Herman Skolnik Award Symposium Honoring Henry Rzepa and Peter Murray-Rust
ACS National Meeting, Philadelphia, PA, August 21, 2012

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

Introduction

This one-day symposium was remarkable for its record number of speakers (23 in all, plus one withdrawn and one replaced by a demonstration). Despite the number of performers, and some unfortunate technical faults, the whole event proceeded on schedule and without serious mishap. Henry Rzepa's own talk was an opening scene-setter. He told a 1992 tale of some molecular orbitals explaining the course of a chemical reaction in 1992. The color diagram of these lacked semantics, and when it had been sent by fax to Bangor, it even lost its color. Months later the work was published,[1] but the supporting information (SI) is not available for this article, and even if it were available electronically, would it be usable? So, how can it be mined for useful data or used as the starting point for further investigation?

By 1994 Henry and his colleagues had recognized the opportunities presented by the World Wide Web.[2,3] The data for a later article[4] do survive in the form of Quicktime and MPEG animations on the Imperial College Gopher+ server but they are semantically poor, i.e., they are interpretable by humans but not by computer. The X-ray crystallography data are locatable using the proprietary identifier HEHXIB allocated by the Cambridge Crystallographic Data Center. Open identifiers such as the IUPAC International Chemical Identifier, InChI, are preferable. It would be better if we had access to semantically rich data that allows reanalysis of the key intermolecular interactions (described in Henry's blog entry of July 5, 2012, http://www.ch.imperial.ac.uk/rzepa/blog/?p=7027). The answer is a hand-crafted XML document with the SI as a "datument": a superset of the main article.[5] Molecules and spectra are expressed in Chemical Markup Language (CML)[6] and presented using a Java applet and scalable vector graphics (SVG). The underlying data for the article are still semantically alive today.

More recently, Henry has used electronic SI as a data repository for the main article.[7] The molecules are expressed in CML and a Jmol applet is used as the presentation layer in the style of an explorable story-board. Quick-response (QR) access to the data in 2011 allowed a re-investigation, with revised conclusions. Datasets should be deposited in digital repositories,[8] using CML where possible, and assigned a handle (equivalent to a digital object identifier, DOI). Metadata can be generated from automated scripts and can be harvested for re-injection into other repositories. In Peter Murray-Rust's Chempound (https://journals.tdl.org/jodi/article/view/5873/5879), the Resource Description Framework (RDF) allows SPARQL (http://www.w3.org/TR/rdf-sparql-query/) semantic queries of data. The repository figshare (http://www.figshare.com) allows users to upload any file format so that figures, datasets and media can be disseminated in a way that the current scholarly publishing model does not allow. Most journals treat such data-rich objects as "gold" Open Access but there are not yet many articles with such data and you may not have permission to mine them, or even know how to find them. Perhaps gold data need their own DOIs in figshare, SPECTRa[8] etc.

Steve Bachrach's Computational Organic Chemistry blog (http://comporgchem.com/blog/) is data-rich, discussion-rich, and archivable. In other work, device-agnostic HTML5 components have been rendered natively in a browser or the epub3 Reader (the new shrink-wrapper), enabling a mobile ecosystem. Talks later in the symposium enlarged on the topics introduced by Henry.

Visualization

The first invited talk was by Bob Hanson of St. Olaf College who described two open source Java applets, Jmol and JSpecView that are used for interactive access to molecules and spectra. Jmol is a viewer for chemical structures in 3D. JSpecView, a viewer for spectral data in the JCAMP-DX format, reads a variety of spectral data types, and has recently been integrated into Jmol. Bob also discussed a proposal for a JCAMP file extension, JCAMP-MOL (http://chemapps.stolaf.edu/jmol/docs/misc/Jmol-JSpecView-specs.pdf), that allows Jmol and JSpecView to read molecular structures, spectra and associated correlation data all from the same file. Two new user-defined data labels add 3D Jmol-readable models to the file and also associate spectral bands with specific IR and Raman vibrations, MS fragments, and NMR signals. The purpose of JCAMP-MOL is to allow for a single file that can be read either by the standalone Jmol application or by twin Jmol and JSpecView applets on a Web page. Clicking on an atom or selecting an IR/Raman vibration in Jmol highlights a band or peak or fragment on the spectrum. Clicking on the spectrum highlights one or more atoms, starts an IR vibration, or displays an MS fragment in Jmol. The specification was implemented successfully in Jmol 12.2.18 early in 2012.

The next speaker was Josef Polak of iChemLabs, the company that produces the ChemDoodle chemical structure environment (http://www.ichemlabs.com/products) focusing on 2D graphics and publishing (a product which, incidentally, was used to create all of the posters, pamphlets and conference books at this ACS National Meeting). Josef described how HTML5 adds new functionality in the browser. Java applets and third-party plug-ins such as Flash are being replaced by HTML5 and WebGL, not least in the open source ChemDoodle Web Components, a Javascript chemical graphics and cheminformatics library allowing users to present publication quality 2D and 3D graphics and animations for chemical structures, reactions and spectra. Beyond graphics, this tool provides a framework for user interaction to create dynamic applications through Web browsers, desktop platforms and mobile devices such as the iPhone, iPad and Android devices. The power of mobile technologies was well demonstrated in Josef's presentation when both projectors failed simultaneously: Josef continued, unfazed, while Kevin Theisen of iChemLabs walked around the room showing the slides on his iPad. The ChemDoodle Web Components library is being used by Henry Rzepa in datuments,[5] in the user interface to Jmol, Open Babel (http://openbabel.org/wiki/Main_Page) and ChemSpotlight (http://chemspotlight.openmolecules.net/), and in various educational applications.

