

Data-driven science - a scientist's view

Peter Murray-Rust

Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, UK

For ease of future access this position paper is in XHTML (not PDF) and is issued under a Creative Commons attribution license

History and motivation

Many classic results in science have come from the analysis of existing knowledge already available in the open literature. The chemical classic is Mendeleev's proposal of the [law of periodicity](#) [1] where he states:

The law of periodicity was thus a direct outcome of the stock of generalisations and established facts which had accumulated by the end of the decade 1860-1870: it is an embodiment of those data in a more or less systematic expression.

The data on which the law depended were not gathered systematically but published throughout the chemical literature and in many languages and symbolic formulations. Many of the sources would have taken months to locate and analyse and the metadata - the experimental conditions - were critical for establishing data quality.

In 1974 I visited Jack Dunitz's chemical crystallography laboratory in Zurich and carried out a data-driven study on chemical geometry using a similar approach. We were interested in the distortions of YMX_3 groups and whether they could be interpreted in terms of chemical reactions. The only search method was manual searches of paper-based tables of contents hoping that retrieved papers (often interlibrary loans) might contain structures of interest. For each successful paper (and there were about 100) numeric data had to be typed up and special programs written to compute and analyse the geometry. It took several months and fortunately science was kind - there was a strong and convincing relationship ([original ACS publication](#)) (2).

For many years, therefore, we have worked to create systems that can automatically read the current chemical literature, aggregate the data, add semantics and metadata and allow scientific hypotheses to be tested. More ambitiously it is possible for the system to extract patterns or unusual observations from which new hypotheses might be constructed. This is reflected in our OSCAR and CrystalEye systems (below). Our thesis is that the current scientific literature, were it to be presented in semantically accessible form, contains huge amounts of undiscovered science. However the apathy of the academic, scientific and information communities coupled with the indifference or even active hostility and greed of many publishers renders literature-data-driven science still inaccessible. The remainder of the document explores the organizational, political, and technical challenges.

Hypopublication and data-driven science

Many areas of science are published in a fragmented form rather than the systematic data collection of (say) astronomy, particle physics, or (some) genomics. I use the neologism "hypopublication" ("hypo-" = "below", "low", or "insufficient") to emphasize the inadequacy of current publication protocols and the lack of hyperlinking or aggregatability. For example about 2 million chemical compounds are published each year (about half in patents) with insufficient semantics, metadata or hyperstructure. Vast effort is required to create useful data from these, and the current commercial processes seriously disadvantage the whole of science. It should now be possible to publish a fairly complete scientific record of an experiment, yet the current publication process continues to emphasize the "article" at the expense of the data. The article summarises the experiment and gives the essential impact factor (market indicator for tenure and funding) - the data are often missing or so emasculated as to be useless. It is the film review without access to the film.

Many scientific disciplines require publication - in textual form - of sufficient data for the experiment to be evaluated (though frequently not enough to allow replication). Some communities laudably insist on machine-parsable data including much bioscience (genomes, protein sequences and structures) and crystallography. Over the years they have managed to coerce the publishers to require authors to provide this information. If all communities did this, for all major kinds of data, then literature-driven science would become a reality. Note, however, that some publishers (such as ACS and Wiley (3)) see such data as their property. Although "facts cannot be copyrighted", these publishers continue to insist on this and one senior representative recently told me that this was so they could "sell the data". To try to counter this I am promoting the concept of [Open Data](#) - including a mailing list offered by SPARC. The STM publishers have agreed that factual data is not copyrightable, but there is generally indifference in the academic and information communities to the importance of insisting on this.

It is important to stress that "Open Access" - as currently practised - does not promote Open Data. The Budapest and other declarations make it clear that Open Access involves free, unrestricted access to all the data for whatever legal purpose. In practice, however, publishers ban robotic indexing of sites, cut off subscribers whom they opine are downloading too much content, and continue to copyright facts. The politicisation and complexity of the Open Access struggle means that Open Data currently has little community recognition and support. Yet Open Data is the single most important problem in data-driven science.

