Herman Skolnik Award Symposium 2016

Honoring Stephen Bryant and Evan Bolton

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

Introduction

Stephen Bryant and Evan Bolton were selected to receive the 2016 Herman Skolnik Award for their work on developing, maintaining, and expanding the Web-based National Center for Biotechnology Information (NCBI) PubChem database, and related software capabilities and analytical tools, to enhance the scientific discovery process. NCBI is part of the United States National Library of Medicine (NLM), a branch of the National Institutes of Health (NIH). A summary of Steve and Evan's achievements has been published in the *Chemical Information Bulletin.* They were invited to present an award symposium at the Fall 2016 ACS National Meeting in Philadelphia, PA. They invited twelve speakers:

L to R: Valery Tkachenko, Roger Sayle, Leah McEwen, Steve Heller, Wolf-Dietrich Ihlenfeldt (partially obscured), Yulia Borodina, Peter Linstrom, Steve Bryant, Marc Nicklaus (at front), Evan Bolton (at back), Steve Boyer, Daniel Zaharevitz, Christoph Steinbeck. Not pictured: Michel Dumontier (inset)

Developing databases and standards in chemistry

Steve Heller was the first speaker, with an amusing scene-setting talk. He admitted that his secret in getting to where he is now was "luck, luck, luck". He disliked chemistry lab work; he was at the right place at the right time with the right people; he worked with supportive people; and he planned for who would take over the work next. If the problem were just technology, someone would have solved it already. The real problem is always cultural and political, not technical. Steve had the good luck to be at NIH to collaborate with Hank Fales and Bill Milne; at the Environmental Protection Agency (EPA) with Morris Yaguda, when EPA started using mass spectrometry to identify pollutants; at the National Institute of Standards and Technology (NIST) with Steve Stein, when CAS stopped providing Registry Numbers to the NIST Mass Spectrometry database; and to be retiring just when Ted Becker and Alan McNaught thought that the International Union of Pure and Applied Chemistry (IUPAC) needed to move into the 21st century of chemical structure representation.

The NIH/EPA/NIST mass spectrometry database[1,2] originated at MIT (with Klaus Biemann), and was run at NIH in the 1970s using a modification of Richard Feldmann's search software. Control moved to EPA, and eventually to NIST in the 1980s. NIST was the right home for the database: NIST now collects a few million dollars a year in mass spectrometry database royalties. The NIH/EPA Chemical Information System (CIS)[3] was a collection of chemical structures with links to various databases supporting environmental and scientific needs. It also had a number of analysis and prediction programs. All the databases had CAS Registry Numbers[4] as their link. The CIS worked for a number for years, but never had the full support of the government or of ACS. It died in the mid-1980s; it was a bit ahead of its time.

Steve's next example of luck dates back to November 1999 when he and Steve Stein seeded the idea of a chemical identifier. The right people in this case were the IUPAC International Chemical Identifier (InChI) team: Steve himself, Alan McNaught, Igor Pletnev, Steve Stein, and Dmitrii Tchekhovskoi. InChI [5] was developed as a freely available, non-proprietary identifier for chemical substances that can be used in printed and electronic data sources, thus enabling easier linking of data compilations, and unambiguous identification of chemical substances. It is a machine-readable string of symbols which enables a computer to represent a compound in a completely unequivocal manner. The InChI algorithm normalizes chemical structures and includes a "standardized" InChI, and the hashed form called the InChIKey. InChI is easy to generate, expressive, unambiguous and unique and it does not require a centralized operation. It enables structures to be searched by Internet search engines using the InChIKey.

InChI is not a replacement for any existing internal structure representations, but an addition to them. Its value is in finding and linking information. The proof of its success is in its widespread adoption.[6] All the major structure drawing programs have incorporated the InChI algorithm in their products. There are millions of InChIs in large chemical databases. Regardless of controversies and differing opinions, InChI has been more widely adopted than SMILES. Currently, the InChI algorithm can handle neutral and ionic organic molecules, radicals, and some inorganic, organometallic, and coordination compounds. Steps to expand it to handle more complex chemical structures are underway, under the auspices of the InChI Trust.

Finally, Steve had the luck to join the PubChem Advisory Board, and worked with the right people, Steve Bryant and Evan Bolton. The database now contains nearly 92 million compounds, 223 million substances, and 1.2 million bioassays, and related data and publications. More than 100,000 searches are carried out every day, by 1.6 million unique users in a month. The success of PubChem, like that of InChI, is measured by its widespread use.

Two decades of open chemical data at the Developmental Therapeutics Program (DTP) at the National Cancer Institute (NCI)

The talk by Daniel Zaharevitz of NCI also covered freely available chemical and biological data. A history of DTP/NCI was posted on the Web on the 50th anniversary of the Cancer Chemotherapy National Service Center (CCNSC), which was set up in 1955. Until 1990, transplantable mouse tumors were used and gram quantities of test substances were needed. After that, human tumor cell lines in culture (the "NCI-60" cell lines) were used and only milligram quantities of test substances were needed.

The philosophy behind the National Chemotherapy Program[7] was one of hundreds of independent investigators who were not required to collaborate. Indeed, over the last ten years, 42,301 compounds have been submitted from 1,477 different groups. Consequently data and decision making have been compartmentalized, and data systems development has reflected this compartmentalization. There was little pressure to apply any standardization.

From the 1970s until 2000 the Drug Information System was part of the CIS Structure and Nomenclature Search System (SANSS). Since 2000 there has been a Web interface for compound submission, accepting structures in only molfile format. Before 1994 there was no policy for making chemical structures publicly accessible. Data release was avoided if possible because of the costs and difficulties involved, and because there was no perceived advantage. In 1994, 127,000 structures for which there was a CAS Registry Number were made available via FTP, after SANSS connection tables had been converted to molfiles, and CORINA had been used to generate 3D coordinates. Since 2000, molfiles have been extracted from a newer internal system, and structures are released about once a year on a Web page. In June 2016 there were 284,176 open NCI structures, but there are many versions of "NCI structures" around, including multiple depositions in PubChem.

DTP compound submissions are now performed online. The submitter must register as a user and the submission must include structures, which are subjected to consistency checks (with the Chemistry Development Kit, CDK), and stereochemistry consistency checks (with InChI). A material transfer and screening agreement is signed electronically, and, nowadays, the confidentiality period is limited to three years. Submitters are given access to screening results and to COMPARE analysis. Researchers can request samples or plated sets from a collection of about 100,000 compounds, if they submit a material transfer agreement electronically, and pay for shipping.

There is no science without communication, including communication with a more general audience, as well as with those immediately involved. Despite the barriers to widespread communication, it is

important to do *something*. Note also that good communication of data is hard work, and attention to detail is critical.

The earliest plans[8] for PubChem recognized the need for significant resources to store and disseminate data. NLM was a natural choice for this function, and Steve Bryant was brought in early in the implementation process. Evan Bolton came in when the nuts and bolts implementation started. When PubChem went live, about a third of the structures and all of the biological data were from DTP. In less than 20 years the world of open chemical structures has gone from about 100,000 compounds in a single file to millions of structures being freely available in a searchable database.

In future, more applications will be built based on PubChem data. "Chemical awareness" should be integrated into the publication process, especially peer review. In future, data consistency will be improved, and we will be more able to know the context for structures and data, and to find out which similar structures are known and which assays have been run on them. Researchers will use predictive tools more as a measure of surprise than as a substitute for measurements.

Using InChI to manage data

To explain the usefulness of InChI, Peter Linstrom of NIST started by defining a problem as follows. "I have data about a substance and my colleague has data about a substance. Are these substances the same so that we can combine the data about them? Are we talking about well-defined molecular species?" The term "well-defined" can mean different things to different people. A well-drawn structure can precisely identify a molecule, but there are issues with formats and drawing conventions. Drawing a structure from a name by itself does not improve identification because additional information is often required to improve specificity. Moreover, sometimes we do not have a "well-defined" molecular structure. This is a general problem which cannot be solved for a significant portion of historical data.

