It is run by a for-profit organization, funded by publishers etc. The reviews themselves are peer-reviewed, but there is no journal editor in control of the peer review process. As of July 2014, 176 manuscripts have received 381 peer reviews, and there have been 942 peer review evaluations. Rubriq (http://www.rubriq.com/) also supports one peer review process for multiple journals. It offers independent, double-blind peer review and manuscript submission, recommends a suitable journal, and provides a score-card for reviews. The reviewer is rewarded financially, while the author pays for the R-score etc. Axios Review (http://axiosreview.org/) is another independent review service in ecology and evolutionary biology. The author aims at a top choice journal and "plays safe with others". (Few manuscripts completely fail to be published, rather they are resubmitted, after rejection by one journal, to a less prestigious journal, in the so-called "journal cascade".) eLife, BioMedCentral, PLoS and EMBO have started a peer review consortium in which papers are redirected with reviewer reports. Reviewers are anonymous to the author, but anonymity is optional for the journal cascade editor. SciRev (https://scirev.sc/) claims to be

"speeding up scientific knowledge": authors rate journals on efficiency and seek an efficient journal. An editor can compare his or her own journal with competitors.

So-called "mega-journals" (*PLoS ONE*, *PeerJ*, and *eLife)* are another trend. eLife arose from the San Francisco Declaration on Research Assessment (DORA, http://am.ascb.org/dora/). It is opposed to journal Impact Factors and runs a consolidated, pre-publication peer review service. Researchers can read and publish in *eLife* for free; the journal is supported by the Howard Hughes Medical Institute, the Max Planck institutes, and the Wellcome Trust. As of July 2014, it had published 488 articles, 85 of them in biochemistry. *PeerJ* is a peer-reviewed journal and preprint server in biological and medical sciences. It offers cheap, lifetime accounts (pre-paid). An editor handles pre-publication peer review. Some reviews (about 40%) are not anonymous. It links to Publons (*vide infra*). As of July 2014, it had published 476 articles and 433 preprints.

Frontiers (http://www.frontiersin.org/) was launched in 2007 by scientists from the Swiss Federal Institute of Technology, Lausanne, with a major investment by Nature Publishing Group. Its journals are largely in medicine etc., but also in chemistry, earth science, ecology and evolution. It offers open peer review in two phases (independent and interactive); anonymity is not allowed. Analytics automatically track views and downloads. The Frontiers evaluation system allows an entire community to score a paper. *Frontiers* has published 20,000 articles in 45 community-driven journals. The business model is author-pays. The community also shares jobs postings. Other communities or social networks include Academia.edu (http://www.academia.edu/) and ResearchGate (http://www.researchgate.net/), where millions of researchers share papers, see analytics, and follow other people. The data repository figshare (http://figshare.com/) has collaborative space in the cloud.

A number of centralized commenting platforms have sprung up. PubMed Commons (http://www.ncbi.nlm.nih.gov/pubmedcommons/) enables authors to share opinions and information about scientific publications in PubMed. Publons (https://publons.com/) collects peer review information from reviewers and publishers, produces reviewer profiles with publisher-verified peer reviews, and handles pre- and post-publication peer review. Authors (and a few peers) are notified of comments, and reviewers get credit in the form of DOIs. As of July 2014, 1,954 reviewers had produced 4,247 reviews. The service is free to academics. PubPeer (https://pubpeer.com/) offers anonymous post-publication peer review. Users can comment on any scientific article with a DOI, or on an arXiv preprint, adding comments to a centralized database. Authors and other interested parties are alerted to comments. Journal Lab (http://www.journallab.org/) handles open summaries and peer review of PubMed papers, but anonymity is optional. It also offers journal clubs and discussions.

A few "publishing platforms" such as Faculty of 1000 (http://f1000.com/), ScienceOpen (https://www.scienceopen.com/) and The Winnower (https://thewinnower.com/) offer a wider range of services. Faculty of 1000 has F1000Prime (http://f1000.com/prime) literature filtering, and F1000Research (http://f1000research.com/), an open access journal and journal club, with open post-publication peer review. It has published 522 articles. F1000Posters (http://f1000.com/posters) was launched recently.

