Herman Skolnik Award Symposium Honoring Prof. Alexander (Sandy) Lawson
Information Services in Chemical Sciences: Perspectives
ACS National Meeting, Denver, August 30, 2011

A report by Wendy Warr (wendy@warr.com)

When Sandy Lawson initially planned this symposium he said that he wanted to look to the future rather than the past, and that his theme was "change". The first two speakers addressed change as it impacts publishers. Bob Massie, President of CAS, talked about long-term implications for learned societies. The STM industry has prospered in the digital age when many other industries have been upended. One sign that it has done so is the ability to maintain prices. Music and book stores have closed; the circulation of most newspapers is in decline. Book publishing cannot control its economics. Asking why STM publishing has prospered, Massie suggested that executive leadership may have been a factor, and some of those leaders have been winners of the Herman Skolnik Award. On the other hand, he noted the observation by Warren Buffett that when an executive with a reputation for brilliance tackles a business with a reputation for bad economics, it is most often the reputation of the business that remains intact. So, while leaders matter, it is more likely we will find elements within the STM business itself to explain its resistance to the ravages of the Web.

Why then has STM publishing weathered the storm brought about by the Internet era? First, it was not advertising funded. In addition it did not allow itself to be "dis-intermediated". Beyond that, the value in STM products lies in the content, not in the users or their interaction with the content. This makes STM products less vulnerable to the viral power of the Web. STM publishing also relates uniquely to its professional marketplace: journals link to career paths and other information tools lock in the specific work functions. So far so good, but where should the industry invest now? Outsell's "five to watch" for the next two years are APIs, tablets, social messaging, HTML5 and MySQL. Massie predicts that primary publishers will continue to emphasize brand prestige and the career connection, while secondary publishers will continue to emphasize "workflow solutions". Will other players with the right technologies (Google and Facebook, for example) be a threat in future?

There are two ways an industry can adapt: retain the essential character but adapt new technologies and approaches, or change into something different. The issue for the future is whether STM publishing will be able to adapt without changing its fundamental nature. Finally, Massie pointed out that this is the "Asian century". Will the Herman Skolnik awardee in 2025 be a Chinese or Indian technologist?

In the second talk, Martin Tanke, Elsevier's Managing Director of S&T Journal Publishing, homed in on the learned journal, or, more precisely, "beyond the journal". Will journals continue to exist? Yes, says Tanke, as long as they truly support the scientist's workflow. Finding information is crucial to the processes involved in applying the scientific method. Faced with information overload, scientists end up repeating the work of others as they move through the disconnected data containers. Journals have a role in integrating the containers. Journal content must be smarter. Publishers must tag and enrich articles, open up publishing platforms, embrace applications, and redefine how journals work on the Web.

Smarter content arises from semantic search and navigation. RSC, Nature and other publishers are all innovating in semantic technologies. One Elsevier example is entity extraction and marking up of reactants, conditions and solvents in a Reaxys synthetic procedure. Enabling the research community to collaborate drives innovation. Tanke showed community-built applications that have appeared since publishers opened up APIs. PANGAEA is a mashup with geographical data. The Genome Viewer on SciVerse Science Direct is part of Elsevier's "Article of the Future" project (http://www.articleofthefuture.com/). It connects articles with the NCBI database. Article of the Future improves readability and discoverability and is extensible. New content is created as a scientist reads an article. The display has three panes: on the left is navigation, in the middle is a "traditional" reading pane, and on the right is a task-based pane that adds value and context.

Interoperability of journal content with raw research data sets is also important. Reflect, winner of the Elsevier Grand Challenge, tags and highlights proteins, genes, chemicals, and Wikipedia; and pulls information from EMBL and other databases. Author-created supplementary data files can be visualized and given added-value functionality with InChIKeys, Google and Reaxys links, all inside an article. The Reflect-Network application within SciVerse Applications and SciVerse ScienceDirect addresses the workflow challenges of life sciences researchers.

A lot has been achieved, said Tanke, but there is still work to be done. Publishers must set up an environment that works across all sciences not just chemistry; continue development of taxonomies and ontologies; devise authoring tools; work with suppliers; and achieve a balance between manual and automated approaches, since fully automated semantics might not scale.