InChI can help because it identifies a molecule based on its structure, and it allows us to ask whether two "well-defined" structures are the same. Also, InChI has a layered design allowing matches to related compounds such as stereoisomers, geometric isomers, and "isotopologues" (compounds that differ only in isotopic composition). In addition, with a little string manipulation we can ask even more questions.

An InChI is hierarchically layered. There are several InChI layer types, each representing a different class of structural information. These include: formula, connectivity, geometric and stereo isomerization, isotopic composition, charge, and protonation state layers. Layers are separated by a forward slash. Consider the two isomers of carvone, the InChIs of which differ only in the stereochemical layer (emboldened in the following). One isomer smells of spearmint and has

InChI=1S/C10H14O/c1-7(2)9-5-4-8(3)10(11)6-9/h4,9H,1,5-6H2,2-3H3/t9-/**m0**/s1.

The other smells of caraway and has

InChI=1S/C10H14O/c1-7(2)9-5-4-8(3)10(11)6-9/h4,9H,1,5-6H2,2-3H3/t9-/**m1**/s1

(The “1S” at the beginning of each string indicates a standard InChI.)

The NIST Chemistry WebBook provides an example of the use of InChI. It combines data from many sources. It is over 19 years old and there are many problems with identifiers from older datasets. Historically, CAS Registry Numbers and other accession numbers were used in matching species, but there were many problems (even the check sums in CAS Registry Numbers were wrong in one case out of ten). Newer data often come with structures and InChI can be used. Moreover, drawing structures can force additional analysis. Nevertheless there are still legacy data with incomplete identifiers (e.g., for stereoisomers and isoanalogues). An example is the species labeled as “gamma-elemene,” where 81 chromatographic retention values in the literature were analyzed, and found to correspond to five different chemical species (with similar mass spectra).[9]

PubChem is a great resource. Apart from the features that we all know and love, there are lesser known features that help disambiguate species. The substance database, separate from the compound database, records the mapping of names to structures by the various people who submitted the data. Partial InChIKey search allows compounds with the same composition and connectivity, but different information in further InChI layers, to be retrieved.

Voltaire said that perfect is the enemy of good. We cannot fix all chemical structure errors without abandoning valuable historical data, and newer data also are not immune to identification problems, but we can make progress where resources permit. There are tools such as InChI and PubChem that can help, but not solve the entire problem. “Zero Defects” was an industrial quality management approach championed in the 1960s and 1970 which was criticized as an exhortation to do something that may not be possible. Total Quality Management, the approach championed by W. Edwards Deming, is based on continuous improvement of systems, driven by measurement. It has been dramatically successful and has succeeded where “Zero Defects” failed. The transition from “econoboxes” in the early 1970s to modern, reliable compact cars did not happen overnight. Similarly, our chemical structure tools are getting better but we still have a long way to go.

Open chemistry resources provided by the NCI computer-aided drug design (CADD) group

NCI has a 60-year history of cheminformatics, starting with the drug development program authorized by Congress in 1955, said Marc Nicklaus, the leader of the NCI CADD group. By 1963, “it became clear the system must track not just individual chemical compounds, but distinct samples of chemical compounds…magnifying the data management problem considerably”.[10] This was a direct antecedent of the concept of separate PubChem Substance and Compound databases. The open NCI structure database was made publicly available in 1994 (see the talk by Daniel Zaharevitz, summarized above). The NCI Database Browser was, in 1998, the first public Web GUI for a large, small-molecule database, with advanced capabilities such as full substructure search. It arose from a collaboration between NCI and Wolf-Dietrich Ihlenfeldt at the University of Erlangen-Nürnberg. The Enhanced NCI Database Browser has 250,250 structure records and about 60 million data points: mostly Prediction of Activity Spectra for Substances (PASS)[11] predictions. Sophisticated search and output options are available.

The CACTVS Web Server offers many services, tools and downloadable datasets centered on small molecules. Apart from the database browser, Marc singled out the Chemical Structure Lookup Service (CSLS, pronounced "sizzles"), the Optical Structure Recognition Application (OSRA), and the Chemical Identifier Resolver (CIR). Developed by Igor Filippov in 2006, CSLS is a "phone book for chemical structures", linking 74 million indexed structures (46 million unique structures) to over 100 databases. OSRA, developed by Igor Filippov in 2007, converts graphical representations of chemical structures in journal articles, patents, or other text, into SMILES. CIR, developed by Markus Sitzmann in 2009, converts one structure identifier or representation into another. Its workflow involves lookups in the CADD group's chemical structure database (CSDB). CSDB contains about 121 million structure records for 85 million unique structures, in 140 databases, including PubChem, and the Sigma Aldrich iResearch Library.

It might be thought that the many large databases now available for CADD are enough, but perhaps we need a new approach. Perhaps we should not design a new molecule, and then ask how it can be made. Instead, we could look into what can be made reliably and cheaply, and then search only among those molecules for new, potentially bioactive compounds, using the usual CADD approaches.

Therefore, Marc's team has begun building the Synthetically Accessible Virtual Inventory (SAVI), using a set of highly predictive and richly annotated rules (transforms) from Lhasa Limited and Lhasa LLC, a set of reliably available and inexpensive starting materials from MilliporeSigma, and the cheminformatics engine CACTVS from Xemistry GmbH.

A parser has been implemented in CACTVS for the CHMTRN/PATRAN retrosynthetic transforms (of which there are more than 2,300), and it has been adapted for the forward-synthetic SAVI approach. Fourteen transforms have been implemented and used in production runs so far. Among the 3.3 million building blocks in sets from Sigma-Aldrich, and other catalogs, 377,484 compounds were identified as highly available, and in their majority annotated with pricing and availability data.

Using 11 "productive" transforms in one-step reactions, a sample subset of about 610,000 compounds was generated in summer 2015, and made available for download. It is annotated with (but not yet filtered by) 54 compound, reaction, and typical drug design properties. As of August 2016, 238 million products have been generated; it is estimated that there might 280 million when the runs are completed. Overlap with PubChem is minimal: more than 99% of the compounds appear to be novel.

Eleven new transforms are being added, and in future, products will be steered toward interesting novel rings and scaffolds. The product files will be offered for download. Multi-step reactions will be investigated in future, and a Web GUI with extensive search capabilities will be developed. Topics of the ongoing work are how the predicted synthetic routes will work in actual syntheses, what filter rate will be needed for truly "interesting" compounds, and how the editing and adding of transforms can be made as easy as possible.

Evolution of open chemical information

Valery Tkachenko of RSC continued the theme of open data in chemistry. Everything changed in 1992 with the arrival of the World Wide Web. Later, PubChem changed the world of chemical information. ChemSpider, a structure-centric hub for Web searching now contains 57 million compounds chemicals from over 500 different sources, and deposition of data is ongoing. It differs from PubChem in that curation and annotation are crowdsourced. ChemSpider has analytical data, text and literature references, and data on compounds and reactions. NextMove Software's text mining software has been used to analyze reactions from the RSC archive of journal articles, output CML, and break down each procedure summary into steps.

We are moving into the world of the Internet of Things and phones with modular, replaceable parts. Gartner has identified the Top 10 Strategic Technology Trends for 2016. Our world is hyperconnected, and connections require standards. The IUPAC "color books" took years to write, and thus data quality issues arose. Evan Bolton has referred to the proliferation of errors in public and private databases as "robochemistry". Manual curation of huge databases is not feasible but automatic quality control systems such as RSC's Chemistry Validation and Standardization Platform (CSVP) can be developed. CVSP allows users to upload chemical structure files which are then validated, and optionally standardized, in preparation for publication or submission to a chemical database. About 200 rules have been encoded, and expressed as XML, to check for errors in, for example, the depiction of stereochemistry. The community can amend these rules. The structure's relationship to names, SMILES, and other identifiers also needs checking.

Knowledge from the past is used to derive wisdom. The Open PHACTS discovery platform has been developed to reduce barriers to drug discovery in businesses and academia. It contains multiple data sources, integrated and linked together so that users can easily see the relationships between compounds, targets, pathways, diseases and tissues. The platform has been used to answer complex questions in drug discovery. It was built in collaboration with a large consortium of organizations involved in drug discovery, and is founded on Semantic Web and linked data principles. RSC developed the chemical data handling software for OpenPHACTS.