ScienceOpen is an open access, research and publishing network, launched on May 29, 2014. It features almost 1.3 million articles from PubMed Central and arXiv, by 2 million networked authors. It will publish all sorts of article types in the sciences, humanities, and social sciences. It offers collaborative pre-publication workspaces where authors can manage draft versions and share files, and easily collaborate on a paper. Authors get almost immediate publication with a DOI. Reviewers get DOIs for their open, post-publication peer reviews. Article metrics are powered by Altmetric. Authors benefit from automatic proofs, easy corrections, and versioning. User roles are allocated based on ORCID publication history. Public and private groups can be constructed.

In the world of open access, open data and open science what might happen next? It certainly seems likely that there will be consolidation (or closure) among the services mentioned above. Does Science Open have anything to learn from ChemWeb.com? ChemWeb was ahead of its time. If Elsevier had hung on to ChemWeb for just a short time longer, it would have had a ready-built community for Article of the Future and its other ventures. More than a decade after the Chemistry Preprint Server closed, chemistry still has a different culture[10,11] from other disciplines, but a number of members of ScienceOpen's scientific advisory board were in the audience, wishing the new venture well.

Reaxys: a digital transformation

Sebastian Radestock of Elsevier Information Systems gave this talk, replacing David Evans of Reed Elsevier Properties, who was indisposed. Sebastian talked about the long road leading to the current version of Reaxys (http://www.elsevier.com/online-tools/reaxys). Reaxys has its origins in the preeminent Gmelin and Beilstein Handbooks, begun by Prof. Leopold Gmelin in 1817 and Prof. Friedrich K. Beilstein in 1881. In the 1980s the focus was on database development: SANDRA (a structure-based tool to locate references in the Beilstein Handbook), and the Gmelin Formula Index were released in 1987-1988, and the Gmelin and Beilstein databases went online on STN and Dialog in 1989-1990.

Since then, the development focus has been the user. CrossFire was launched and improved between 1993 and 1995, and the printed Handbooks were discontinued in 1997-1998. The Patent Chemistry Database was launched in 2005. In 2009, Reaxys was launched, based on Gmelin, Beilstein, and the Patent Chemistry Database. In 2013, Reaxys was completely overhauled and its scope was expanded. In 2014, it was upgraded with re-indexing and concept search.

The 1989 version of Beilstein on STN was text-based; knowledge of the database structure and STN commands was needed; and not all query forms or results were readily accessible. CrossFire Commander in 1996 was a graphical, client-server based system. Access was improved with input forms, and structure and reaction searches could be combined with factual queries, but displays were rather cluttered. The 2014 version of Reaxys has a subject-oriented, customizable, Web-based user interface; concept search is enabled; all types of chemical information needs are supported; and there is one-click access to advanced-query forms for multiple source databases.

The CrossFire database structure had three different, tightly connected contexts: substances and properties, reactions and reaction details, and citations and abstracts. This concept is still valid today: Reaxys is a bibliographic database with more than 46 million records from 16,000 journal titles; it is a

substance database with more than 57 million unique substances and more than 500 million experimental facts; and it is a reaction database with more than 36 million single- and multi-step reactions. The database structure is built around a sustainable chemical substance model for single compounds, component compounds, and Markush compounds.

Chemists like to do graphical structure searching, but Elsevier wondered if they would also like to start searching for a topic by typing text. The company therefore surveyed 700 chemists with a wide range of job roles, years of experience, and worldwide locations. The chemists were from many different areas of interest (14% were in materials chemistry, 9% in organic chemistry, and 4% in electrochemistry, for example), and all sorts of organizations in many sectors.

The survey revealed a definite need for keyword search capabilities. On average, chemists search for chemistry-related information five times a week. Researchers in organic, inorganic, medicinal and organometallic chemistry are the most intensive searchers. Some 70% of the chemists are keyword searchers (in that they spend more than 60% of their time searching for keywords); 10% are structure searchers; 20% search for both keywords and structures. Structure search is highest among organic and inorganic chemists.

