The Future Information Needs of Pharmaceutical and Medicinal Chemistry

Royal Society of Chemistry (RSC) Chemical Information and Computer Applications Group (CICAG)

Monday 28th November 2011

A summary by Wendy A. Warr

This one-day meeting was ably organised, and chaired, by Yvonne Pope of CAS and Stuart Newbold of Astex Pharmaceuticals. Bookings and finance were handled by Diana Leitch, who also chaired the concluding session with questions and answers. Over 60 people attended including bursary students from Sheffield and Southampton.

Background to the pharmaceutical industry

The proceedings began with my own talk setting the scene about the pharmaceutical industry, if not about its information needs. It takes about thirteen years to discover and launch a new drug: 3-6 years for discovery and preclinical, 6-7 years for phases I-III, and 0.5-2 years for review by the Food and Drug Administration and scale up to manufacturing. Only one of about 50,000 compounds entering lead identification will make it to market; one in 250 survives the journey from development candidate to regulatory approval. The number of drugs entering phases I, II, and III fell by 47, 53, and 55% respectively in 2010 (according to Thomson Reuters *Pharmaceutical R&D Factbook 2011*), and 55 drugs failed at Phase III in 2008-2010, more than double the number of failures in 2005-2007. It costs about one billion dollars to develop a drug and only two or three in 10 approved drugs produce revenues that exceed average R&D costs.

The entire industry spent $68bn on R&D in 2010. The annual R&D spending of major pharmaceutical companies is declining very slightly but overall, R&D spending is still increasing: in short more research is being done outside major pharma. According to Deloitte and Thomson Reuters, 10 out of 12 top pharma companies saw their internal rate of return on R&D spending fall from 11.8% in 2010 to 8.4% in 2011, while the cost of bringing a drug to market increased by 21%. Despite the large sums spent on R&D it seems that number of new molecular entities (NMEs) approved by the FDA is decreasing, although some would claim that the mid-1990s were boom years and that the annual number of NMEs considered over many years does not really decline appreciably.

The industry's revenues continue to rise (sales totalled $856 billion in 2010) but percentage growth over the previous year has been falling significantly since 2003. The falling compound annual growth rate is just one of the threats that pharma faces. Others are generic competition (big pharma will lose $60bn in the next 5-6 years from patent expiries), price pressures, and crowded markets.

The good news, according to Deloitte and Thomson Reuters, is that there is more value from product commercialisation than there are losses from late-stage failures, and that non-R&D costs have declined.

To combat high costs in future, R&D organisations will share capabilities in non-competitive R&D areas: the Pistoia Alliance, for example, aims to ensure that the inability to exchange information securely and understandably will not be an obstacle to innovation.

The many financial problems faced by pharma have led to a rash of mergers and acquisitions (M&As), but M&As have a negative impact on R&D (see for example, LaMattina, J. L. The impact of mergers on pharmaceutical R&D. *Nat. Rev. Drug Discovery* **2011**, *10*, 559-560). The rationalisation needed after a merger, and the disruption, may mean that momentum is lost in research; growth is largely from cost savings. M&As are not a long-term solution.

Cost cutting, including closing sites and downsizing, is another strategy. In 2009 the industry lost more than 60,000 jobs; in 2010 more than 50,000. Restructuring in terms of creating smaller, more innovative organisational units has also been tried. In-licensing of drugs from outside of large pharma has been a big trend for a number for years but the *Pharmaceutical R&D Factbook* suggests that drugs developed in-house have a 20% better chance of making it to market than in-licensed drugs. Diversification is another strategy: Novartis, for example, has diversified into biologics, antibodies, vaccines and diagnostics.

“Personalised medicine”, i.e., aiming the drug at only that small proportion of the market that will really benefit, has led to much interest in biomarkers and diagnostics, but better drugs for smaller markets will inevitably have higher prices. Alliances and outsourcing, and targeting emerging markets are also being pursued, as are portfolio management techniques and life cycle management.