A high percentage of raw data is lost in the science data publishing workflow. Horizon 2020 is a very large EU research and innovation program. It already mandates open access to all scientific publications; from 2017, research data are open by default, with possibilities to opt out. In the era of Uber, transportation is now a commodity. Will scientific data become a commodity by 2020? How will publishers cope? Authorities have moved from centralized to decentralized to distributed, as we have moved into the hyperconnected world. We are on a verge of a new technical revolution; RSC is excited, and is ready to ride high on the wave of data science developments.

Open chemical information at the European Bioinformatics Institute

Christoph Steinbeck of EMBL-EBI looked back to his early years as a natural products chemist, and recounted what has happened since the old days of access to Beilstein and CAS in 1992. There were no open source software libraries for cheminformatics in those days, but there were computer-assisted structure elucidation (CASE) systems.[12,13] Christoph sold his CASE software to Bruker and it got buried. He learned that successful science requires data and software to be free and open.

So in 2000 he and his co-workers began work on an open source library for bioinformatics, cheminformatics, and computational chemistry written in Java: the Chemistry Development Kit (CDK).[14-16] Sixteen years later, it is a well-established, mature code base (564,171 lines of code), maintained by a large development team; 16,521 commits have been made by 115 contributors.

Christoph's database years really began when he moved to EMBL-EBI, although his open database NMRShiftDB [17,18] was written earlier. It contains 50,000 compounds and their spectra. Christoph's current research interest is documenting the metabolomes of all species on the planet. To coin Donald Rumsfeld's phraseology, "known knowns" can be found in databases, "known unknowns" can be found using NMRShiftDB, but "unknown unknowns" are dark matter. Too many metabolomes are not known.

EMBL-EBI has many important databases, Chemical Entities of Biological Interest (ChEBI) and ChEMBL being just two of them. ChEBI is a freely available dictionary of molecular entities focused on small chemical compounds. The molecular entities are either products of nature or synthetic products used to intervene in the processes of living organisms. ChEBI incorporates an ontological classification, whereby the relationships between molecular entities or classes of entities and their parents or children are specified. ChEMBL is an open data resource of binding, functional and ADMET bioactivity data for a large number of druglike compounds.[19] The types of data reported in PubChem and ChEMBL are distinct and complementary. To maximize the utility of the two datasets EMBL-EBI has worked with the PubChem group to develop a data exchange mechanism.

It is estimated that there are about 8.7 million eukaryotic species on earth, of which 1.2 million have been identified and classified. Three or four thousand complete species genomes have been sequenced. What about completed metabolomes? Steinbeck's team has argued that the time is now right to focus intensively on model organism metabolomes.[20] They have proposed a grand challenge to identify and map all metabolites onto metabolic pathways, to develop quantitative metabolic models for model organisms, and to relate organism metabolic pathways within the context of evolutionary metabolomics.

Species metabolomes are now being assembled through data sharing in metabolomics. MetaboLights[21-23] is an EMBL-EBI database for metabolomics experiments and derived information. It is cross-species, and cross-technique, and covers metabolite structures and their reference spectra as well as their biological roles, locations and concentrations, and experimental data from metabolic experiments. Christoph's team has reported one dataset[24] in the data publication *Scientific Data*.

History and the future of tools and software components for working with public chemistry data

Wolf-Dietrich Ihlenfeldt's CACTVS software suite has been an integral component of the PubChem software since the beginning. It handles structure searching, 2D structure layout and image rendering, submission checking, property computation, hashcodes, and a sketcher application. CACTVS is not used only in PubChem. The CACTVS scripting toolkit (solutions in Python or Tcl) is free for academia, and can be used in database cartridges and in KNIME nodes. It can give access to more than 50 Internet chemistry data sources.

One of the reasons CACTVS works particularly well with PubChem is PubChem's forward-looking design, including the PUG, Entrez E-utilities and REST interfaces which make it possible to access structured data by software without resorting to HTML page scraping. Additionally, CACTVS has some inherent advantages in performing these tasks: much of the PubChem engine is based on CACTVS, and CACTVS understands the native PubChem ASN.1 data formats for structures and assays, so it can process the original data content of PubChem, without format conversion losses. It is also possible to send native toolkit structure encodings directly to the PubChem query engine, which opens up query functionality which cannot be expressed by any standard structure query exchange formats, such as SMARTS or Query molfiles (which are, of course, supported by the query interface). An example of such advanced query functionality which will be made accessible on the PubChem side in the near future is querying for ring attributes which are not atom attributes, such as the overall ring atom formula, substituent counts and classes, and similarly also for ring systems, and even user-defined atom groups.

PubChem uses CACTVS hashcoding as a primary key (one-to-one mapping of hashcode to the PubChem compound identifier, called a CID); for mapping between CID and PubChem substance identifier (SID), for related compound links, and for a similarity boost scheme. The hashcodes are currently 64-bit pseudo-random numbers, but soon will be 128-bit. Computation is based on configuration-dependent atom seeds, and neighbor-coupled, atom-centric xor-feedback shift registers. The hashcodes are fast to compute: faster than SMILES and much faster than InChI. They are of constant length, and are independent of ring set, aromaticity system, and formal charge localization. Database performance is outstanding: identity is looked up on a fully indexed database field. PubChem variants of the codes include with or without stereochemistry, and with or without isotope labels, on the submitted structure, standardized structure, or canonical tautomer, but there are many more possible seed variants not used in PubChem.

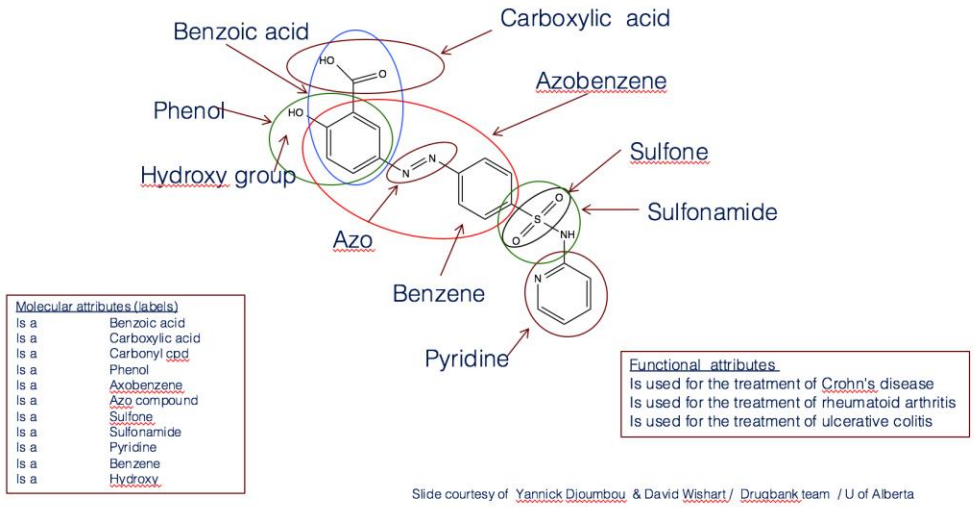
Hashcodes link structures to closely related compounds which agree at least in fragment connectivity. Wolf-Dietrich is exploring more advanced options, hashing structure relationships relevant to medicinal chemistry, for example, linking structures with similar ring systems and substituent fragments at sites of interest, and using various fragment and generalized hashes. He calls this PogoChem and a proof-of-concept is available. Users simply click on a structure and query results appear instantaneously.

In one option, ring system variants are produced by generalizing ring system atoms. There is one hashcode per ring system. Ring system size, and heteroatom count are stored for the similarity score. In another option ring systems or bridges are resized by excising unsubstituted atoms between substitution or fusion points, individually or in combination. This time there are 1-10 hashcodes per ring system. It is also possible to cut bonds, and compute a hash for the fragments. These are stored with bond information and basic fragment statistics. This leads to about 50 topology-filtered hashcodes per compound. Storing 5 billion records, at 56 bytes per record is no problem.

Wolf-Dietrich concluded by saying that PubChem is a great resource, in the hands of a capable team. It is still evolving at a fast pace, and it continues to inspire new ideas of how to access and analyze its contents.