The searching pattern for both structure search and keyword search is similar. The top four use cases are:

- find reviews, introductory articles and other starting points for research
- find the very latest information on a certain topic
- search and retrieve substance properties, and
- compile a comprehensive survey of the published literature.

Keyword searchers more often search for reviews and introductory articles, properties, and comprehensive surveys of the published literature. There are some use cases which are ideally supported by structure searching.

Based on these results, Elsevier developed Ask Reaxys to be more than just a text search. Text input is analyzed to identify all possible queries; all queries are assigned a probability factor; and the query exceeding a certain probability factor threshold is automatically selected, or the user is prompted to select a query manually from the list of possible queries. Each word, or group of words, can be classified as bibliography, compound, concept, date, or keyword, or it can be ignored. An example is “electrical conductance of titanium”. This query could be “electrical conductance” as a concept and “titanium” as a compound; or it could be “electrical conductance” and “titanium” as keywords with “of” ignored. The former option is translated into a combined structure and factual query, and this is executed in the substances context (on a substances tab in the Reaxys interface). The latter is translated into a pure keyword query, which is executed against all text fields in the citations context. To improve the relevancy of the results, all citations and abstracts have been indexed, and enriched with additional Elsevier keywords.

For text input analysis, the input string is tokenized and each token is annotated using

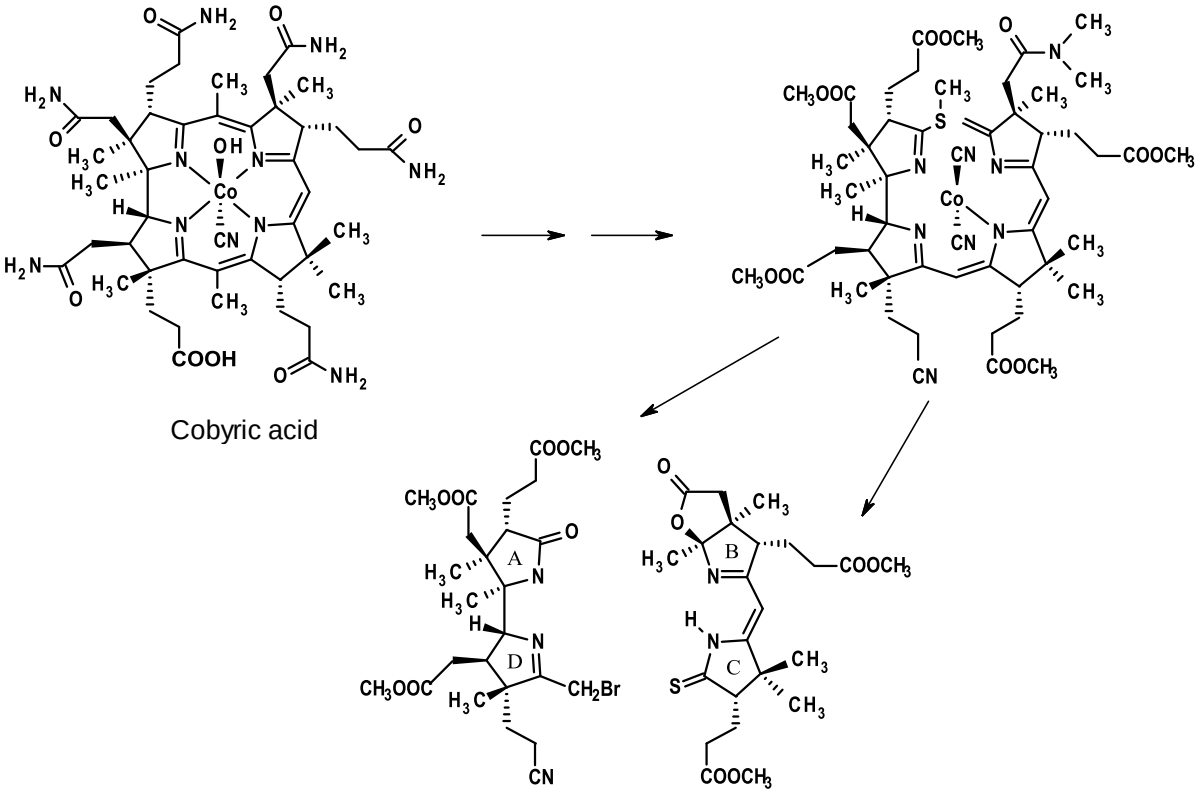
- a proprietary tool for chemical entity recognition
- a regular expression for third party registry numbers and InChIKeys
- a list of author names
- a regular expression for dates
- a list of words that can be ignored, and
- a large chemistry taxonomy.