Authoring and ELNs

Alex Wade of Microsoft Research talked about the Chemistry Add-in for Word, "Chem4Word" (http://research.microsoft.com/en-us/projects/chem4word/), a joint initiative of Microsoft Research and the University of Cambridge, the goals of which are to simplify the task of authoring a chemical document and to do so in such a way that the document is semantically meaningful, facilitating

downstream tasks such as publisher's workflow, entity extraction and semantic applications. Chem4Word is an open source tool that chemically enables Word, allowing direct search of structural repositories and insertion of structures directly into documents. Structures can be locally manipulated within Word and are stored in CML format. Alex explained the nature of Office Open XML files, and demonstrated the chemical editing and re-use cycle: loading structures into Word, from a gallery in Chem4Word itself or from PubChem (http://pubchem.ncbi.nlm.nih.gov/), editing structures, getting CML data back out of a document, and using and sharing the data in Chemistry for SharePoint.

A talk by Jeremy Frey of the University of Southampton also concerned the sharing of data. His team's first approach to the semantic electronic laboratory notebook (ELN) was the Smart Tea project,[9] so-called because, in order to gain a better understanding of the chemist's experimental design and execution process, the team made tea as a chemistry experiment. This early work, at the start of the e-science revolution, pushed the boundaries of the use of RDF, schemas and ontologies. "More Tea" used a tablet interface and RDF World but these hardware and software technologies still did not have the necessary power. LabTrove (http://www.labtrove.org/) is a more flexible ELN and data management system facilitating the capture of information and the use of this information in a collaborative environment. Jeremy's team has implemented a system ("Blogjects") to "blog" information from instruments: the Smart Research Framework (SRF) LabBroker middleware gets the data into the trove before the users even look. "Tweetjects" is another option. The ELN pages can now be read by both humans and computers, using XHTML (http://www.w3.org/TR/xhtml1/) and (RDFa http://www.w3.org/TR/xhtml-rdfa-primer/). Barcodes can be incorporated too, and LabTrove can be linked to SharePoint, using RSS, Atom, and the Open Data protocol (OData http://www.odata.org/).

The difference between Jeremy's system and other approaches is that the data are associated with the proposed scientific endeavor prior to or at the point of creation rather than by annotating the data with commentary after the experiment has taken place. This means that scientists and their peers can recreate and adapt the experiment repeatedly having already automated the processes and instrument settings. Prospective provenance describes a scientific experiment that *will be enacted*; retrospective provenance describes the scientific experiment that *was enacted*. Recording provenance allows the experiment itself to be embedded within the literature.

One weakness of the current system is the lack of support for existing external vocabularies and data models. Blog^3 (and TeaTrove^3) will have even greater user focus and semantic rigor. Blog³ provides an extensible plug-in architecture that enables authentication and authorization; in-line preview and search-engine indexing for all data; an integrated vocabulary and schema-editing environment; and export of all data in a variety of formats.

Simon Coles, also of Southampton University, continued the theme, talking about the ELN in academia. The Dial-a-Molecule Grand Challenge (https://connect.innovateuk.org/web/dial-a-molecule1/) addresses the problem of efficiently making molecules in days, not years. ELNs could be a response to this challenge. Other drivers are information overload, and government and funding agency initiatives to encourage researchers to share data openly. Repositories such as Dryad (http://datadryad.org/) and figshare (http://www.figshare.com) allow data to be published in their own right. Citation of data

through DataCite (http://www.datacite.org), for example, promises attribution and recognition for data publication.

An academic ELN should support a range of data acquisition techniques at different scales; promote access to data, sharing and reuse; enable discovery of results in related disciplines; facilitate access to data underpinning publications; enhance communication across the community; and support long-term preservation. ELNs currently on the market are primarily concerned with the protection of intellectual property and are very poor at supporting academic practice. The solution is to turn the ELN into a publishing platform in its own right with a protocol by which a range of existing platforms and resources can make the content available, based on simple, structured metadata. A number of repositories and alliances already exist and a number of people involved in them got together to produce a "lowest common denominator" solution, easy to implement on any platform, that can nevertheless be made more sophisticated at a later stage.

The multi-layered approach included a knowledge layer, with "core" metadata, an information layer, with "contextual" metadata and a computation layer with "detail" metadata. Through the knowledge layer, users can discover what is being made available, whether it is of interest, and whether it can be accessed. The information layer determines the granularity at which data should be made available and the computation layer determines whether the information can be processed automatically. Two case studies illustrate the entry point for layers two and three. One is LabTrove (described by Jeremy Frey earlier). The other is an extension to the IDBS e-Workbook plug-in that enables deposition of 2D structures directly into the Royal Society of Chemistry (RSC) database ChemSpider (http://www.chemspider.com). This could be extended to more content, such as spectra, reactions and properties. Simon's team is developing examples of automatic accessing and processing of data in ELNs layers two and three, and is encouraging wider academic use of ELNs. They will also mine theses and patents and investigate getting data out of old notebooks. The semantic ELN, Blog^3, described by Jeremy Frey, and "iPad in the Lab" are other works in progress.

Blogs

Continuing the blog theme, Steven Bachrach of Trinity University listed a number of examples. Peter Murray-Rust's blog (http://blogs.ch.cam.ac.uk/pmr/), Derek Lowe's *In The Pipeline* (http://pipeline.corante.com/), Paul Bracher's *ChemBark* (http://blog.chembark.com/) and *The Chemistry Blog* (http://www.chemistry-blog.com) provide opinion and news. Some blogs such as James Ashenhurst's *Master Organic Chemistry* http://masterorganicchemistry.com are for teaching. Paul Docherty's *Totally Synthetic* (http://totallysynthetic.com/blog/) and Steve Bachrach's own *Computational Organic Chemistry* http://comporgchem.com/blog/ publish article reviews. Henry Rzepa's blog (http://www.ch.ic.ac.uk/rzepa/blog/) features original research. Blog aggregators include Egon Willighagen and Peter Maas' *Chemical Blogspace* (http://cb.openmolecules.net/) and Jan Jensen's *Computational Chemistry Highlights* (http://www.compchemhighlights.org/).