PubChem a resource for cognitive computing

Stephen Boyer of the IBM Almaden Research Center has collaborated with OntoChem, the University of Alberta, NIH, EMBL-EBI, and others on a chemical ontology approach to addressing drug discovery. Their work with chemical ontologies identifies a family of molecular attributes that define a molecule and explores how those attributes might be used for identifying functional attributes based on molecules with similar structure activity. An example of their use of molecular attributes can be seen below, illustrated by assignments within the target molecule (Azulfidine) of benzoic acid, carboxylic acid, carbonyl compound, phenol, azobenzene, azo compound, sulfone, sulfonamide, pyridine, benzene, and hydroxyl groups:



In this example of Azulfidine, assignments are also made for functional attributes, for example, “it is used for” the treatment of Crohn’s disease, rheumatoid arthritis and ulcerative colitis.

The process begins by converting a compound name to SMILES. From the SMILES, molecular attributes (also known as molecular descriptors or chemical labels) such as “hydroxy” or “benzoic” or “phenyl” are generated. Steve’s team submitted about 1.4 million SMILES strings from ChEMBL to two different auto-classification systems to make a ChEMBL ontology database with two computer-generated chemical ontologies: ClassyFire (written by David Wishart of the University of Alberta and Ph.D. student Yannick Djoumbou Feunang) and OntoChem (Lutz Weber).

Steve then used this database in a multi-step process. He queried it for a gene or target of interest (“XYZ”); created a set of candidate compounds with reported activity for XYZ; refined the candidate set to create a training set of compounds (e.g., with EC_{50} <30); scored and ranked the molecular attributes; and then used those results to query the ChEMBL database minus the candidate set and the training set. He thus identified 100 compounds with potential activity, exclusive of the candidate or training sets.

Steve reported two experiments. The first concerned MDM2 (mouse double minute 2 homologue), a protein that in humans is encoded by the MDM2 gene. The key target of MDM2 is the p53 tumor suppressor. Steve carried out a sample analysis, using the two chemical ontologies, to predict compounds that may have MDM2 activity, scored with a chi-squared test. In ChEMBL, 20,558 molecules have activity for MDM2, but only 27 of these have IC_{50} < 30 nM. He compared the top 100 compounds identified by ClassyFire with the top 100 compounds identified by OntoChem, generated with the parameters of the top 10 labels, assay minimum = 30, and corpus count cut off = 300,000. He found 57 predicted compounds in common between the two ontologies. Not having a laboratory, he was unable to test any of these compounds, but he did find structure activity data in numerous patents that had 26 compounds with reported assay data for MDM2, and some of them matched compounds in his set of 57 potential actives.

Steve’s second example concerned SGLT2 (sodium/glucose cotransporter 2) inhibitors that reduce blood glucose levels and have potential use in the treatment of type II diabetes. Thirty compounds with assay data for SGLT2 were derived from the ChEMBL database, but only 12 had EC_{50} < 10 nM. Using these 12 molecules as a training set, the team identified several new molecules as possibly having SGLT2 activity. A search of patents and the scientific literature confirmed that several of the identified compounds had reported significant activity as SGLT2 inhibitors.

Steve closed with some final thoughts on innovation. Steven Johnson[25] coined the term “hummingbird effect” to describe how an innovation in one field ends up triggering changes that seem to belong to a different domain altogether. Innovations arise from the “adjacent possible” (a term Johnson borrows from the theoretical biologist Stuart Kauffman): you get railroads when it is railroading time, and not before, even if some prescient inventor sketches them out far in advance, and they open up all kinds of new possibilities.

SPL and openFDA resources of open substance data

Yulia Borodina is in the Office of Health Informatics at the U.S. Food and Drug Administration (FDA/OHI). Her talk concerned "bulk" open data. Machine-readable data are extracted from text or legacy databases, harmonized, and coded in a machine readable format. To provide data interoperability you need a data standard, and then you harmonize the data according to the standard, and ensure that the standard is publicly available (and, ideally, freely available). Unfortunately, you may have to wait 50 years until the community adopts the standard. To support data reuse you can provide direct downloads and APIs, and let the user decide how to select and analyze the data.

Structured Product Labeling (SPL) is a document markup standard approved by Health Level Seven (HL7) and adopted by FDA as a mechanism for exchanging product and facility information. It covers health informatics, cheminformatics, and bioinformatics. It has many applications: Yulia concentrated on substances. SPL is a universal (not data-specific) exchange standard, with reusable data types, coded data elements, and data-specific validation procedures. Drug manufacturers and distributors submit SPL to FDA, and FDA makes a product SPL file with substance, pharm class, billing unit, and product concept index files. Data are output to the FDA Online Label Repository, the National Library of Medicine's DailyMed website, and the public data warehouse, openFDA.

Substances in products can be small molecules, proteins, nucleic acids, polymers, organisms, parts of organisms, or mixtures. Definitions of non-confidential substances from the FDA Substance Registration System are available in SPL format, with unique ingredient identifiers (UNII). The data for over 50,000 chemical substances, and over 5,000 biological ones, are compliant with the Identification of Medicinal Products (ISO IDMP 11238) standard, and are available from DailyMed and openFDA. The IDMP standard defines "what" (e.g., proteins are to be defined by sequence) and the SPL standard defines "how" (e.g., UNII, molfile, InChI, and InChIKey for small molecules). Yulia showed the content of some SPL Substance Index Files for various types of substance. SPL data have been integrated into PubChem.

The concept of openFDA is to index high-value, high priority, and scalable public datasets (e.g., medical device reports, drug adverse events, and food recall enforcement reports), and to format and document the data in developer- and consumer-friendly standards, and make those data available via a public-access portal that enables developers to use them in applications quickly and easily. openFDA allows direct downloads and APIs. Substance and Pharm Class SPL index files can be downloaded, and some substance SPL fields associated with a product label are available in JavaScript Object Notation (JSON) format via API. openFDA allows users to carry out statistical applications around adverse events, such as the likelihood ratio test-based method for signal detection in drug classes. Interactive open-source applications available on https://open.fda.gov/analytics/ demonstrate how openFDA APIs can be used for epidemiological research, combined with powerful statistical tools built by the openFDA community.

Building a network of interoperable and independently produced linked and open biomedical data

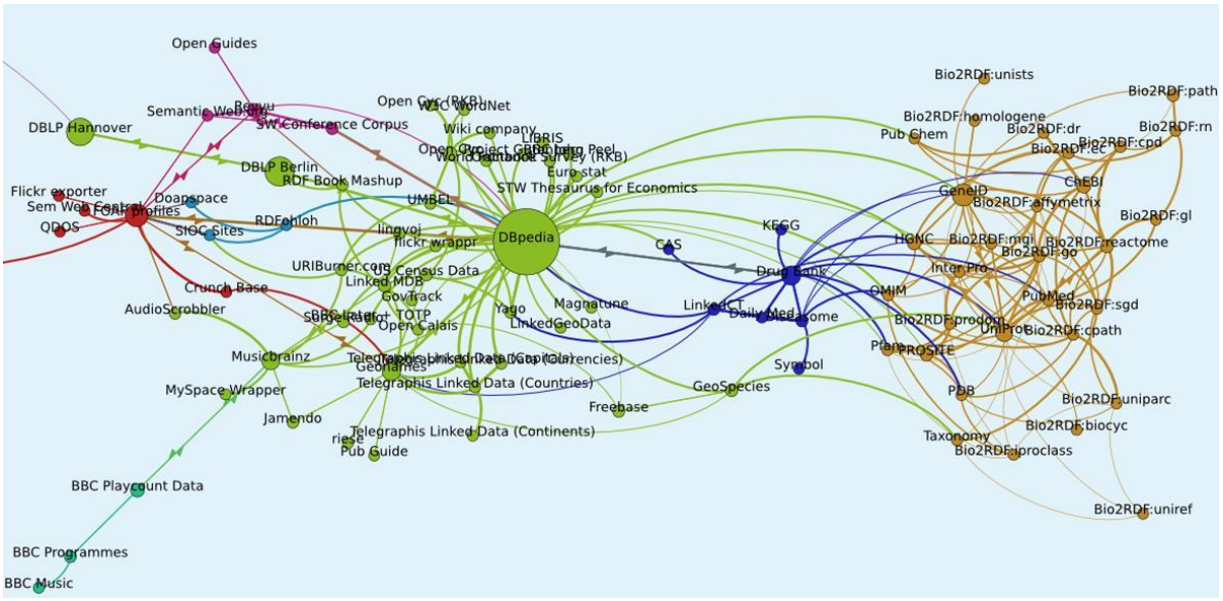
Michel Dumontier of Stanford University, and his co-workers, develop tools and methods to represent, store, publish, integrate, query, and reuse biomedical data, software, and ontologies, with an emphasis on reproducible discovery, which necessitates data science tools and methods, and community standards. Data need to be "FAIR",[26] that is, findable, accessible, interoperable, and reusable.