Text input analysis also involves annotation clean-up. Query translation takes the annotated text input and returns an advanced database query string.

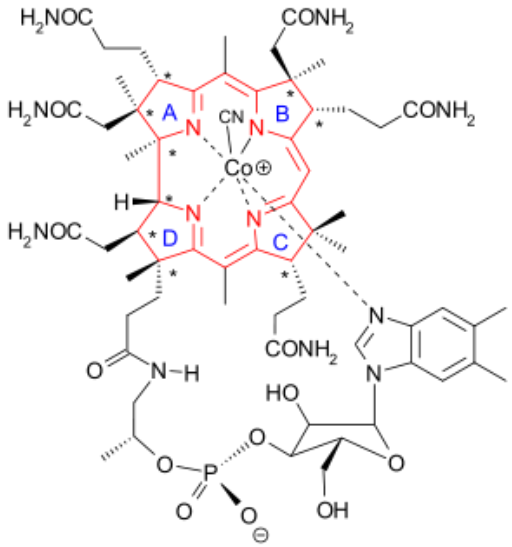
ReaxysTree, a new chemical taxonomy, has been developed. Tree development started with looking at the Reaxys data structure (the field codes) and the database content. The tree was further extended using keywords from more than 46 million records from 16,000 journal titles. Synonyms and spelling variants were also added to each term. ReaxysTree contains more than 15,000 concepts with more than 40,000 synonyms. It is poly-hierarchically structured and organized in several chemistry-related facets. There are broad and narrow terms, each with a unique label ID, language, date, source, type, case sensitivity, etc. Elsevier's taxonomy construction rules were followed. ReaxysTree is used for indexing the content in a continuous process of application and learning. After automatic indexing has been carried out, statistical analysis reveals new synonyms and additional terms, which are subjected to editorial work, and then fed back into ReaxysTree before further automatic indexing. The Ask Reaxys keyword search functionality is now prominently available at the top of the Reaxys query page (with a Google-like appearance). ReaxysTree can also directly be accessed, searched and browsed.

Award address

Finally, Bert gave his own presentation. In the last century, publications about total syntheses in journals would usually have full experimental details. A remarkable exception was the total synthesis of vitamin B12, considered a landmark in organic synthesis, involving the research groups of Robert Burns Woodward at Harvard, and Albert Eschenmoser at ETH Zürich. The synthetic target was actually cobyric acid, because this compound had already been converted to Vitamin B12.[12] Hence total synthesis of cobyric acid would amount to a formal synthesis of Vitamin B12.



There are two variants of the total synthesis, differing in their overall strategy of creating the corrin system (in red in the vitamin structure below).



The variant collaboratively pursued closes the macrocyclic corrin ring between rings A and B (the "A/B variant"), while the synthesis accomplished at ETH achieves the corrin ring closure between rings A and D by a photochemical process (the "A/D variant"); the final steps toward cobyric acid were jointly carried out at Harvard and ETH, using material from the respective variants. Woodward reported on the A/B variant in lectures published in 1968, 1971, and 1973, culminating in the announcement of the total synthesis of the vitamin in his lecture[13] at the IUPAC Conference in New Delhi, in July 1972. Eschenmoser discussed the ETH contributions to the A/B variant in his Centenary Lecture,[14] published in 1970, and presented the approach to the photochemical A/D variant of the B12 synthesis at the 23rd IUPAC Congress in Boston, published in 1971 (http://e-collection.library.ethz.ch/eserv/eth:8691/eth-8691-01.pdf).[15] A full report on the photochemical variant is given in a *Science* article[16] which is an extended English translation of an article based on a lecture[17] by Eschenmoser.