Tony Williams of RSC spoke next, mainly about collaboration in the cloud, and open data. ChemSpider has been called the "Google and Wikipedia of Chemistry". Its vision is linking all chemistry on the Internet. Its roles are hosting and exposing data for the community and curating and validating chemistry-related data. ChemSpider is just one of many Internet resources that can be searched by chemical name, structure skeleton, molecular formula etc. Unfortunately, errors proliferate because of data sharing between the databases. Some public databases are "trusted" as primary sources and that trust is granted without investigation. Williams says you should never trust a public domain database. Indeed, he believes that you should never trust *any* database: always ask questions.

ChemSpider is curated in a never-ending, crowdsourced effort. Data curation is tough and, sadly, the "crowd" is small: only 131 people have ever become involved, but that does include a few "master curators". Reciprocal curation is thus a good idea. Identifier dictionaries of InChIKeys and synonyms allow curators of other databases to check if their data match ChemSpider's. DrugBank is already using this facility. Batch validation also works (for example, checking if there is a count for chlorine in the molecular formula of a compound that has "chloride" in the name). To validate spectra ChemSpider has constructed a game called SpectralGame (http://spectralgame.com) and a learning tool (http://spectraschool.rsc.org).

The community can also contribute reactions to ChemSpider SyntheticPages but submissions have been few to date. Williams suspects that one problem is that chemists fear that a SyntheticPages entry will be

considered “prior publication”, preventing formal publication in a high impact journal. ChemSpider would grow if it had access to supporting information from journal articles. Williams ended his talk with a plea for data to be open. Examples include OpenPHACTS (http://www.openphacts.org/) and RSC LearnChemistry. Williams is convinced that more collaboration can benefit us all.

Jan Brase of TIB (Technische Informationsbibliothek Hannover, the German National Library of Science and Technology) speaking on behalf of Uwe Rosemann, addressed change in libraries. The Digital Agenda for Europe outlines policies and actions to maximize the benefit of the digital revolution for all. Supporting research and innovation is a key priority for the Agenda and is essential if Europe wants to establish a flourishing digital economy by 2020. Brase said that the answer to the deluge of data is not to turn off the tap but to build boats.

TIB’s GetInfo portal is being opened up to data, maps, movies, PowerPoint files, graphs etc. The data can be held elsewhere as long as there is a persistent link. TIB has been issuing such links in the form of DOIs for data sets since 2005, and in 2009 it was a co-founder of DataCite (http://datacite.org/) which helps researchers to find, access, and reuse data. It aims to establish easier access to scientific research data on the Internet, increase acceptance of research data as legitimate, citable contributions to the scientific record, and support data archiving that will permit results to be verified and re-purposed for future study. Now TIB is addressing new media types, visual search and visualization. Challenges are to ensure quality and preservation, and migration to even newer media.

PROBADO is a visual search in architecture. Brase gave an illustration of indexing based on room connectivity graphs, after which visual searches such as “buildings with 15 rooms over three floors” can be carried out. Graphical queries, such as drawing a chair, are possible. In chemistry, CLiDE (http://www.simbiosys.ca/clide/) and chemOCR (http://infochem.de/mining/chemocr.shtml) extract chemical structures that are held as images, and produce live structures. TIB is collaborating with Thieme on publication of research data by assigning DOIs to data such as spectral peaks that occur in articles published by Thieme.

What if you could just draw a curve and search for curves “just like this one”? TIB is working with the Fraunhofer Institute IGD (Institut für graphische Datenverarbeitung) and the Technical University of Darmstadt on query by example and query by sketch in visual search. Note that the curve for cell phone use in India might be the same as that for the weather in Hawaii. Brase concluded by saying that the ultimate goal of dissemination of scientific and technical information is interlinking and search across all digital assets. The methods may have changed but the mission remains the same.

The next two papers illustrated changes in chemistry, the central science, and in particular how chemistry reaches out into biology. Robert Glen of the University of Cambridge gave an academic viewpoint and Torsten Hoffmann of Roche an industrial viewpoint. Glen’s team has faced the challenge of probing a new found target without any prior knowledge of suitable pharmaceutically active molecules. In particular they tackled Apelin, a difficult target, for which there were no small molecule leads. Apelin is a G-protein coupled receptor (GPCR) which is a potent vasoconstrictor.