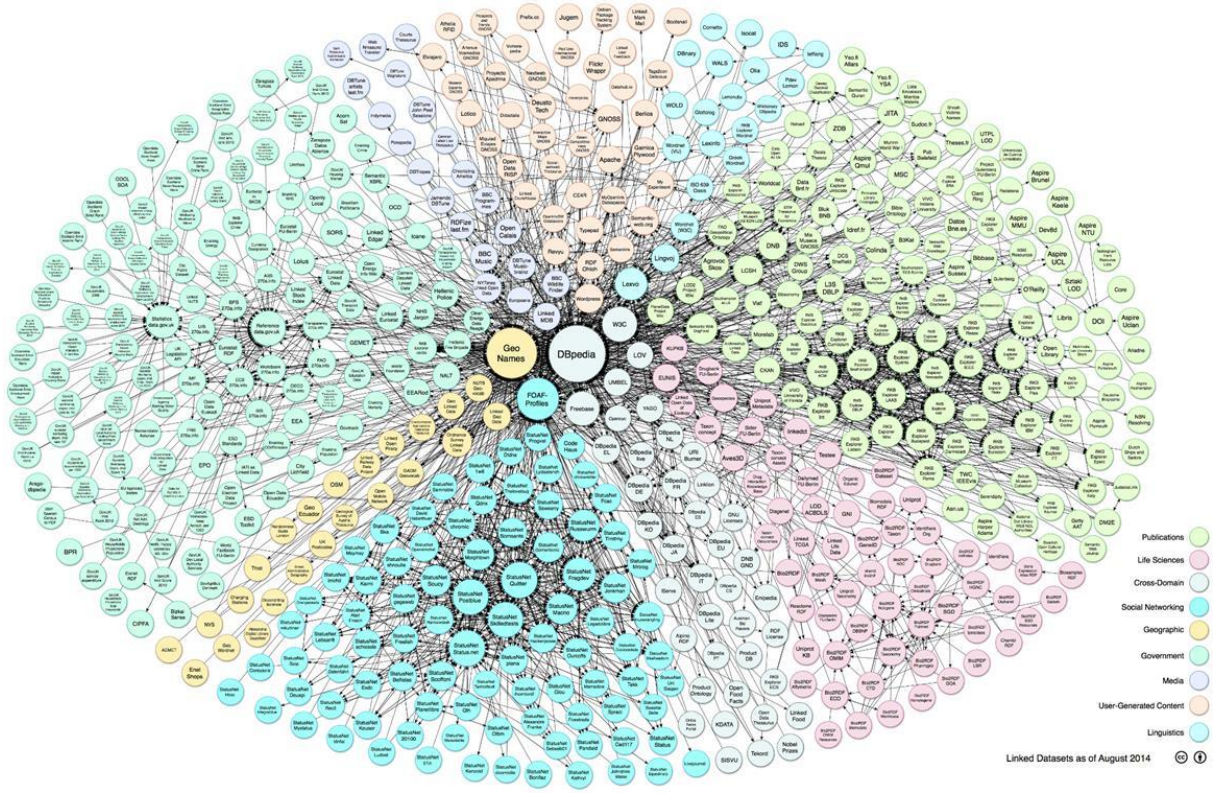
The Semantic Web is the new global web of knowledge: it has standards for publishing, sharing and querying facts, expert knowledge and services, and a scalable approach for the discovery of independently formulated and distributed knowledge. Linked Data offers a solid foundation for FAIR data: entities are identified using globally unique identifiers (URIs); entity descriptions are represented with a standardized language (resource description framework, RDF); data can be retrieved using a universal protocol (HTTP); and entities can be linked together to increase interoperability.

Bio2RDF is an open source project to unify the representation and interlinking of biological data using RDF: it transforms silos of life science data into a globally distributed network of linked data for biological knowledge discovery. It shows how datasets are connected together. Queries can be federated across private and public Protocol and RDF Query Language (SPARQL) databases. A graph-like representation is amenable to finding mismatches and discovering new links.[27] EbolaKB[28] is an example using linked data and software.

In current, unpublished research on network analysis and discovery, Michel's team is examining whether they can implement an open version of PREDICT[29] using linked data. HyQue,[30,31] for hypothesis validation, is a platform for knowledge discovery that uses data retrieval coupled with automated reasoning to validate scientific hypotheses. It builds on semantic technologies to provide access to linked data, ontologies, and Semantic Web services, uses positive and negative findings, captures provenance, and weighs evidence according to context. It has been used to find aging genes in nematodes, and to assess cardiotoxicity of tyrosine kinase inhibitors

The network of linked data goes beyond biology. Michel displayed a network from about 2007, and the linking open data cloud diagram as of August 2014, to show how rapid has been the expansion over domains:





EMBL-EBI have been producing RDF for two years, PubChemRDF was released more than two years ago, and NLM has released a beta version of Medical Subject Headings (MeSH) RDF linked data, but lack of coordination makes Linked Open Data chaotic and unwieldy. There is no shortage of vocabularies, ontologies and community-based standards. The National Center for Biomedical Ontology (NCBO) manages a repository of all publicly available biomedical ontologies and terminologies. The NCBO BioPortal resource makes these ontologies and terminologies available via a Web browser and Web

Services. The NCBO Annotator service takes as input natural-language text and returns as output ontology terms to which the text refers. The Center for Extended Data Annotation and Retrieval (CEDAR) project relies on the BioPortal ontology repository and the NCBO Annotator. CEDAR is making data submission smarter and faster, so biomedical researchers and analysts create and use better metadata. Through better interfaces, terminology, metadata practices, and analytics, CEDAR optimizes the metadata pathway from provider to end user.

PubChem engaged the community to reuse and extend existing vocabularies. Semanticscience Ontology (SIO) is an effective upper level ontology, with over 1,500 classes, and 207 object properties. Chemical Information Ontology (CHEMINF)[32] is a collaborative ontology that distinguishes algorithmic, or procedural information from declarative, or factual information, and renders of particular importance the annotation of provenance to calculated data.

Large scale publishing on the Web across biomedical datatypes is possible. Hubs such as NCBI and EMBL-EBI now integrate data, but there is need for global coordination on all data types. Standard vocabularies must to be open, freely accessible, and demonstrably reused. Worldwide data integration formats such as RDF can improve linking of data, and some toolkits that are easier to deploy will provide standards-compliant, linked data. The development and use of standards by PubChem, and others, brings us closer to an interoperability ideal, but much more work is needed to support computational discovery in a reproducible manner.

Chemical structure representation in PubChem

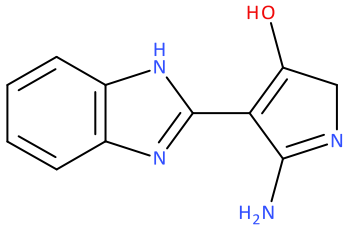
A unique and invaluable feature of the architecture of PubChem is the distinction between the deposited structures (substances) and the normalized structures (compounds), and the retention of both. This feature allowed PubChem to avoid the early mistakes of CAS, said Roger Sayle of NextMove Software. PubChem Substance contains about 209.6 million structures; PubChem Compound contains about 91.7 million structures. The PubChem standardization service aims to determine when two chemical structures are the same.