Two recent examples illustrate post-publication peer review by blog. As a result, initially, of blogging in *Totally Synthetic*, a paper on reduction by sodium hydride[10] was withdrawn for scientific reasons; and a

paper with claims about dinosaurs in space[11] was criticized for self-plagiarism and exaggerated claims in several blogs before being withdrawn by the author on the grounds of similarity to his earlier publications. Steve himself has good reasons other than altruism for blogging. He surveys the literature to provide currency to his book and assist in writing the second edition. His blog also forms the basis of a series of review articles for the RSC and demonstrates the use of blogging in chemical communication. Blogging faces pressure from other social media but it is hard to envisage Twitter as an effective chemical communication medium. Altmetrics (an alternative to journal Impact Factors) and journal review overlay may establish a professional benefit to blogging in future.

Statistics and Property Prediction

Egon Willighagen at Maastricht University gave his presentation remotely. His take-home message was that you can improve your property prediction, training, and validation by adopting semantic pipelines.[12] This means using open look-up lists, dictionaries, and ontologies; removing format limitations; linking to data from other domains; and using calculation provenance. CML is semantic, flexible, and embeddable in HTML and RSS, but it is limited to XML. JavaScript Object Notation (JSON, http://www.json.org/) and Terse RDF Triple Language (Turtle, http://www.w3.org/TeamSubmission/turtle/) are alternative formats to XML for transmitting data between a server and a Web application. They enable linked data. RDF is an open standard, independent of format and database technology, and embeddable in HTML. It can be queried using SPARQL (http://www.w3.org/TR/rdf-sparql-query/). A federated query extension allows execution of queries distributed over different SPARQL endpoints.

One application is a computational toxicity assessment platform[13] generated from integration of two open science platforms related to toxicology: Bioclipse, which combines a scriptable, graphical workbench environment for integration of diverse sets of information sources, and OpenTox, a platform for interoperable toxicology data and computational services. A second application (unpublished) is Egon's work on nanotoxicity carried out in Stockholm last year, using SPARQL to link a wiki to the R statistics environment. Another project in progress is the Open Pharmacological Concepts Triple Store (Open PHACTS, http://www.openphacts.org),[14] a knowledge management project of the Innovative Medicines Initiative (IMI, http://www.imi.europa.eu/).

Rajarshi Guha of NIH discussed the benefits of integrating cheminformatics with statistical software, specifically the Chemistry Development Kit (CDK, (http://sourceforge.net/apps/mediawiki/cdk/index.php?title=Main_Page) and R. R is an environment for modeling that contains many prepackaged statistical and mathematical functions. It is also a matrix programming language that is good for statistical computing. Cheminformatics capabilities include statistics and machine learning and R is well suited to these. There is thus a case for "cheminformatics in R".

CDK provides chemical and more complex objects, input and output of various molecular file formats, fingerprint and fragment generation, rigid alignments, pharmacophore searching, substructure searching, SMARTS support, and molecular descriptors. Rajarshi has implemented CDK (http://github.com/rajarshi/cdk; http://sourceforge.net/projects/cdk/) in R using the rJava package,

providing access to variety of CDK classes and methods, and idiomatic R. Currently in rcdk you can access atoms and bonds and get certain properties and 2D and 3D coordinates, but since rcdk does not cover the whole CDK API you might need to drop down to rJava level, and make calls to the Java code, in some cases.

Rajarshi outlined some applications. The fingerprint package implements 28 similarity and dissimilarity metrics, allowing enrichment studies and comparison of datasets.[15] 2D structure images can be visualized. A typical QSAR workflow can be followed. The PubChem (http://pubchem.ncbi.nlm.nih.gov/) and ChEMBL (https://www.ebi.ac.uk/chembl/) databases can also be accessed directly within R using their public APIs. Published QSAR models may even become reusable: reproducible data mining is encouraged because DB and HTTP access ensures that an analysis can always be up to date if required.

Open Chemistry

In the final talk of the morning session, Marcus Hanwell of Kitware criticized the proliferation of black box, proprietary codes in chemistry. There is a need for open tools and open standards and more papers should be including data. The Open Chemistry project (http://www.openchemistry.org/) is a collection of open source, cross platform libraries and applications for the exploration, analysis and generation of chemical data. Kitware is developing three independent applications: the Avogadro[2] structure editor, Molequeue for running local and remote jobs, and ChemData for storing, annotating and searching data. Avogadro (http://avogadro.openmolecules.net/)[16] is an open source molecule editor and visualizer designed for cross-platform use in computational chemistry, molecular modeling, bioinformatics, materials science, and related areas. The Avogadro library is a framework providing a code library and application programming interface (API) with 3D visualization capabilities. The Avogadro application provides a rich graphical interface using dynamically loaded plug-ins through the library itself. The application and library can each be extended by implementing a plug-in module in C++ or Python. By using the CML file format as its native document type, Avogadro seeks to enhance the semantic accessibility of chemical data types. HDF5 (http://www.hdfgroup.org/HDF5/) will be used to store "heavy data" (e.g., for quantum mechanics).Kitware distributes its products using the very open Berkeley Software Distribution (BSD) license.