“Product lifecycle management” is a technique for extending the lifetime of a drug already on the market, for example by finding new indications for it. “Repurposing” or “repositioning” an existing drug presents less risk than launching a new one. In repurposing there has been a move from serendipitous to systematic approaches: literature driven techniques and mathematical models. Other strategies are reformulations, combination drugs, selling the drug “over the counter” (OTC) rather than on prescription, branded generics, and techniques for extending the life of a patent. Steve Arlington of Pricewaterhouse Coopers has good news for those of us in cheminformatics when he talks of “making R& D more virtual”. He recommends semantic technologies, computer-aided molecular design and predictive biosimulation.

Drug discovery in the 1970s involved unplanned innovation and serendipity. Drugs were based on natural products. Researchers used their intuition and carried out random screening. Workflow was linear and informatics was only peripheral. The 1990s saw advances such as genomics and proteomics, the emergence of technologies such as high throughput screening and combinatorial chemistry, a growth in the knowledge of protein structures, and progress in bioinformatics and cheminformatics. Today we start with knowledge of a biological target, and maybe a known protein structure, and screen the fewest compounds needed. Vast quantities of data are gathered and informatics is of strategic importance, supporting decision making. Multi-disciplinary teams share knowledge. Druglikeness is predicted in accordance with the “fail early, fail cheap ” maxim.

Carolyn Buck Luce of Ernst & Young uses a Web 2.0 metaphor to explain the changes in industry strategy. Pharma 1.0 used the blockbuster model: the focus was on the top line, i.e., reliance on high revenues from one or two highly successful drugs. Pharma 2.0 uses the strategies addressed in this talk and focuses on the bottom line. Pharma 3.0 will concentrate on delivering health outcomes, being customer-centric and being payer-insightful.

Use of information

Clearly all these changes in the climate have had an impact on information services. Frank Cooke of GlaxoSmithKline (GSK) gave an industrial viewpoint. Information Resources at GSK is a group within R&D IT. It has two main accountabilities: Published Information (with resources in "The Library" community on the intranet) and Competitor and Scientific Intelligence. Frank is concerned with Published Information. In keeping with trends at other pharma, GSK has closed its physical library and is using sophisticated tracking and metrics for electronic resources. Most companies are making more use of RSS (GSK is developing Outlook and Web Parts options) and are outsourcing non-confidential patent searching. Pharma are late adopters when it comes to operating systems and browsers: it may take GSK 2-3 years to move 10,000 people to Internet Explorer and Windows 7.

Most big pharma now have enterprise licenses for SciFinder and Reaxys. These products are seen as complementary tools, with different indexing policies and significant overlap. Moving to Web access has meant quicker updates and revisions but there are issues with dependencies on Java and the variety of chemical drawing packages supported.

There are major differences in approaches to alerting services; GSK has stopped most mediated alerts. Companies also vary in their approaches to text mining but vendors are becoming interested. GSK has made a significant effort in this field in the past but little work is currently ongoing. The company will begin looking into mobile technologies in 2012.

Published Information has a much reduced headcount and is still under pressure to move applications out of the group. The environment is much simpler compared to five years ago, with an emphasis on users helping themselves and cross-support between sister teams. The in-house CrossFire chemistry tool is now externally hosted, the Product Literature database has been eliminated (Embase and SciVerse Scopus are used instead), CI Workbench has been broken up and replaced with separate tools; competitive intelligence alerts and maintenance have been dramatically reduced; and all biomedical alerts have been dropped. A few browser plug-ins and Reference Manager remain in-house; a replacement for Reference Manager on the Web was not deemed cost effective.

Many of these streamlining activities are being driven by budgetary pressures and are supported by robust metrics. Metrics such as COUNTER provide only a high level view of usage. It is, in addition, important to understand how many people are using each product, the numbers using specific components of each product, whether multiple product components are used within a search session, what proportion of the target audience use a product, which alternative products are in use, and how overlapping products are being used. Frequent users and unique users can be distinguished.

GSK's Customer Activity Tracking product (CATS) is used in studying marketing and penetration, department cost-share modelling, checking the effectiveness of training events, looking at the relative use of features within one or similar applications (e.g., reaction searching in SciFinder and Reaxys), determining the "most read article"' in each department, studying the use of PDF *versus* HTML for electronic journal articles (HTML is beginning to take off), and for "right-sizing" in contract negotiations. Unix scripts are used to process vendor metrics and update the CATS Oracle database semi-automatically; firewall cache logs are parsed automatically.