The approach of Glen's team was to model the receptor and associated endogenous ligands to understand the criteria for binding (and mechanism of action) combined with compound selection to optimize chemical structures for efficacy and affinity. Using this approach and rational design they have discovered novel agonists, partial agonists, antagonists, and some micromolar small molecule leads.

Using previously published alanine scanning data (Fan, X. *et al.* Structural and Functional Study of the Apelin-13 Peptide, an Endogenous Ligand of the HIV-1 Coreceptor, APJ. *Biochemistry*, **2003**, *42*(34), 10163–10168) they investigated the changes in the biological activity and linked this to structural and dynamic features of both ligand binding and receptor dynamics. They also constructed cyclic peptides and used NMR and constrained MD to study the shape of the peptides. Analysis was done with replica exchange molecular dynamics. A beta-turn at the RPRL motif was important for binding affinity (Macaluso, N. J. M.; Glen, R. C. Exploring the RPRL' Motif of Apelin-13 through Molecular Simulation and Biological Evaluation of Cyclic Peptide Analogues. *ChemMedChem* **2010**, *5*(8), 1247-1253). Analogues were synthesized, pharmacophores were generated, and molecular dynamics was used to study them.

Glen and co-workers analyzed the binding modes, interactions, and dynamics and related these to potency, agonism and antagonism. They then attempted to make an antagonist by stabilizing the antagonist conformation and they designed peptides with two "anchor" groups and a variable linker. The first competitive antagonist was discovered. Biased agonism is a new emerging concept in GPCR pharmacology. Glen used molecular dynamics to show the differences between apelin (a full agonist) and a biased agonist; he showed the audience the motion of the 7-helix and associated -C terminal loop which seems to be associated with biased agonism.

His team has combined this methodology with access to ethically sourced human tissue, an approach which eliminates many of the problems associated with animal testing and which also allows investigation of not only healthy, but diseased tissue. Drugs can then be targeted at the diseased state, which is more relevant in a clinical setting. As an aside, I was interested to note the multidisciplinary, international nature of Glen's research team.

Torsten Hoffmann of Roche subtitled his talk "if we only knew what we already know". This was a good way of expressing the challenge of knowledge capture and retrieval in medicinal chemistry. He started by explaining the workflows involved in carrying out medicinal chemistry research. He did this with reference to RG1678, a potent and selective GlyT1 inhibitor for the treatment of schizophrenia. A benzoylpiperazine hit was identified through high throughput screening, but it contained a nitro group (with potential for mutagenicity) and it had some undesirable properties that needed to be improved in lead optimization. A methylsulfone replacement was found for the nitro group, and, after SAR explorations, a series of compounds was found which had good overall physicochemical properties, high metabolic stability, oral activity, and no undesirable cytochrome P450 and off-target activity. The scaffold was also patentable. RG1678, in this series, has an excellent overall profile. It was safe and well tolerated in Phase I trials and had an excellent pharmacokinetic profile; in Phase II it improved the negative symptoms of patients with schizophrenia; and phase III studies are ongoing.

Hoffmann identifies three types of knowledge in discovery chemistry. Explicit knowledge can be shared in the form of hard data, scientific formulas, codified procedures or universal principles. External knowledge is communicated in journals, books, and at conferences. Tacit knowledge is *personal* and hard to formalize; Roche Chemistry Knowledge (ROCK) is a unique knowledge capturing tool for capturing tacit knowledge.

An editorial board controls the ROCK submission and review process. The knowledge can be browsed, or substructure-searched in combination with keywords. Scientists can retrieve both the knowledge and the experts involved within a few mouse-clicks. ROCK creates a knowledge sharing culture, fosters a life-long learning attitude, and offers a reward and recognition scheme. It is regularly used by medicinal chemists at all Roche sites.