Artificially Intelligent Chemists

Peter Murray-Rust opened the afternoon session with some thoughts on building artificially intelligent chemists. He was helping to build a knowledge base for the Dial-a-Molecule Grand Challenge (https://connect.innovateuk.org/web/dial-a-molecule1/) but found that many publishers were unwilling to allow him to mine their content. There was interest in artificial intelligence (AI) in the 1970s but over the next 35 years little progress was made. Some early examples are Ralph Christoffersen's work on quantum pharmacology[17] and Malcolm Bersohn's work on retrosynthesis.[18] In those days knowledge bases depended on look-up, heuristics, rules, logic, brute force, tree pruning and computing chemical reality. Nowadays most of the tools we need are available but the will to use them is not there. Peter presented a diagram of the 2012 knowledge base, and perception and communication of the

transformed knowledge. Knowledge is represented in CML, ontologies and other domains. AI means putting all the components together.

Peter discussed a chemical application of John Searle's Chinese room thought experiment (http://en.wikipedia.org/wiki/Chinese_room). The experiment supposes that there is a program that gives a computer the ability to carry on an intelligent conversation in written Chinese. If the program is given to someone who speaks only English to execute the instructions of the program by hand, then in theory, the English speaker would also be able to carry on a conversation in written Chinese. However, the English speaker would not be able to understand the conversation. Here are Frog and Zog asking Magic Chemical Panda a chemical question and getting an answer:

There is no "Magic Chemical Panda" in Peter's box (http://vimeo.com/48280639). Chemical names are found by look-up and if the precise name is not found, the rule book is used to manipulate symbols and relate ethanoic to ethanoate, say. The Open Parser for Systematic IUPAC Nomenclature (OPSIN) name to structure software,[19] is a symbol manipulation system with a rule base.

Peter's team has also worked on CML and Chem4Word in the intelligent laboratory: Ami,[20] uses image recognition, voice recognition, sensors and RFID tags. Peter continues to capture semantics "by stealth" and he uses patents because publishers have prevented him from mining the journal literature. "Open" means *really* open and not pretending that your API is open. It is possible to make revenues from open source software: Kitware, ChemDoodle and GGA Software have proved this.

Computational Chemistry and NMR

Peter has been working with the Environmental Molecular Sciences Laboratory (EMSL) at Pacific Northwest National Laboratory (PNNL) on enriching the NWChem open source computational chemistry software (http://www.nwchem-sw.org/) with CML. Wibe ("Bert") de Jong was unable to present a talk

about this in person but Marcus Hanwell deputized. NWChem now generates semantic data, enabling Avogadro to extract and visualize NWChem semantic output. The team has completed a CML generator for Gaussian basis function based quantum methods based on the FOX library (http://fox-toolkit.org/), using an infrastructure based on PNNL's Extensible Computational Chemistry Environment data generator (http://blogs.ch.cam.ac.uk/pmr/2011/11/02/searchable-semantic-compchem-data-quixote-chempound-fox-and-jumbo/). Work currently in progress aims to get all NWChem data stored into CML output file, to reduce the CML data by avoiding replication, and to integrate CML with the appropriate format for bigger data blocks. Then plane wave capability will be made semantically rich. Another goal is to use Peter's JumboConverter to convert old NWChem output files into CML, and store them in MyEMSL. The CML CompChem dictionary and conventions are being extended to enable integration of NWChem and NMR data which can be accessed and visualized in MyEMSL through EMSLHub.

In another EMSL talk, Karl Mueller addressed the subject of NMR data. EMSL is collaborating with the Australian Commonwealth Scientific and Industrial Research Organization (CSIRO) in an NMR project. PNNL has about 12 very large NMR instruments but the data have not been captured well in the past. Karl gave one example of an experiment in which he was involved.[21] He showed diagrams of the workflows for translating and processing raw data from an experiment and for simulating and processing raw data from calculations in Gaussian, NWChem, etc. He also showed some screenshots from a potential MyEMSL Workbook for NMR. The team initially planned to continue updating the JCAMP NMR dictionary with relevant terms and definitions, to update the JCAMP parser, to test the output and to begin working on code to extract binary data for Agilent and Varian.

To make further progress the development of a repository for NMR data must address three important issues: the large number of different NMR experiments in existence, many with multiple versions and variations; the intricate processing steps often required to convert raw time domain data into usable spectra (and the need for a detailed record); and the large number of divergent NMR data formats. A proper record of an NMR experiment must contain original digitized numerical values, information about the source instrument, and saved instrument parameters, all in a standardized file format. The processed spectrum (as saved by the experimentalist) should include software version and processing parameters, in a standardized file format. The standardized file for a high-level experiment description should include sample, pulse sequence, magnetic field, detected isotope, decoupled and undetected isotopes, pulse times, delays, phase cycles, and temperature, etc. and interpretation, and instrument-parameter to experiment-parameter translation.

New approaches such as blogging are also of interest so Karl has been collaborating with Jeremy Frey in the use of LabTrove. To put all these approaches together, community buy-in, and partnerships are being developed with other national facilities in multiple countries, other NMR data model efforts, and NMR spectrometer companies.