One of the challenges for the future is providing affordable access to key literature sources at GSK's joint venture Stevenage Bioscience Catalyst (http://www.stevenagecatalyst.com/), an independent bioscience facility due to open in 2012. Frank concluded with some amusing anonymous quotations and the lessons that they illustrate. "*Artificial Intelligence is no match for natural stupidity*": no matter how user friendly and smart we think a system is someone out there will find it difficult. "*There are 10 types of people in the world: those who understand binary, and those who don't*": information professionals often assume that everyone else understands what they are talking about. "*Give a person a fish and you feed him for a day; teach him to use the Internet and he won't bother you for weeks*": we often give people tools and don't hear from them, so how do we know they're using the tool properly or getting real value from it?

The ChEMBL database

Anne Hersey of the European Bioinformatics Institute (EMBL-EBI, http://www.ebi.ac.uk/) described one specific tool which gives users access to curated structure-activity (SAR) data. ChEMBL (https://www.ebi.ac.uk/chembl/) is an open access database of druglike, bioactive compounds. It contains 2D structures, calculated properties (e.g., log*P*, molecular weight and Lipinski parameters) and abstracted bioactivities (e.g., binding constants; pharmacology; and absorption, distribution metabolism, excretion and toxicity (ADMET) data). The data are manually abstracted from the primary published literature on a regular basis, then further curated and standardised. Version 12 of the database contains 5.6 million bioactivity measurements for more than 1 million compounds and 8700 targets, of which 5526 are protein targets. A subset of about 350,000 compounds from PubChem (http://pubchem.ncbi.nlm.nih.gov/) is integrated. Access is available through a Web-based interface, data downloads and Web Services at https://www.ebi.ac.uk/chembldb.

Target data can be searched by keyword or protein sequence, or the target classification hierarchy can be browsed. Compound data can be searched with lists of keywords, SMILES strings, compound identifiers, or chemical structure. Assay data are searched by keyword. Anne presented two case studies. In the first she studied what is known about chemical structures that bind to the protein ZAP70, and what is known about their potency, selectivity and ADMET properties. The bioactivity data report has links to Uniprot (http://www.uniprot.org/), for sequence data, and the Protein Data Bank (PDBe, http://www.ebi.ac.uk/pdbe/) which holds 3D structures. The second case study started with a substructure search. Anne then retrieved bioactivities for the hit compounds and looked for other data such as clinical trials, 3D structures, and drug data. She pointed out links to Chemical Entities of Biological Interest (ChEBI, http://www.ebi.ac.uk/chebi/), ChemSpider (http://www.chemspider.com/),

DrugBank (http://drugbank.ca/), PDBe (http://www.ebi.ac.uk/pdbe/), PubChem (http://pubchem.ncbi.nlm.nih.gov/) and Wikipedia (http://www.wikipedia.org/).

Standardising the representations of chemical structures presents challenges. Software packages interpret structures differently, particularly when converting between structure formats. EBI's procedures have to be largely automated and inevitably some mistakes will be introduced. The standardisation protocol is based on the FDA Substance Registration System User's Guide (http://www.fda.gov/ForIndustry/DataStandards/SubstanceRegistrationSystem-UniqueIngredientIdentifierUNII/default.htm). Standard InChI (http://www.iupac.org/inchi/) is used to identify unique structures and merge tautomers. Parent compounds and salts are registered with different ChEMBL identifiers, and relationships are recorded in the database. Stereochemistry is recorded if it is known.

Integration with other resources also presents challenges. The database has to be maintained and updated, and data from different sources have different business rules. More data sources are becoming available and decisions have to be made about whether to integrate the data or link to the source. Scalability is an issue. So EBI decided to set up an external system for cross-referencing chemical structures and their identifiers between databases. By using standard InChI to compare chemical structures across databases, this system, "UniChem", enables tracking of identifier-to-structure assignments over time.