Hoffmann next showed a truly fascinating movie of the perceptive pixel technology in use at Roche (http://www.youtube.com/watch?v=MIz_25ehzs0). Molecules on cards can be moved around in a touch screen interface, like moving CD covers in iTunes. Chemists can scribble on a card, can erase a section of a molecule, can create piles of cards, and name and save them, and can drag an avatar onto a card pile. It is not only chemistry that can be handled: it is also possible to drag and drop clinical data in the interface.

In the future, users expect next neighbor analyses by user-defined similarity searching, visualization enhancements and navigation options of search results in 2D and 3D, pattern recognition tools for broadest possible knowledge access, and intuitive human-computer interfaces for portable devices with wireless access.

The next paper covered some other futuristic aspects of capturing and reusing data. Talking about enriched research documents at the cutting edge, Rudy Potenzone of SciencePoint Solutions wonders why we are not focusing on authoring tools now that the “e-paper” has arrived and we are on the verge of a major revolution. Authoring technology enables scientists to create elaborate versions of the results of research, capturing the full context of research in progress: the formal scientific report, and the very methods used, with a full data repository, and complete workflows. The resulting documentation offers the information for completely reproducible results.

The scientific e-paper will help to improve the quality of science, facilitate the intellectual transfer of the core discoveries, fully document the provenance of the research, and preserve the knowledge with complete context. Services such as visualization and analysis will be easily accessible on top of the data. This heralds an era of accessible, reproducible research.

When Microsoft introduced OpenXML in Office 2007 it paved the way for workflow options and add-ins such as Chem4Word (http://research.microsoft.com/chem4word), GenePattern (http://GenepatternWordAddin.codeplex.com), the Research Information Centre Project (a virtual research environment for SharePoint, http://ric.codeplex.com/), and the chemical Semantic Web in oreChem (http://research.microsoft.com/en-us/projects/orechem/). There are also several commercial data sharing and analysis services. Harvard’s Dataverse Network project (http://thedata.org) enables data archiving and preservation through re-formatting, standards and exchange protocols. It provides

control and recognition for researchers through data management, branding and formal data citation. Workflow and pipelining tools include Taverna, KNIME and Pipeline Pilot. Taverna is integrated with the myGrid open suite of tools (http://www.mygrid.org.uk/) designed to "help e-scientists get on with science and get on with scientists". The Project Trident Scientific Workflow Workbench (http://tridentworkflow.codeplex.com/) is built on Windows Workflow Foundation.

Potenzone concluded that the e-paper will provide a significantly more capable platform for science for the scientific community, but it will further erode the *status quo* system of rewards and tenure. As far as the business of science is concerned, e-papers mean that publishers must evolve into "cool providers" of tools and "hot" distribution centers; A&I companies will need to redefine their role; and software vendors have a real opportunity, if they can adapt. The developments Potenzone discussed offer opportunities for improving reproducibility of scientific results; for data sharing and collaboration; for reliable maintenance of provenance; for faster availability and efficient query tools; for controlled access to data; for finding related data and research partners; for assuring that data will be preserved; and for improved knowledge transfer.

The final invited talk was my own. Since my brief was to summarize the preceding papers it seems pointless to repeat myself in a written report that summarizes all the talks. Readers who want to hear the live and more humorous version can access the official ACS recording. What must be recorded here in print are the many tributes paid to Sandy Lawson himself. The official accolade lists his main achievements but during the symposium much was said about him as a person. He was praised as a gentleman, and a mentor, and someone who has survived with a reputation for impeccable ethical behavior despite the industry controversies that arose during his long career.

In my talk I picked out a few topics such as new interfaces (noting that no one mentioned multilingual search) and data reuse and open data. The underlying theme of the day was not "databases" but access to, and reuse of "data". I am sure that our awardee will not mind my mentioning that Reaxys gives immediate access to actionable data, although Reaxys was not the central theme of his talk. In his award address Sandy Lawson discussed some challenges and opportunities in preserving the scientific record. As more and more information gets published, the time-pressed reader feels more and more of a need for a focus on relevance. Chemistry databases are not passive media, they are interactive and have considerable underused opportunity to improve on pinpointing relevance *via* a specific description of a series of observations in words to which the users can respond.