Natural Language Processing and the Semantic Web

Lezan Hawizy was indisposed on the day of the symposium and a video was shown of her presentation about natural language parsing for semantic science. ChemicalTagger is an open source package for

"understanding" organic chemistry experiments, developed by Peter Murray-Rust's group, using natural language processing (NLP) approaches. Tools available include Open Source Chemistry Analysis Routines (OSCAR)[22] and OPSIN.[19] ChemicalTagger converts flowing text into structured text. Processes such as dissolve phase, purify phase and yield phase are marked up in the chemical procedure. Components of ChemicalTagger include tokenizers, which split a sequence of text into individual tokens; taggers, which assign parts of speech to each token; a parser which groups tagged tokens into phrases; and a role identifier which assigns roles to the parsed phrases. Taggers include OSCAR for chemical entities, RegEx for chemistry-related entities, and OpenNLP (http://opennlp.apache.org/) for English entities. The parser has a rule-based grammar for molecules, amounts etc. The role identifier assigns action roles (e.g., "dissolving") to phrases, and roles such as "solvent" to molecules. The role identifier was evaluated using 50 experimental paragraphs by comparing the effort of four annotators with each other and with ChemicalTagger, using the Dice coefficient to measure similarity. There was about 90% agreement between human and machine tagging.

Daniel Lowe has expanded the work to chemical reactions. The software identifies experimental sections, uses ChemicalTagger with an additional OPSIN tagger to produce structured data, associates chemical entities with quantities, assigns chemical roles, and carries out atom-atom mapping. Daniel extracted 424,621 reactions from 65,034 patent documents. Hannah Barjat has developed an additional tagger, ACPTagger, for use with the open access journal *Atmospheric Chemistry and Physics*. Lezan showed some visualization features of the resulting system, including geolocations mapped onto a map of the world.

Materials informatics requirements are substantially different from small molecule informatics: while structural representations of small molecules often contain enough information for the development of structure-property relationships, this is frequently not the case for complex materials. Often an account of the provenance of a material must be added to the chemical representation of a material. Additionally, materials data are usually generated in "native vernaculars": non-portable formats, which do not easily allow for data exchange. To make these data widely accessible, they must be converted to formats with both human as well as machine comprehensible standard syntax and semantics.

Nico Adams of CSIRO has used a complete Semantic Web toolstack, from XML dialects to axiomatically rich ontological models in Web Ontology Language OWL (http://www.w3.org/standards/techs/owl#w3c_all), in the development of modern materials information systems. Nico showed an example of Polymer Markup Language (PML) and the ChemAxiom ontology for polymerization and he produced graphical representations that describe a chemical procedure. Synthetic robots produce a log file that can be decoded by the manufacturer but Nico had to put in some effort to convert the information into an ontology and graph. Unfortunately the robot does not know what chemistry went into the robot. This has to be caught elsewhere. Nico uses ChemicalTagger.

Janna Hastings of EBI started her talk with her conclusions: classification conveys the type for data; the Semantic Web makes data of all types available, open and interlinked; and classification using OWL ontologies dramatically enhances the potential of the chemical Semantic Web. The subject and object in

an RDF triple are types. Molecules are small and three-dimensional. Their structures can vary according to their environment. We say they have the same type when they share important properties. All caffeine molecules have type caffeine. There are many different ways to represent a molecule: by InChI, by a reference number, by a ball and stick model, and so on. None of these is, in itself, a molecule; all these describe and approximate. All data are representations. Science aims to make discoveries of general rules about the things that the data are about. Classification puts the scientific knowledge into the data. RDF is a technology for data representation and OWL is a technology for classification.

Ontologies encode expert domain knowledge in a hierarchically organized format that a machine can process. One such ontology for the chemical domain is ChEBI.[23] ChEBI provides a classification of chemicals based on their structural features and a role- or activity-based classification. An example of a structure-based class is "pentacyclic compound" (compounds containing five-ring structures), while an example of a role-based class is "analgesic", since many different chemicals can act as analgesics without sharing structural features.

ChEBI has been applied to annotation of chemicals in biological contexts and for diverse tasks of chemical discovery including metabolic network gap prediction, but its growth has been limited to the throughput of manual annotation. A recent publication[23] describes the requirements for structure-based, automated classification; the analysis of structure-based features of chemical classes in ChEBI; and mapping to existing OWL-based technology and cheminformatics-based approaches. Another publication[24] describes feature and maximum common substructure detection for a group of chemicals, asserts class definitions logically using OWL and SMARTS, and demonstrates automated classification using OWL reasoning.

Exploration and Analysis

In the pre-Google era, Henry's team wrote an indexing and search engine called ChemDig;[25] in the post-Google era, Geoffrey Hutchison at the University of Pittsburgh has built ChemSpotlight (http://chemspotlight.openmolecules.net), using Spotlight (the desktop search feature of Apple's OS X operating system) plus Open Babel[26] and about 300 lines of code. ChemSpotlight is a metadata importer plug-in for Mac OS X, which reads common chemical file formats using the Open Babel chemistry library. Spotlight can then index and search chemical data: molecular weights, formulas, SMILES, InChI, fingerprints, etc. The data are kept as native files with a separate index. The current version (with about 800 more lines of code) allows freely rotatable 3D views of molecules and 2D views of ChemDraw and molfile formats, thanks to the ChemDoodle WebComponents. Geoffrey refers to ChemSpotlight as an "undatabase" because it has no (visible) database or SQL. It stores fingerprints, and number of atoms, bonds, and residues, PDB and SDfile keywords and properties, calculation keywords, and calculation results. Geoffrey presented a new genetic algorithm approach with Spotlight for designing new molecules for organic heterojunction solar cells, by calculating electronic and optical properties, and a synthetic score, for virtual libraries of more than a million compounds. His take-home message was that "undatabases" and ChemSpotlight, integrated into user-friendly tools, work well for big data.

Brian McMahon of the International Union of Crystallography (IUCr) talked about crystallographic publishing in the semantic age. The Semantic Web adds value (and meaning) to data in IUCr journals online through linking, allowing navigation, search, provenance, accreditation and access to related data and literature. Dynamic textual annotation of IUCr article content currently gives links to the *Online Dictionary of Crystallography* and the *IUPAC Gold Book*. The layout in HTML tables implies some semantics and can communicate meaning to another application (e.g., Jmol to highlight a selected bond).