Semantic Web technology

After Anne's talk, a very different topic was introduced, that of Semantic Web technology in the laboratory, and electronic laboratory notebooks (ELNs). Although the research is carried out in academia, the underlying themes also have relevance in industry. Jeremy Frey of the University of Southampton emphasised the importance of metadata, and the need to *plan* experiments as well as to capture the data as they are detected during the course of the planned experiment. No longer should researchers in the laboratory be expressing regrets such as:

"*If only I knew exactly how she did this experiment.*"

"*I wish I had recorded things at the start the way I do now."*

"*I know all this supplementary information could be useful but will people really remember the format? Is it worth all the hassle?*"

"*I wish I could get the numbers from this graph: the PDF is not much use.*"

About 6 years ago Jeremy's team originally demonstrated their "Smart" ELN concepts in the Smart Tea project. In order to gain a better understanding of the chemists' laboratory experiences, and of the experimental design and execution process in particular, the team made tea as a chemistry experiment. Building on the insight gained from their e-science research, the team has extended the model of a digital infrastructure for supporting research to incorporate the familiar social networking paradigm of

the online blog, and the ELN system has developed into LabTrove (http://www.labtrove.org/) which is now also used at the University of New South Wales and elsewhere.

The experimental process begins with a plan. In the United Kingdom, chemists are required to fill in a COSHH assessment form (under the Control of Substances Hazardous to Health Regulations 2002) which articulates a safe plan for the experiment. COSHH plans are an ideal basis for driving an ELN. With just a little extra work the plan becomes a useful way to provide context for the digital record; methods are as important as the data. Metadata, essential for provenance, are collected right from the beginning of the experiment.

The Semantic Interoperability of Metadata and Information in unLike Environments (SIMILE) project at MIT (http://simile.mit.edu/wiki/SIMILE:About), with participation from the World Wide Web consortium (W3C, http://www.w3.org/) and Hewlett Packard, aimed to create interoperability among digital assets, schemas, metadata and services. Like related work at Southampton, it made use of the Resource Description Framework (RDF, http://www.w3.org/RDF/). RDF extends the linking structure of the Web to use Uniform Resource Identifiers (URIs) to name the relationship between things as well as the two ends of the link. This is usually referred to as a "triple".

Jeremy is a collaborator in the oreChem Project (http://research.microsoft.com/en-us/projects/orechem/), which aims to integrate chemistry scholarship with the Semantic Web. He gave an example of SharePoint integration; DeepZoom technology (http://www.microsoft.com/silverlight/deep-zoom/) is used as an alternative to lots of Web pages. BlogMyData (http://www.blogmydata.org/) is a virtual research environment (VRE) used in collaborations between scientists from many different disciplines and institutions. It combines the capabilities of LabTrove and the Godiva2 data visualisation system (http://behemoth.nerc-essc.ac.uk/ncWMS/godiva2.html) which allows users to browse interactively through large environmental datasets using only a Web browser (http://eprints.soton.ac.uk/164533/1/TuesT4BlowerBlogMyData.pdf). Southampton's blog[3] (http://blog3.rubyforge.org/) is a blogging engine specifically designed for the Semantic Web.

The RSC's free chemical database ChemSpider (http://www.chemspider.com) has added RDF functionality to its interface, in collaboration with the University of Southampton's School of Chemistry: ChemSpider is a Linked Data (http://www.w3.org/standards/semanticweb/data) source for oreChem. The oreChem core ontology describes three concepts: the planning of scientific experiments, their enactment, and causality of data products that are used and generated during the enactment. Planning (prospective provenance) describes a prospective experiment that will be enacted; enactment (retrospective provenance) describes a scientific experiment that was enacted.

As an example, the structure determination workflow in eCrystals (http://ecrystals.chem.soton.ac.uk, a repository for crystal structure determinations) can be described by an oreChem plan. Jeremy also briefly outlined some laboratory middleware applications: capturing sensor information remotely ("LabBroker") and Second Harmonic Generation experimental control and review (http://middleware.chem.soton.ac.uk/shg/). The Smart Research Framework project (SRF,

http://www.mylabnotebook.ac.uk/) will enhance and deploy the LabTrove, blog[3], and LabBroker services to a shared virtual infrastructure ("in the cloud").