Consider, for example, implicit and explicit hydrogens. Ethanol (PubChem CID 702) has been deposited 1569 times with six different explicit atom counts, and thus, six different SIDs. All have the same SMILES and InChI. Nitrobenzene (PubChem CID 7416) has been deposited as 164 distinct substance depositions, with five SIDs, two with molecular formula C6H5NO2, and the others with extra hydrogens: $C_6H_6NO_2^+$, $C_6H_6NO_2^-$, and $C_6H_7NO_2$. To complicate matters, BIOVIA 2017 changed the interpretation of CTfiles (the default valences of some neutral main group elements have changed); this affects 342,689 SIDs and 213,097 CIDs. PubChem is inconsistent on protonation, but generally protonation state is preserved.

A major challenge in chemical databases is aromaticity; two compounds that differ in Kekulé forms are the same molecule. A significant novel innovation in cheminformatics was Evan Bolton's development of a "canonical" Kekulé SMILES form of a molecule. This enabled PubChem to avoid the early mistakes of Daylight Chemical Information Systems. Different chemistry toolkits (and chemists) differ in opinion on

which ring systems are aromatic and which are not, hence PubChem's wish to remain "neutral" by only providing non-aromatic SMILES. Unfortunately, Evan's algorithm aromatizes all conjugated cycles, and not just those associated with the smallest set of smallest rings, a computationally demanding requirement. PubChem does not restrict aromaticity to 4n+2 Hückel aromaticity; thus conjugated ring systems such as pentalene are deemed aromatic.

Tautomers are normalized. Thus 4-(phenylazo)-1-naphthalenol (CAS RN 3651-02-3), a case of classic tautomerism, has only one CID (5355205), but there are two InChIs, one for each tautomer. Unfortunately not all tautomers are handled so well: four tautomers of this molecule are recorded:



PubChem follows InChI in breaking bonds to metals. It currently handles 109 of the 118 elements in the periodic table. PubChem registration confirms that any specified isotope has been observed experimentally. Hence $^{7}CH4$ is rejected, but $^{8}CH4$ (which has an exceptionally short half- life) is allowed. Another quirk is that PubChem does not normalize mononuclidic isotopes. Hence fluoromethane has CID 11638, while fluoromethane with ^{19}F has CID 58338844. PubChem rejects chlorine dioxide, and carbide anions, but it accepts disulfur dioxide (O=S=S=O) which is stable for only a few seconds.

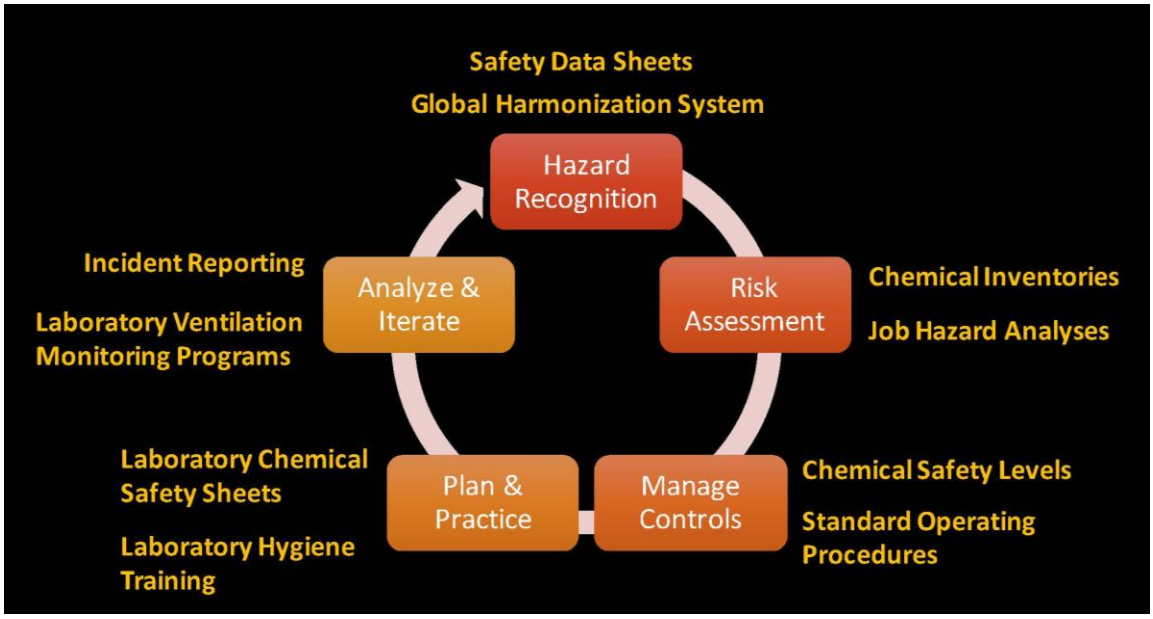
It is one of the innovations of PubChem that it explicitly stores relationships (such as having similar 3D shape) in the database. Given a CID, you can find all similar CIDs based on Tanimoto similarity, for example, but you can also find all the tautomeric forms provided by depositors by following the links from CID to SID. Likewise, there are internal links (backwards and forwards) between mixtures and their components, and between isotopes of a compound, and between enantiomers of a compound.

PubChem allows depositors to specify advanced representations of molecular structures such as inorganics and organometallics via SD tags. Quadruple, dative, complex and ionic bonds can be specified with the non-standard bond option; hydrogen, resonance, bold and Fischer bonds, and close contacts can be specified with the bond annotations option. Relatively few depositors make use of these options.

Roger concluded by saying that PubChem represents the current state-of-the-art in chemical structure representation.[33-35] Under the surface, unseen to most users, are many technical and scientific innovations that have enabled PubChem to scale to contain nearly 100 million compounds. From simple design decisions such as the substance versus compound distinction, to breakthroughs such as canonical Kekulé SMILEs, the architecture of PubChem contains a treasure trove of cheminformatics innovations, covering normalization, tautomers, mixtures, 2D fingerprints and similarity, substructure search, biopolymers, text mining, and much more.

iRAMP and PubChem: of the people, for the people

Leah McEwen of Cornell University gave a talk on synergies between chemical safety and information literacy skills. In 2015, the ACS Committee on Professional Training (CPT) released an updated version of Undergraduate Professional Education in Chemistry: ACS Guidelines and Evaluation Procedures for Bachelor's Degree Programs. These guidelines include a description of six skill sets that undergraduate chemistry majors should develop, two of them being chemical literature and information management skills, and laboratory safety skills. Laboratory safety skills can be viewed as a specific "use case" of information literacy skills.[36] The CPT safety guidelines describe a RAMP model[37] to organize safety information in a consistent way that is transferable, scalable and sustainable as laboratory work evolves. RAMP is an acronym for the initial letters of the four core principles of safety: Recognizing hazards, Assessing risks of hazards, Minimizing hazards, and Preparing for emergencies. The iRAMP project was begun in 2014 by the ACS Divisions of Chemical Information (CIF), and Chemical Health and Safety (CHAS).[36,38] The "i" of iRAMP signifies the iterative nature of the chemical safety decision cycle:



The research laboratory environment is complex, involving chemicals, biological agents, and radioactive materials, with five levels of Occupational Safety and Health Administration (OSHA) controls. The information environment is also very complex. Questions that safety professionals need to ask have been listed by the ACS Committee on Chemical Safety, Safety Advisory Panel. Data supporting chemical risk assessment are detailed in a National Research Council (NRC) work[39], but there are many challenges for the information community. Many chemicals lack critical data. The diversity of substance forms that impact chemical reactivity is broad. Data are scattered across many sources. Reporting standards are variable and most data are not machine-readable.

The research practices described by the American Library Association, Association of College and Research Libraries Framework for Information Literacy for Higher Education reflect a process of iterative

critical inquiry that can be used to address these questions about the chemical information available to be used in risk assessment, and the most effective process for identifying, compiling, analyzing, and applying this.[36]

A PubChem Laboratory Chemical Safety Summary (LCSS) for a compound is based on the format described by the National Research Council.[39] LCSS provides a convenient consolidated view of an open Internet search on chemical hazard information, with non-authoritative sources filtered out and available documentation on the context of each data point. It became clear that PubChem could help chemists fill out an NRC safety form. The University of California has produced a pilot mobile app, UC Chemicals, a cloud-based chemical inventory management tool, which allows tracking of containers using a barcoding system. Chemical and safety information, such as hazard codes and first aid, are automatically populated from PubChem and other sources.