The Crystallographic Information File (CIF)[27] information interchange standard has informed the structural content of CML. CIF was designed from the outset as an extensible standard, and now covers many areas of crystallography. It forms the basis for integrated data and publishing workflows linking laboratories, data repositories, publishers and databases, and has been an important factor in improving the quality of published crystal structures. The CIF publishing editor pubCIF (http://journals.iucr.org/services/cif/publcif/) is a desktop application for formatting and validating CIFs. CIF acts as a vehicle for article submission; checkCIF (http://checkcif.iucr.org) can be used to validate the structural model. An enhanced figures toolkit (http://submission.iucr.org/jtkt) brings an article alive by creating Jmol enhanced figures. The CIFs in SI for non-IUCr articles on the Web can be loaded into the IUCr visualization tool. The metadata about instrument, refinement etc. is available. CheckCIF can be run on the SI. Brian concluded his presentation with some charts showing where CIF sits in the data flow in crystallography and the publication flow in IUCr journals.

Kitware has developed a new open-source application, ChemData (part of the Open Chemistry project), to facilitate the exploration and analysis of large chemical datasets. Kyle Lutz described the program features of which include a variety of 2D plotting techniques, such as traditional scatter plots, parallel coordinates charts, and scatter plot matrices. Similarity relations between molecules can be explored using a range of graph-based visualization methods. Multiple querying and filtering functions allow users to locate molecular data relevant to their work.

ChemData is a native C++ application built with the user interface framework Qt (http://qt-project.org/). It uses the NoSQL database MongoDB (http://www.mongodb.org/) as a semantic data store, focusing on cheminformatics and assessment of chemical properties such as QSAR data. Computational chemistry data are stored directly in the file store, and semantic data are extracted to facilitate search and analysis. ChemData uses the Visualization Toolkit (VTK, http://www.vtk.org/) for 2D and 3D dataset visualization. Molecular structure, geometry, identifiers and descriptors are stored as a single "ChemicalJSON" object. JSON is used as the data interchange format, rather than XML/CML, because it is more compact, it is the native language of MongoDB, and it is easily converted to a binary representation. Initial work is in progress for using Web-based visualization and analysis tools. ParaViewWeb (http://paraviewweb.kitware.com/PW/) accesses the MongoDB database and will provide a collaborative remote Web interface for 3D visualization with ParaView as a server. ParaView (http://paraview.org/) is an open-source, multi-platform data analysis and visualization application.

InChI and Databases on the Web

Stephen Heller, the project manager for InChI (http://www.iupac.org/home/publications/e-resources/inchi.html), outlined the significance of this standard. InChI is a non-proprietary, machine-readable string of symbols which 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 appropriate software. InChI is not a registry system. It is not a replacement for any existing internal structure representations; it is in addition to what one uses internally. Its main value to most organizations is in linking information. Like a barcode, it is not designed to be read by humans. The InChIKey has been designed so that Internet search engines can search and find the links to a given InChI. To make the InChIKey the InChI string is subjected to a compression algorithm to create a fixed-length string of upper-case characters. Steve showed examples of Google searches for an InChI and an InChIKey, and of Henry Rzepa's QR smartphone app for InChI.

The InChI Trust (http://www.inchi-trust.org/), a UK charity, was formed to develop and improve on the current InChI standard, further enabling the interlinking of chemistry and chemical structures on the Web. InChI is a truly international project with programming in Moscow, computers in Germany, incorporation in the UK, and a project director in the United States. Collaborators from over a dozen countries, from academia, pharma, publishing, and the chemical information industry, have all offered senior scientific staff to develop the InChI standard. InChI is a success because organizations need a structure representation for their content so that it can be linked to and combined with other content on the Internet. InChI provides an excellent return on investment. It is a public domain algorithm that anyone, anywhere can freely use.

ChemSpider (http://www.chemspider.com/) would not have been possible without InChI. Valery Tkachenko of the RSC put it into perspective. We live in the world of Web 2.0; a connected world of social networks, mobile communications and Internet TV; a big data world with semantic content and new interfaces. Data is king and NoSQL is the new data model approach. Data flows in and can be structured, searched, linked and navigated. Data and code are distributed and self-sustained in the cloud. Federated systems take precedence over standalone solutions. Sophisticated human computer interfaces and pervasive machine to machine interfaces prevail. Yahoo, Google, Facebook and YouTube are huge islands on the Internet map; why are chemical domains so insignificant?

ChemSpider is a database and search engine for small organic molecules, their properties, names and synonyms, and spectra. It is an aggregator of information from online resources as well as a host of data extracted from RSC scientific articles. Over the past five years over 26 million chemicals together with a diverse array of associated data have been deposited. The online database is open to community deposition, annotation and curation and, as a result, has expanded into a rich resource to contribute to a Semantic Web of chemistry. ChemSpider provides access to its data *via* Web Services and as RDF. There is an extensive infrastructure: a computer farm and components. Standard interfaces such as Simple Object Access Protocol (SOAP), Representational State Transfer (REST), JSON, RDF and SPARQL are used. Automated validation and standardization procedures are now being developed. ChemSpider provides the chemistry services supporting the Open PHACTS project (http://www.openphacts.org/),[14] a

semantic project serving the life sciences community to facilitate the linking of chemical and biology data and enable drug discovery.