Seventy-seven postdoctoral students, but no Ph.D. students, worked on the project at Harvard between August 1961 and December 1975. Twelve Ph.D. students and 14 postdoctoral students worked on the project at ETH between September 1960 and August 1974. Research records consist of 67 postdoctoral reports (not publicly accessible) and 48 individual experimental procedures at Harvard, and 12 Ph.D. theses (publicly accessible) one diploma thesis (not publicly accessible) and 6 postdoctoral reports (not publicly accessible) at ETH. Unfortunately theses are not indexed in *Chemical Abstracts* and most European theses are not even covered by *Chemical Abstracts* or *Dissertation Abstracts*.

Bert found that a SciFinder search for "total synthesis of vitamin B12 (total synthesis of cobyric acid)" does not retrieve the significant publications by Woodward and Eschenmoser, although they can be found in Google. Web of Science finds Eschenmoser's paper in *Science*[16] and is the only database to find two papers in Chimia.[18,19] (These references are just one short abstract about two talks given at a Swiss Chemical Society Meeting, so it is perhaps not surprising that they are not found in SciFinder.) Scopus finds Eschenmoser's paper in *Science*,[16] a paper in Japanese that is not found by SciFinder, although it is in *Chemical Abstracts*, and a paper by Wintner with recollections of ETH.[20] Note the absence of papers by Woodward. Woodward's 1972 lecture *is* found in Scopus and SciFinder if you include a space between B and 12.

In summary, the primary publication record is incomplete: Harvard has no experimental details and ETH has them only in theses. In the secondary literature, too many publications are missed altogether and those that *are* covered are hard to retrieve. The tertiary publication record (plus the Web and Wikipedia) is often incomplete, and the change of paradigm exemplified by the two variants of the total synthesis is not recognizable.

In a lecture given at Wesleyan University on September 29, 1972, Woodward gives the only complete list of both Harvard and ETH co-workers ever made public so far; 50 Harvard postdoctoral students are not mentioned in other lectures published by Woodward. Only the Harvard-ETH A/B route is mentioned in this 1972 lecture; the A/D alternative is not. The tapes and slides of this lecture are no longer available, but Bert has a shorthand transcript by Eschenmoser's secretary, Miss H. Gächter (now Frau Zass).

The ETH publication project started by Eschenmoser and the Zasses in 1979 is without precedent. It seeks to produce a high-quality, fully documented record of 2,123 man-months of work, 3,732 pages of postdoctoral reports and procedures, and 1,889 pages of Ph.D. theses. Bert started by applying for all the Harvard records and recording individual reaction steps; 75 out of 77 postdoctoral students are now covered; two Russians are not included, for lack of information. Bert showed pages and pages of reports, listings, strategies, and handwritten records. Attribution is given to scientists for the specific work they did, with detailed comments; everyone's work should be acknowledged. The summaries for 238 compounds, with nomenclature and spectra, were at first recorded on edge-notched cards, in different colors to distinguish Harvard and ETH. Reaction pathways, and flow charts for synthesis pathways and reaction details were drawn up. Compound-centered manuscripts and modular, standalone, experimental descriptions were produced. Literature references were listed and standards were drawn up for the data, solvents, reagents etc. All this work was patiently typed by Frau Zass and then retyped when corrections were needed.

Between 1979 and 1986, 599 handwritten pages were typed on a Remington typewriter and later retyped into Macintosh Word; 210 pages were processed on an Olivetti ETV 300 and later converted into Macintosh Word. From 1984 a computer graphics program was used. By 1983 the Harvard ring A/D work was finished; ETH rings B/C and D were finished in 1986. Bert had other work to do for ETH at this time, including most non-routine online searches, but the project was resumed in January 1990. The Olivetti work was converted in 1990. Corrections to the A/D seco-corrin records were completed by June 1991, and retyping of the Remington material was finished in February 1992. In 2006, work began to add theses to the ETH collection. By February 2009, the Macintosh Word 5.1 files were converted to Windows 97-2003. Unfortunately for the project, Bert and Eschenmoser had an enormous number of other commitments between 1984 and 2012, and the final steps carried out at Harvard and ETH are still missing. Bert is now resuming work on the project (in 1979, his first cheminformatics project) and he intends to finish it as a tribute to all the chemists involved, and in particular to Albert Eschenmoser.