There are, however, some key gaps that iRAMP must address. These include resolvable identifiers for mixtures; associating the Global Harmonization System (GHS) with supporting data (a sort of "Rule of Five" for hazards would be good to have, as most compounds have not been classified); mapping chemical concepts to process conditions; mapping procedures to chemical, equipment, and process hazards; and empirical data from incidents.

iRAMP aims to build a "flexibly structured ecosystem of data, workflow tools, and domain expertise, mapped to the essential commonalities of the use cases and content, connected by good information management practices".[38] PubChem enables reuse of data in applied contexts, based on open data, open mission, open process and open collaboration, for the public good. Together, iRAMP and PubChem can build an ecosystem, of the people, by the people, for the people.

Open chemical information: where now and how?

Evan Bolton gave the award address on behalf of both awardees. Many people think that cheminformatics is a solved problem. "Open" is now a popular adjective: open learning, open access, open data, open government, open source, and so on. "Open" was much less of an "in" word when PubChem was conceived. There is still little openness when it comes to scientific data. There is still a lot to be done in the open space. For example, openness is not widespread in drug discovery. We have to empower researchers with ready access to information so that they do not repeat work that has already been done.

PubChem is an open archive; the data are free, accessible, and downloadable. Information is uploaded by depositors, it is normalized and displayed, and it can then be downloaded by other researchers. Algorithms carry out the normalization, but sometimes they go wrong and can introduce ambiguity; later processing of this ambiguous data can result in data corruption or error. For example, chemical file format interconversion can be "lossy", such as when converting from SDF to SMILES, where the coordinates are lost and stereo must be perceived by algorithms. Different software packages may "normalize" or convert a chemical structure in different ways. This variation produces tens of different representations of nitro groups and azides in PubChem.

Atom environments have to be standardized. Data clean-up approaches include structure standardization; consistency filtering (name-structure matching, and use of authoritative sources, and hand curated black, gray, and white lists); chemical concepts (groupings of chemical names, setting a preferred concept for a given structure, and a preferred structure for a given concept); and cross-validation via text mining (to gather evidence to support the reported association of a chemical to other entities). A chemical structure may be represented in many different ways (tautomer and salt-form drawing variations are common, for example), and the chemical meaning of a substance may change with context (e.g., the solid form may involve a hydrate, which affects molecular weight when weighing out a substance to make a solution). The boiling point of benzene is both 176.2 ^{o}F and 200-500 ^{o}F in PubChem Compound; the first record is that for benzene, but the second is for coal tar oil (a crude form of benzene). There are many-to-many relationships between chemical concepts and chemical structures.

PubChem is successful because it is inclusive, free, robust, innovative, and helpful. If a chemical exists, you often find it. Evan singled out a few features of PubChem for particular mention. Substances are converted to compounds, but the original information is kept. There is clear provenance, so users can trace from whom the data came. Information is downloadable, and there are extensive programmatic interfaces. PubChem is constantly improved, can handle a lot of abuse, and is sustainable. The PubChem synonym classification was available first in RDF. It indicates the chemical name type, allows grouping of names, and can involve guess work. More authoritative name sources have been added. Most non-classified names are unhelpful (perhaps because of chemical name corruption, or chemical name fragments).

As more data are added, the scalability of PubChem is difficult to maintain. It is not uncommon to reach the limit of technology. For example, PubChem could no longer use SQL databases for some queries due to performance bottlenecks. After examination of noSQL technologies like Solr/Lucene, better approaches were determined. An example of this is PubChem's structured data query (SDQ), which uses the Sphinx search engine to perform the query, but then fetches data from an SQL database. It is a query language with clear logic in concise format, communicating with a JSON object. It features a powerful search ability, a URL-accessible Common Gateway Interface (CGI), and easy application integration.

PubChem faces many challenges. One is growth: 50% of the resources of the project are needed just to keep scaling the system. Government mandates (like the current HTTPS-only edict) necessitate regular migrations. Data clean-up, and error proliferation prevention require constant vigilance: the team uses existing technology where possible, but solutions do not always exist. They must be developed for PubChem to remain scalable.

Chemical structure databases have come a long way since the origins of computerization in the 1960s, and the rise of databases such as CAS REGISTRY and Beilstein in the 1970s. The 2010s are the era of large, open chemical databases of aggregated content, with RESTful programmatic access. These large open collections of tens of millions of chemical structures need methods to lock down the data without curation, otherwise non-curation combined with open exchange of data leads to error proliferation. Digital standards are needed to improve chemical data exchange and chemical data clean-up methods

to prevent error proliferation. Close attention to provenance, and a set of clear definitions for chemical concepts are also needed.

ACS CINF had a data summit at the spring 2016 meeting in San Diego. Ten half-day symposia were held over five days, with over 70 speakers, including experts from different related domains. The summit helped to identify informatics "pain points" for which we need to find solutions. The Research Data Alliance and IUPAC had a follow-up workshop in July at EPA, where a number of projects were discussed. One on chemical structure standardization education and outreach aims to help chemists and other stakeholders to understand the issues of chemical structure standardization. Another, updating IUPAC's graphical representation guidelines, seeks to help chemists to understand the issues of chemical structure standardization, often apparent in chemical depiction. Other recommendations concern open chemical structure file formats, and best practices in normalizing chemical structures. There are plans to develop a small-scale ontology of chemical terms, based on terms in the IUPAC Orange Book as a case study. A project on the IUPAC Gold Book data structure is related to a current effort to extract the content and term identifiers, and convert them into a more accessible and machine-digestible format for increased usability. Finally, a scoping project on use cases for semantic chemical terminology applications will focus on researching the current chemical data transfer and communication landscape for potential applications of semantic terminology.

We are entering a new era: in the 2020s we will have large, extensively machine-curated, open collections, with clear provenance, and standard approaches to file formats and normalization, where errors do not proliferate, and links are cross-validated. Open knowledge bases will emerge that contain all open scientific knowledge that is computable (i.e., inferences can be drawn using natural language questions). By the 2030s machine-based inference will drive the majority of scientific questions, and efficiency of research will grow exponentially by harnessing "full" scientific knowledge.

In all, accurate computer interpretation of scientific information content is paramount. It needs to be at or above the level of the human scientist for this vision of the future to occur. It will be the great achievement of our generation to make this leap forward. Improved chemical information standards and uniform approaches will be critical for it to occur.

Conclusion

After the award address, Rachelle Bienstock, chair of the ACS Division of Chemical Information, formally presented the Herman Skolnik Award to Evan and Steve:

L to R: Rachelle Bienstock, Evan Bolton, Steve Bryant

References

(1) Heller, S. R.; Fales, H. M.; Milne, G. W. A.; Heller, R. S.; McCormick, A.; Maxwell, D. C. Mass spectral search system. *Biomed. Mass Spectrom.* **1974,** *1* (3), 207-8.

(2) Heller, S. R.; Feldmann, R. J.; Fales, H. M.; Milne, G. W. A. Conversational mass spectral search system. IV. Evolution of a system for the retrieval of mass spectral information. *J. Chem. Doc.* **1973,** *13* (3), 130-3.

(3) Heller, S. R. The chemical information system and spectral databases. *J. Chem. Inf. Comput. Sci.* **1985,** *25* (3), 224-31.

(4) Heller, S. R.; Milne, G. W. A.; Feldmann, R. J. Quality control of chemical data bases. *J. Chem. Inf. Comput. Sci.* **1976,** *16* (4), 232-3.

(5) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminf.* **2015,** *7*, 1-63.

(6) Warr, W. A. Many InChIs and quite some feat. *J. Comput.-Aided Mol. Des.* **2015,** *29* (8), 681-694.

(7) Endicott, K. M. The National Chemotherapy Program *J. Chron. Dis.* **1958,** *8* (1), 171.