Chemistry is also available in Wikipedia. Martin Walker of the State University of New York at Potsdam described DBpedia, a project to extract chemical data from Wikipedia. The substance information is in a ChemBox or DrugBox. Traditionally these boxes were used simply for cutting and pasting but the chemistry Wikipedia team has made a machine friendly version using formats such as SMILES and InChI. Now ChemBoxes are more like a database, and it is easier to pull data out. The InChIs for complex molecules can be very long, and this was a hindrance to their use in Wikipedia until "show/hide" became available. "Table creep" could be a problem in data pages; the answer is to put data on a supplementary data page.

Data validation lets the user know if the data are correct. Curation is the ongoing process of fixing errors. In 2008 a validation exercise was initiated and, in collaboration with CAS, 3,500 substances have been validated as having the same name, structure, and CAS Registry Number (CAS RN). Validated entries carry a green check mark. Every old version of an article, with a RevID, is preserved for posterity and can potentially serve as a permanent record of a validated version. To protect validated fields, a bot patrols the pages and logs dubious CAS RN edits, in a system developed by Dick Beetstra of Eindhoven University. Structures present more of a problem since they are loaded from an external file on Wikimedia Commons which can be "invisibly" changed, but, since fall 2010, a modified bot has been looking out for such changes.

Another example of data rich chemistry in a wiki is RSC's LearnChemistry wiki which aims to enrich RSC educational content with data from ChemSpider, and then make it open for educators to contribute their own content. ChemSpider provides data on structures, physical properties, spectra, etc. Martin and his colleagues wanted to make the data presentation more suitable for students, including high school students, and cut out all the content that beginner students would not use. LearnChemistry includes laboratory experiments, tutorials and guides, substance pages, quizzes, and project and collaboration pages. Users can share their own educational materials such as homework problems and laboratory procedures.

Conclusion

Bobbie Glen of the University of Cambridge summed up Peter and Henry's contributions to the Semantic Web of chemistry. Traditionally, science involved two main pillars: theory to generate hypotheses and experimentation to test them. In modern science, theories are complex, data volumes are large, and experimental teams are often international collaborations. We can add a third pillar, e-Science, to manage these new realities of science.[28] For e-science we need open data and standards; glue ware for computation and analysis, interfaces that encompass the "system"; access control to data and intellectual property, collaboration methods that allow analysis, dialogue and data exchange; data and data analysis tools for "big data"; scalable, physically realistic algorithms; infrastructure (networks, high performance computing, and data storage), and metadata and semantics to put it all in context. Biology, chemistry and patents have "big data" e.g., 429,512,389,024 nucleotide bases, 60,475,000 chemical

substances and 150,000,000 pages of European patents. The connections present big opportunities for innovation, but also great challenges. Navigation through all this information is not easy.

Most real chemicals do not exist as connection tables; the sticky, brown stuff in the reaction vessel is not a SMILES. The next generation of chemical information tools should capture the history of the materials and the manufacturing process which went to make up the substance, as well as measured and predicted properties, and that is just a beginning. Peter and Henry's work with CML,[6,29] opens up opportunities to do just this, once we capture the data.[30] The first step is the automated lab: data capture using the human senses integrated into robotic data capture. *Everything* should be stored (minor omissions often mean an unrepeatable experiment) and a knowledge framework is needed (semantics) that gives meaning to the data: any result has to be put in the context of the experiment.

Bobby gave a few examples. The solubility of caffeine varies by orders of magnitude in the literature. Single values tell nothing useful: we need the metadata to tell us what the material was and how the solubility was measured. How flufenamic acid is made determines the aqueous solubility because there are two polymorphs, made under different conditions, with different solubilities. 6,6'-Dinitro-2,2'-diphenic acid exhibits atropisomerism (the conformation is twisted to reveal an enantiomeric structure) so how the material was synthesized needs to be included in the data. Different atropisomers of a compound have different biological activities. Some bicyclo[3.2.0]heptan-6-one derivatives have two forms in a single crystal and in solution because of transannular interactions: how should this "dynamic" molecular structure be represented? Chemistry is not best served by 20th century descriptions of molecules and materials. CML allows the addition of vital metadata within a semantic framework, which adds context, reproducibility and knowledge.

References

1. Baird, M. S.; Al Dulayymi, J. R.; Rzepa H. S.; Thoss, V. An Unusual Example of Stereoelectronic and Entropic Control in the Ring Opening of 3,3 Disubstituted-1,2-Dichloro-Cyclopropenes. *J. Chem. Soc., Chem. Commun*, **1992**, 1323-1325.
2. Rzepa, H. S.; Whitaker B. J.; Winter, M. J. Chemical Applications of the World-Wide-Web. *J. Chem. Soc., Chem. Commun.* **1994**, 1907-10.
3. Casher, O.; Chandramohan, G.; Hargreaves, M.; Leach, C.; Murray-Rust, P.; Sayle, R.; Rzepa, H. S. Whitaker, B. J. Hyperactive Molecules and the World-Wide-Web Information System. *J. Chem. Soc., Perkin Trans 2*, **1995**, 7-11.
4. Camilleri, P.; Eggleston, D. S.; Rzepa, H. S.; Webb, M. L. Intermolecular interactions responsible for the absence of chiral recognition: aromatic C–H ▪▪▪O hydrogen bonding in the crystal structure of 3-chloro-9,13-dibutylamino-1-hydroxypropyl-6-trifluomethylphenanthrene propan-2-ol solvate hydrochloride. *J. Chem. Soc., Chem. Commun.* **1994**, 1135-1137.
5. Rzepa, H. S. Chemical datuments as scientific enablers. *J. Cheminf.* **2012**, *4*, in press. http://www.ch.ic.ac.uk/rzepa/datument/.
6. Murray-Rust, P.; Rzepa, H. S.; Wright, M. Development of chemical markup language (CML) as a system for handling complex chemical content. *New J. Chem.* **2001**, *25*, 618-634.