Lawson cited three statements from Lawson, A. J. The Beilstein Database. In *Handbook of Chemoinformatics: From Data to Knowledge*; Gasteiger, J., Ed.; John Wiley & Sons: Chichester, UK, 2003; Vol. 5, pp 608–628:

"...the quality of a database...lies in the quality of its indexing power, to be able to reduce the suggested list of articles to a manageable number, without discarding relevance..."

"... expect secondary indexing systems...to retreat somewhat from an insular, stand-alone, centre-stage posture... [with] query formulators taking a less prominent role..."

"...much will be driven directly from the source article itself, coupled with a natural language interpreter...to encourage the user to follow up suggestions..."

These three statements in three (almost) consecutive sentences are essentially tying the concept of focus on relevance to a verbal exchange.

Databases of one sort or another are necessary if users are to find documents. A document "does not exist" for the user either when (a) it is not found at all, or (b) when it is lost amongst a host of irrelevant, other documents. In either case it is a question of relevance. Lawson dealt with the first point in his talk about natural language interfaces at the Herman Skolnik Award symposium for Guenter Grethe (Lawson, A. Question, query and relevant response: pick any two. Abstracts of Papers, 222nd ACS National Meeting, Chicago, Il, United States, August 26-30, 2001, CINF-062). In the current talk he concentrated on the second point: a very common problem facing all users of primary and secondary information systems.

The focus on relevance is more about inspecting answer sets, and less about formulating queries. Relevance is a judgment that lives in the eye of the user. Trying to express relevance exactly in query formulation requires chemists to "know" the nature of the answers (and how they are expressed) before even looking. Lawson believes therefore that an *answer set* should "know" about chemistry and be able to correlate the inspection of the set for relevance to the user, and be able to respond "what it thinks" in conversational terms. Lawson's approach is based on remote dynamic interpretation of database metadata, by a Visual Basic Web Service, an app, that is communication between two electronic devices over a network. Document content (words and graphics) can be converted into a document summary (in words and graphics). Lawson says this should be done even though there are titles and abstracts because titles and abstracts are *fixed* (on the historical unique focus of the author) but the user's focus on relevance is *not fixed*: it is based on the user's current thought and current query.

Lawson presented a proof of the concept. His prototype is bidirectional, i.e., it can "listen" and "talk" directly to a database system (here Reaxys). The requirements are to analyze a document and present its results for inspection in two ways: either as a synthetic backbone, or as a list of methodologies. The prototype worked well on two papers, one of which had an abstract and one of which did not, but the more interesting test was application to a "crowd" of documents.

In this example the app was applied to not one document, but to 101 documents all at once. Lawson formulated a query which even after refinement resulted in 840 reactions from 101 publications. This is far too many publications to be read or scanned. Each publication has its own single focus. If "relevance" is equivalent to [(user's focus) AND (publication focus)], which publications should Lawson now read? The focus on relevance is about inspecting answer sets. The analysis is automatic and a black box for the purposes of the present talk.

Lawson's engine listed the top seven methods in the whole hit set, and has the ability to call exactly these methods up from Reaxys. The app does not "know" everything: five out of the seven methods were named (e.g., "amidation with nitriles") but two were not. In practice the app is currently about

60% successful. Lawson viewed each of the seven methods, one by one, and judged them in the light of his own relevance view. This reduced the hit set to two methods from seven publications. The final fine-tuning uses manual inspection by focusing on the method as a percentage of the total article content. Two documents stood out with a high focus on relevance. This is 2% of the original 101 publications.

Using the engine, one preparative method was chosen, and there were just two documents to read. Had the searcher hoped to find these documents by use of keywords, title, or abstracts in the (Reaxys) query, he or she would have had to guess correctly and settle for some combination of generic chemical name and transformation terms. With Lawson's engine, the user merely has to recognize the term when he or she sees it, supported by the structures presented. This talk was not really about a new method or feature of any particular database; rather it was about general principles of communicating relevance in graphics and also words, reducing the load on users, by removing some aspects of "you need to know" at query formulation, and harnessing the power of databases from outside of the formal database user interface. Lawson concludes that both principles can contribute significantly to preserving the scientific record as a live entity moving forward.