(8) Austin, C. P.; Brady, L. S.; Insel, T. R.; Collins, F. S. NIH Molecular Libraries Initiative. *Science* **2004,** *306* (5699), 1138.

(9) Zenkevich, I. G.; Babushok, V. I.; Linstrom, P. J.; White V, E.; Stein, S. E. Application of histograms in evaluation of large collections of gas chromatographic retention indices. *J. Chromatogr. A* **2009,** *1216* (38), 6651-6661.

(10) Milne, G. W. A.; Miller, J. A. The NCI Drug Information System. 1. System overview. *J. Chem. Inf. Comput. Sci.* **1986,** *26* (4), 154-9.

(11) Poroikov, V. V.; Filimonov, D. A.; Ihlenfeldt, W.-D.; Gloriozova, T. A.; Lagunin, A. A.; Borodina, Y. V.; Stepanchikova, A. V.; Nicklaus, M. C. PASS Biological Activity Spectrum Predictions in the Enhanced Open NCI Database Browser. *J. Chem. Inf. Comput. Sci.* **2003,** *43* (1), 228-236.

(12) Steinbeck, C. LUCY - A program for structure elucidation from NMR correlation experiments. *Angew. Chem., Int. Ed. Engl.* **1996,** *35* (17), 1984-1986.

(13) Steinbeck, C. SENECA: A platform-independent, distributed, and parallel system for computer-assisted structure elucidation in organic chemistry. *J. Chem. Inf. Comput. Sci.* **2001,** *41* (6), 1500-1507.

(14) Guha, R.; Howard, M. T.; Hutchison, G. R.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J.; Willighagen, E. L. The Blue Obelisk - Interoperability in Chemical Informatics. *J. Chem. Inf. Model.* **2006,** *46* (3), 991-998.

(15) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003,** *43* (2), 493-500.

(16) Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. L. Recent developments of the Chemistry Development Kit (CDK) - an open-source Java library for chemo- and bioinformatics. *Curr. Pharm. Des.* **2006,** *12* (17), 2111-2120.

(17) Steinbeck, C.; Krause, S.; Kuhn, S. NMRShiftDB - constructing a free chemical information system with open-source components. *J. Chem. Inf. Comput. Sci.* **2003,** *43* (6), 1733-1739.

(18) Steinbeck, C.; Kuhn, S. NMRShiftDB – compound identification and structure elucidation support through a free community-built web database. *Phytochemistry* **2004,** *65* (19), 2711-2717.

(19) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012,** *40* (D1), D1100-D1107.

(20) Edison, A. S.; Hall, R. D.; Junot, C.; Karp, P. D.; Kurland, I. J.; Mistrik, R.; Reed, L. K.; Saito, K.; Salek, R. M.; Steinbeck, C.; Sumner, L. W.; Viant, M. R. The time is right to focus on model organism metabolomes. *Metabolites* **2016,** *6* (1), 8/1-8/7.

(21) Haug, K.; Salek, R. M.; Conesa, P.; Hastings, J.; de Matos, P.; Rijnbeek, M.; Mahendraker, T.; Williams, M.; Neumann, S.; Rocca-Serra, P.; Maguire, E.; Gonzalez-Beltran, A.; Sansone, S.-A.; Griffin, J. L.; Steinbeck, C. MetaboLights-an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* **2013,** *41* (D1), D781-D786.

(22) Kale, N. S.; Haug, K.; Conesa, P.; Jayseelan, K.; Moreno, P.; Nainala, V. C.; Spicer, R. A.; Williams, M.; Salek, R. M.; Steinbeck, C.; Rocca-Serra, P.; Li, X.; Griffin, J. L. MetaboLights: An Open-Access Database Repository for Metabolomics Data. *Curr. Protoc. Bioinformatics* **2016,** *53*, 14.13.1-18.

(23) Salek, R. M.; Haug, K.; Conesa, P.; Hastings, J.; Williams, M.; Mahendraker, T.; Maguire, E.; Gonzalez-Beltran, A. N.; Rocca-Serra, P.; Sansone, S.-A.; Steinbeck, C. The MetaboLights repository: curation challenges in metabolomics. *Database* **2013,** *2013*, bat029.

(24) Beisken, S.; Earll, M.; Baxter, C.; Portwood, D.; Ament, Z.; Kende, A.; Hodgman, C.; Seymour, G.; Smith, R.; Fraser, P.; Seymour, M.; Salek, R. M.; Steinbeck, C. Metabolic differences in ripening of Solanum lycopersicum 'Ailsa Craig' and three monogenic mutants. *Sci. Data* **2014,** *1*, 140029.

(25) Johnson, S. *How We Got to Now. Six Innovations That Made the Modern World*; Riverhead Books: New York, NY, 2014.

(26) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J. J.; Appleton, G.; Dumon, O.; Groth, P.; Strawn, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da, S. S. L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Rocca-Serra, P.; Sansone, S.-A.; Gray, A. J. G.; Goble, C.; Grethe, J. S.; Heringa, J.; Kok, R.; t, H. P. A. C.; Hooft, R.; Kuhn, T.; Kok, J.; Lusher, S. J.; Mons, B.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Roos, M.; Thompson, M.; van, S. R.; Schultes, E.; Sengstag, T.; Slater, T.; Swertz, M. A.; van, d. L. J.; van, M. E.; Mons, B.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016,** *3*, 160018.

(27) Hu, W.; Qiu, H.; Dumontier, M. Link Analysis of Life Science Linked Data. In *The Semantic Web - ISWC 2015: 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part II;* Arenas, M.; Corcho, O.; Simperl, E.; Strohmaier, M.; d'Aquin, M.; Srinivas, K.; Groth, P.; Dumontier, M.; Heflin, J.; Thirunarayan, K., Staab, S., Eds.; Springer International Publishing: Cham, 2015; pp 446-462.

(28) Kamdar, M. R.; Dumontier, M. An Ebola virus-centered knowledge base. *Database* **2015,** *2015*, bav049.

(29) Gottlieb, A.; Stein, G. Y.; Ruppin, E.; Sharan, R. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.* **2011,** *7*, 496-504.

(30) Callahan, A.; Dumontier, M. Evaluating Scientific Hypotheses Using the SPARQL Inferencing Notation. In *The Semantic Web: Research and Applications: 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012. Proceedings;* Simperl, E.; Cimiano, P.; Polleres, A.; Corcho, O., Presutti, V., Eds.; Springer: Berlin, Heidelberg, 2012; pp 647-658.

(31) Callahan, A.; Dumontier, M.; Shah, N. H. HyQue: evaluating hypotheses using Semantic Web technologies. *J. Biomed. Semantics* **2011,** *2* (2), 1-17.

(32) Hastings, J.; Chepelev, L.; Willighagen, E.; Adams, N.; Steinbeck, C.; Dumontier, M. The Chemical Information Ontology: provenance and disambiguation for chemical data on the biological semantic web. *PLoS One* **2011,** *6* (10), e25513.

(33) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. PubChem: integrated platform of small molecules and biological activities. In *Annual Reports in Computational Chemistry;* Wheeler, R. A., Spellmeyer, D. C., Eds.; Elsevier: Amsterdam, 2008; Vol. 4, pp 217-241.

(34) Hahnke, V. D.; Bolton, E. E.; Bryant, S. H. PubChem atom environments. *J. Cheminf.* **2015,** *7*, 41/1-41/37.

(35) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound databases. *Nucleic Acids Res.* **2016,** *44* (D1), D1202-13.

(36) Stuart, R. B.; McEwen, L. R. The Safety "Use Case": Co-Developing Chemical Information Management and Laboratory Safety Skills. *J. Chem. Educ.* **2016,** *93* (3), 516-526.

(37) Hill, R. H.; Finster, D. C. *Laboratory Safety for Chemistry Students*; Wiley: Hoboken, NJ, 2010.

(38) McEwen, L. R.; Stuart, R. B. Meeting the Google Expectation for Chemical Safety Information. Chemical Risk Assessment in Academic Research and Teaching. *Chem. Int.* **2015,** *37* (5-6), 12-16.

(39) National Research Council *Prudent Practices in the Laboratory*; National Academies Press: Washington, DC, 2011.