7. Marshall, E. L.; Gibson, V. C.; Rzepa, H. S. A Computational Analysis of the Ring-Opening Polymerization of *rac*-Lactide Initiated by Single-Site β-Diketiminate Metal Complexes: Defining the Mechanistic Pathway and the Origin of Stereocontrol. *J. Am. Chem. Soc.* **2005**, *127*, 6048–6051.
8. Downing, J.; Murray-Rust, P.; Tonge, A. P.; Morgan, P.; Rzepa, H. S.; Cotterill, F.; Day, N.; Harvey, M. J. SPECTRa: the Deposition and Validation of Primary Chemistry Research Data in Digital Repositories. *J. Chem. Inf. Model.* **2008**, *48*, 1571-1581.
9. Hughes, G.; Mills, H.; de Roure, D.; Frey, J.; Moreau, L.; schraefel, m. c. [*sic*]; Smith, G.; Zaluska, E. The semantic smart laboratory: a system for supporting the chemical eScientist. *Org. Biomol. Chem.* **2004**, *2*, 1-10.
10. Wang, X.; Zhang, B.; Wang, D. Z. Reductive and Transition-Metal-Free: Oxidation of Secondary Alcohols by Sodium Hydride. *J. Am. Chem. Soc.* **2011**, *133*, 5160–5160.
11. Breslow, R. Evidence for the Likely Origin of Homochirality in Amino Acids, Sugars, and Nucleosides on Prebiotic Earth. *J. Am. Chem. Soc.* **2012,** *134*, 6887-6892.
12. Willighagen, E. L.; Wehrens, R.; Buydens, L. M. C. Molecular chemometrics. *Crit. Rev. Anal. Chem.* **2006**, *36*(3-4), 189-198.
13. Willighagen, E. L; Jeliazkova, N.; Hardy, B.; Grafstrom, R. C; Spjuth, O. Computational toxicology using the OpenTox application programming interface and Bioclipse. *BMC Research Notes* **2011**, *4*, 487.
14. Williams, A. J; Harland, L.; Groth, P.; Pettifer,S.; Chichester, C.; Willighagen, E. L; Evelo, C. T.; Blomberg, N.; Ecker, G.; Goble, C.; *et al.* Open PHACTS: Semantic interoperability for drug discovery. *Drug Discovery Today* **2012**. Available online June 6, 2012.
15. Guha, R.; Schürer, S. C. Utilizing high throughput screening data for predictive toxicology models: protocols and application to MLSCN assays. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 367–384.
16. Hanwell, M. D.; Curtis, D. E.; Lonie, D. C.; Vandermeersch, T.; Zurek, E.; Hutchison G. R. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminf.* **2012**, *4*, 17.
17. Christoffersen, R. E.; Angeli, R. P. Quantum Pharmacology. In *New World Quantum Chem., Proc. 2nd Int. Congr.*; Pullman, B, Parr, R., Eds.; Reidel: Dordrecht, The Netherlands, 1976; pp 189-210.
18. Bersohn, M. Syntheses of drugs proposed by a computer program. In *Computer-Assisted Drug Design*; ACS Symposium Series 112; American Chemical Society: Washington, DC, 1979; pp 341-352.
19. Lowe, D. M.; Corbett, P. T.; Murray-Rust, P.; Glen, R. C. Chemical name to structure: OPSIN, an open source solution. *J.Chem. Inf. Model.* **2011**, *51*, 739-753.
20. Brooks, B. J., Thorn, A. L.; Smith, M.; Matthews, P.; Chen, S.; O'Steen, B.; Adams, S. E.; Townsend, J. A.; Murray-Rust, P. Ami - the chemist's amanuensis. *J. Cheminf.* **2011**, *3*, 45.
21. Bowers, G. M.; Ravella, R; Komarneni, S.; Mueller K. T. NMR Study of Strontium Binding by a Micaceous Mineral. *J. Phys. Chem. B*, **2006**, *110*, 7159-7164.
22. Jessop D. M.; Adams, S.; Willighagen, E. L.; Hawizy, L.; Murray-Rust, P. OSCAR4: a flexible architecture for chemical text-mining. *J. Cheminf.* **2011**, *3*, 41.

23. Hastings, J.; Magka, D.; Batchelor, C.; Duan, L.; Stevens, R.; Ennis, M.; Steinbeck, C. Structure-based classification and ontology in chemistry. *J. Cheminf.* **2012**, *4,* 8.
24. Chepelev, L. L.; Hastings, J.; Ennis, M.; Christoph Steinbeck, C.; Michel Dumontier, M. Self-organizing ontology of biochemically relevant small molecules. *BMC Bioinformatics* **2012**, *13*, 3.
25. Gkoutos, G. V.; Leach, C.; Henry S. Rzepa, H. S. ChemDig: new approaches to chemically significant indexing and searching of distributed web collections. *New J. Chem.* **2002**, *26*, 656-666.
26. O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: an open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.
27. Hall, S. R.; Allen, F. H.; Brown, I. D. The Crystallographic Information File (CIF): a New Standard Archive File for Crystallography. *Acta Crystallogr*. **1991**, *A47*, 655-685.
28. *The Fourth Paradigm. Data-Intensive Scientific Discovery*. Hey, T., Tansley, S, Tolle, K., Eds.; Microsoft Research: Redmond, WA; 2009.
29. Murray-Rust, P.; Rzepa, H. S. CML: Evolution and Design. *J. Cheminf.* **2011**, 3, 44.
30. Glen, R. C. Computational chemistry and cheminformatics: an essay on the future. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 47-49.