Conclusion

The symposium was ably chaired by Andrea Twiss-Brooks. After Bert's award address, Judith Currano, chair of the ACS Division of Chemical Information, formally presented the Herman Skolnik Award:

References

(1) Bergstrom, T. C.; Courant, P. N.; McAfee, R. P.; Williams, M. A. Evaluating big deal journal bundles. *Proc. Natl. Acad. Sci. U. S. A.* **2014,** *111* (26), 9425-9430.

(2) *Are Chemical Journals Too Expensive and Inaccessible? A Workshop Summary to the Chemical Sciences Roundtable*. Heindel, N. D.; Masciangioli, T. M.; Schaper, E. v., Eds.; The National Academies Press: Washington, DC, 2005.

(3) Galilei, G. *Discorsi e dimostrazioni matematiche intorno à due nuove scienze*. Elzevir: Leiden, The Netherlands, 1638.

(4) Boyle, R. *The Sceptical Chymist*. J. Cadwell for J. Crooke: London, England, 1661.

(5) Morgan, H. L. The generation of a unique machine description for chemical structures - a technique developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965,** *5* (2), 107-113.

(6) McAfee, A.; Brynjolfsson, E. *The Second Machine Age*. W. W. Norton & Company: New York, NY, 2014.

(7) Warr, W. A. Communication and communities of chemists. *J. Chem. Inf. Comput. Sci.* **1998,** *38* (6), 966-975.

(8) Warr, W. A. Evaluation of an Experimental Chemistry Preprint Server. *J. Chem. Inf. Comput. Sci.* **2003,** *43* (2), 362-373.

(9) Warr, W. A. Social Software: Fun and Games or Business Tools? *J. Inf. Sci.* **2008,** *34* (4), 591-604.

(10) Velden, T.; Lagoze, C. Communicating chemistry. *Nat. Chem.* **2009,** *1* (9), 673-678.

(11) Velden, T.; Lagoze, C. The extraction of community structures from publication networks to support ethnographic observations of field differences in scientific communication. *J. Am. Soc. Inf. Sci. Technol.* **2013,** *64* (12), 2405-2427.

(12) Friedrich, W.; Gross, G.; Bernhauer, K.; Zeller, P. Synthesen auf dem Vitamin-B12-Gebiet. 4. Mitteilung Partialsynthese von Vitamin B12. *Helv. Chim. Acta* **1960,** *43* (3), 704-712.

(13) Woodward, R. B. The total synthesis of vitamin B12. *Pure Appl. Chem.* **1973,** *33* (1), 145–178.

(14) Eschenmoser, A. Centenary Lecture. (Delivered November 1969). Roads to corrins. *Q. Rev., Chem. Soc.* **1970,** *24* (3), 366-415.

(15) Eschenmoser, A. Studies on Organic Synthesis. *Pure Appl. Chem. Suppl.* **1971,** *2*, 69-106.

(16) Eschenmoser, A.; Wintner, C. E. Natural Product Synthesis and Vitamin B12. *Science* **1977,** *196* (4297), 1410-1420.
(17) Eschenmoser, A. Organische Naturstoffsynthese heute Vitamin B12 als Beispiel. *Naturwissenschaften* **1974,** *61* (12), 513-525.
(18) Fuhrer, W.; Schneider, P.; Schilling, W.; Wild, H.; Shreiber, J.; Eschenmoser, A. Totalsynthese von Vitamin B12: die photochemische Secocorrin-Corrin-Cycloisomerisierung. *Chimia* **1972,** *26* (6), 320.
(19) Maag, H.; Obata, N.; Holmes, A.; Schneider, P.; Schilling, W.; Schreiber, J.; Eschenmoser, A. Totalsynthese von Vitamin B12: Endstufen. *Chimia* **1972,** *26* (6), 320.
(20) Wintner, C. E. Recollecting the Institute of Organic Chemistry, ETH Zürich, 1972–1990. *Chimia* **2006,** *60* (3), 142-148.