Open PHACTS

The next talk, by Richard Kidd of RSC, was also concerned with the Semantic Web. He reported on the Open Pharmacological Concepts Triple Store (Open PHACTS, http://www.openphacts.org), a knowledge management project of the Innovative Medicines Initiative (IMI, http://www.imi.europa.eu/). IMI is a partnership between the European Community and the European Federation of Pharmaceutical Industries and Associations (EFPIA, http://www.efpia.org/Content/Default.asp?). There are 22 partners in the Open PHACTS project, including 8 pharmaceutical companies, 3 biotechs, and RSC. They are 8 months into this 36-month project.

"Information tombs" have been built to handle all sorts of data (*in vivo*, drug pipeline, literature, patents, news, safety etc.), with their own languages and metadata, resulting in a lack of interoperability. Pharmaceutical companies are accessing, processing, storing and re-processing public domain drug discovery data and for this they need to develop and apply a set of robust standards and implement those standards in a semantic integration platform, the "Open Pharmacological Space" (OPS). The Open PHACTS infrastructure project will also deliver services to support drug discovery programmes in pharma and in the public domain. Although an open system is being built, it must be able to accommodate non-open components in the real world. The guiding principles are "open access, open usage, open source".

OPS services sit on top of a semantic fabric, involving Linked Data. Syntactic normalisation is applied to the various open data sources and the OPS Linked Data in RDF triples are annotated with ontologies in a semantic normalisation process that leads to Linked Knowledge. The main architecture, technical implementation and primary capabilities are being driven by a set of prioritised research questions. Prioritised data sources are being defined based on the main research questions. Three exemplars will be developed to demonstrate the capabilities of the OPS system and to define interfaces and input and output standards. Three use cases have been defined to benchmark the OPS system towards current standard workflows in data retrieval and mining. The data retrieval and data and text mining applications must provide answers to relevant research questions. The interrogation model, graphical user interface (GUI) and interactivity, and presentation of results are being developed.

Seventy three exemplary research questions have been identified and prioritised by the pharma partners. Analysis of these showed that chemical substructure searching, chemical similarity searching, sequence and similarity searching, and bioprofile similarity searching functions would be required. Some prioritised data sources are ChEMBL (https://www.ebi.ac.uk/chembl/), DrugBank (http://drugbank.ca/), ChEBI (http://www.ebi.ac.uk/chebi/), PubChem (http://pubchem.ncbi.nlm.nih.gov/), ChemSpider (http://www.chemspider.com/), the Human Metabolome Database (http://www.hmdb.ca/), and WOMBAT (http://www.sunsetmolecular.com/index.php?option=com_content&view=article&id=15&Itemid=10) in chemistry; and EntrezGene (http://jura.wi.mit.edu/entrez_gene/), the Human Genome Organisation (HUGO) Gene Nomenclature Committee gene names (http://www.genenames.org/), Uniprot

(http://www.uniprot.org/), InterPro (http://www.ebi.ac.uk/interpro/), SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/), Wikipathways (http://wikipathways.org), Online Mendelian Inheritance in Man (OMIM, http://www.ncbi.nlm.nih.gov/omim) and the International Union of Basic and Clinical Pharmacology (IUPHAR) database (http://www.iuphar-db.org/) in biology. Chosen ontologies are the Gene Ontology (AmiGO, http://amigo.geneontology.org/cgi-bin/amigo/go.cgi), the Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg/) the Ontology for Biomedical Investigations (OBI, http://obi-ontology.org/page/Main_Page), the Bioassay Ontology (http://bioassayontology.org) and the Experimental Factor Ontology (EFO, http://www.ebi.ac.uk/efo/).

Agile development is being used and a working "lash up" system has been produced, constrained to technologies in the consortium plus a few data sources. It is focused on just two of the prioritised research questions. The objectives of the exercise were team building, performance and scalability analysis, proving whether the system works for two questions, and using the demonstration to recalibrate the build tasks in order to better respond to user requirements.

Technologies being used in the lash up are the Large Knowledge Collider (LarKC) platform for massive distributed incomplete reasoning (http://www.larkc.eu/), lsp4all (http://netnationen.dk/lsp4all.dk/), ChemSpider (http://www.chemspider.com/), ConceptWiki (http://conceptwiki.org/index.php/Main%20Page) for biomedical concepts and the BridgeDB (http://www.bridgedb.org/) identifier mapping framework for bioinformatics applications. LarKC was populated by loading RDF from source, but there are plans for an automated system that uses semantic site map standards. The demonstration is available at http://www.youtube.com/OpenPHACTS; GUIs in this were Utopia (http://utopia.cs.man.ac.uk/), Pathvisio (http://www.pathvisio.org/), and a generic interface from Lundbeck which will be open source. Results at http://iswc2011.semanticweb.org/fileadmin/iswc/Papers/PostersDemos/iswc11pd_submission_19.pdf show how real pharmacological queries can be answered over 4 billion text mined triples.

Next, an architecture for the whole system must be done and services (such as entity identification and resolution and representing similarity, Open Researcher and Contributor ID (ORCID, http://orcid.org/), and DataCite (http://datacite.org/) developed. Models include RDF and the nanopublication specification (http://www.w3.org/wiki/images/4/4a/HCLS$$ISWC2009$$Workshop$Mons.pdf). Tender documents for commercial storage providers will be prepared. It is hoped that the first public prototype will be released in September 2012.

Talks will take place with major potential partners for the project. OPS Community Workshops are key vehicles for building the OPS community and encouraging wider engagement. These are hosted twice a year, and focus on different aspects of drug discovery, the technology used, data sharing, sustainability, licensing and practical applications. The Pistoia Alliance is active in many areas of the pharma software pipeline, including the "Information Ecosystem". The Pistoia project "Semantic Enrichment of Scientific Literature" (SESL, http://www.pistoia-sesl.org) is directly related to Open PHACTS. Open PHACTS is seen as important to the Pistoia mission.

Publisher perspective

From the open and the experimental, the meeting turned to "essential" information services. Paul Peters gave a publisher perspective with an emphasis, he admitted, on a single publisher. Since Paul is a sales director for CAS, we could guess which publisher's perspective would be given. I have taken the liberty of adding some thoughts of my own to try and broaden the coverage. CAS, Paul told us, "is committed to doing what no one else does: providing the most comprehensive coverage of the world's disclosed chemistry. In summary he told us that "CAS is the world's authority for chemical information. CAS provides the most complete, curated, quality controlled, and current coverage…CAS is the preferred source for chemical information when customers need to understand complex chemistry and gain advantages to make breakthroughs". Could this talk have been a sales pitch? You might very well think that; I couldn't possibly comment. (As a conflict of interest statement I have to add that I am an ACS Editor.)

CAS' highly skilled abstractors and indexers add value, for example, by adding titles in English instead of machine-translated ones. CAS staff speak and read more than 50 languages, and CAS' coverage of Asian literature is growing. Increasingly, new chemical discoveries are being disclosed through patents: the percentage of new compounds added to CAS REGISTRY from patents has increased from 14% in 1976 to 46% in 2010. I have noticed how CAS has responded to competition: how soon would reaction relevancy ranking, experimental procedures from patents and journals, new sorting options and SciPlanner have been added to SciFinder had it not been for the market success of Reaxys? Similarly, I say to myself, the improvements in CAS' patent services that Paul described must surely have spurred on by developments at Thomson Reuters, who are now supplying their Markush package as in-house data feeds for use, for example, with ChemAxon's tools (see https://www.chemaxon.com/ugm-presentations/2011-us/#meeting-report). The drivers that Paul actually listed were the increasing numbers of new substances disclosed in the patent literature; unique substances continuing to be found in chemical catalogues, chemical libraries and Internet sources; and the fact that the Pacific Rim, especially Asia, is increasingly productive.

Frank had mentioned the decrease in the number of information professionals employed in the pharmaceutical industry and the trend towards end user searching. The one area where experts are still in demand is patent searching. To meet this need, a completely new STN will be phased into the market beginning in 2012, featuring project-oriented workflow, combined text and structure queries, simultaneous query and results interaction, real-time analysis of results, and virtually no system limits.

It is a pity that Bob Massie himself could not have been there to talk in a more generic sense about changes in the STM industry. He also has referred to the "Asian century". He has predicted that primary publishers will continue to emphasise brand prestige and the career connection, while secondary publishers will continue to emphasise "workflow solutions" (see http://bulletin.acscinf.org/node/256). It is interesting to compare his predictions with those of the final speaker, Bill Town of Kilmorie Consulting.

Reflections on the past and future

Bill's talk was an ideal conclusion to the proceedings, in that he started with some fascinating history that must have appealed to many in the audience, but moved on to some brave predictions about the

future of the pharma, software and publishing industries. He summarised some of his personal technology milestones from his university's acquiring its first computer in 1964 through to the launch of his own company's Web site in 2004. Apparently he got his first email address in 1979: CAS' wgt22@xtrn.org I assume. I too remember CAS email addresses but I am sure it was later than 1979 when I joined the mystical department 22. In 1986 Bill's company developed STN Express: the software has lasted 25 years but over the next few years a new platform will be phased in, as outlined by Paul Peters earlier. Technology is an enabler: when Bill was at school he used to phone his girlfriend from a red phone box, pressing button A (reader, he did not marry her); now he runs a business from home, holds video conferences, and has a social life that has become a hybrid of contacts between real and virtual friends.

Telecommunications enable global companies and their operations. Spare capacity in networks reduces costs, enables outsourcing, and leads to completely new business models. In the publishing industry, continued growth in numbers of journals and articles published is fuelled by the "publish or die" syndrome and the growth of science in Asia. Technology has enabled a shift from print to predominantly electronic publishing; "electronic paper" still predominates but semantic markup is making an impact. The library funding crisis has led to a demand for new business models. Open access publishing continues to advance but, in chemistry, it is still far from making a major impact. Outsourcing to the Philippines and India is reducing the production costs for typesetting and indexing.

Bill presented the Commercially Available Chemicals Index (CAOCI) as a case study. It began life as a collaborative project proposed in 1972/73 by some members of CNA(UK). Original members of the project in 1974 were Pfizer, Glaxo (including. Allen & Hanbury's), Boots, ICI, Beecham and Wellcome. In the mid 1980s, the file was acquired by MDL and became the Available Chemicals Directory (ACD). It is instructive to look at the fate of some of the original participants: ICI demerged its pharma and agrochemicals divisions to form Zeneca, Glaxo acquired Wellcome and BASF acquired Boots; and later Astra and Zeneca merged, as did GlaxoWellcome and SmithKline Beecham. Pfizer now comprises the Warner Lambert (Agouron, Farmitalia, Jouvenal, and Parke-Davis), Pharmacia (Monsanto, Searle, Sugen, and Upjohn) and Wyeth (American Cyanamid, American Home Products, A. H. Robins, and Genetics Institute). Accelrys now comprises MSI (Biodesign, Cambridge Molecular Design, Polygen, Biosym, and Biocad), Synopsys, Oxford Molecular (Biostructure, CAChe, Chemical Design, HDI, and PSI (Fein Marquart), GCG, Intelligenetics, and Cambridge Combinatorial), Synomics, SciTegic, Symyx (MDL (ORAC, and OHS)), and Contur Software!

Looking into the future is difficult (Bill's slides are available for those who want to see some amusing howlers) but Bill hazarded some predictions nevertheless. In the publishing industry, open access publishing will become the predominant business model within 10 years; some abstracting and indexing services will disappear or mutate; there will be more mergers of publishers; and one or more major publishers will be owned by an Indian company. In the pharma industry, the number of new chemical entities approved each year will remain about the same; the trend to outsource research to India and China will continue; there will be more mergers; and one or more major pharma companies will be owned by an Indian company. In the software industry HTML5 will be released in the 21st century; new software paradigms will emerge; new companies will appear and old companies will merge or die; and

"Brazil, Russia, India and China" (BRIC countries), or African countries, will dominate the industry. Pre-competitive and open collaborative efforts (such as Open PHACTS, OpenTox and the Pistoia Alliance) will thrive and some will succeed.

Conclusion

This was an interesting meeting and one that deserved a better attendance. Sadly, in the discussion session at the end there was an emphasis on the uncertain future for scientific information professionals and the poor career prospects for current chemistry students in chemistry *per se*. I was sorry that the discussion did not end on a more enthusiastic note. It seems to me from the wealth of material presented during the day that the future could be very interesting indeed if we are prepared to adapt and survive.