I categorize data-driven science as requiring that the data be Open and be re-used by someone not connected with the original author without the need for permission or correspondence. Legal, syntactic, semantic and ontological questions must be answered by the data themselves. Where the data are collected systematically in a funded project these requirements are often specifically addressed, but in hypopublication this is rare. It is commonest in bioscience where the communities have developed standards for data deposition (PDB, SwissProt, etc.), where publishers require authors to deposit data and where there are major organisations (RCSB, NCBI, EBI, etc.) which aggregate it. Note that each type of data is well-defined, stable over time, and generally self-contained so that systematic aggregation is possible. Bioscience is one of the best examples of data-driven science and many bioinformatics laboratories do much science without ever going near a laboratory. In similar vein Sam Motherwell and I (1978) developed software to analyse the author-contributed crystal structures aggregated in the Cambridge Crystallographic Data Centre (4) from chemical hypopublications. Subsequently over a thousand publications have been based on this approach.

Requirements for data-driven science

I re-emphasize the paralysing effect of non-Open Data. The next most serious problem is that most scientific data are anyway irretrievably lost before publication. Most scientists do not expect - or do not want - anyone else to re-use their data and therefore wittingly or unwittingly make it very difficult to do so. John Davies (crystallographic service, Cambridge) has estimated that 80% of all the structures he does are never published and are gradually lost. This catalysed the JISC-funded SPECTRA project which not only corroborated this figure but estimated that in other areas (computational chemistry and spectroscopy) the figure could be > 99%. We estimate that, even at non-commercial rates, most large chemistry departments spend millions on collecting data which is subsequently lost to science. There is virtually no realisation by chemistry funders, departments or scientists themselves of the importance of making data accessible to humans and machines.

The next problem is the active destruction of data by the publication process itself. Many publishers do not provide support for supporting information, or hide it. The requirement to publish in double-column PDF (was this ever requested by the community?) is one of the saddest aspects of "electronic publication". It is almost impossible to recreate structured XML documents (the equivalent of turning a hamburger back into a cow) and it denies the value of semantic publication at the expense of lauding "electronic paper". Moreover many advocates of "Open Access" see "the final PDF" as the only thing that matters. All this has made the quest for data-driven science much harder and painful.

More cheerfully, some publishers of chemistry (such as the International Union of Crystallography and, more recently, the Royal Society of Chemistry) recognise the value of semantic publication. Given active support for Semantic Open Data and Standards, the remaining challenges are mainly technical:

- Data (datuments) must be in XML. This is harder than it seems - many communities are conservative and use the same non-extensible and fragile formats that were introduced 30+ years ago in the FORTRAN era. Thus authors use PDB, Swissprot and CIF in preference to XML as user-friendly tools are not available - in any case they still want to be able to read and edit files in a dumb text editor. Funders must be persuaded of the need to support tools - in chemistry there is almost no funding for this and we rely on virtual community action ([the Blue Obelisk](#)). However this activity is not regarded as "chemistry" and even when successful often counts adversely against tenure.
- Each discipline must create or re-use XML vocabularies. At present these are domain-independent (hypertext (XHTML), math (MathML/OM), graphics (SVG), data, numbers, units, etc. (CML/STMML)) and domain-specific (GML (geoscience), CML (chemistry), with ThermoML (thermochemistry) and AnIML (analytical) in chemical subdomains). The flexibility and generality of each language depends on the degree of acceptance in the community - MathML is well supported, while CML deliberately caters for the lack of information discipline in chemistry.
- Domain metadata is provided through community ontologies. A good example in chemistry is the IUCr's CIF dictionary, which provides definitions, datatypes, constraints and some relationships. In principle ontologies can and should be used to validate data but this is still rare (we have recently implemented a validator for the CIF specification). Assuming that chemists will never agree on a single ontology CML supports an arbitrary number of

namespaced dictionaries. This allows different sub-communities to create their own conventions and, where agreed, to move ontological entries to a higher level.

- XML-aware software. The importance of this cannot be over estimated. We use the IETF mantra - "rough consensus and running code". Too many specifications are written in advance of implementations and many turn out to be too complex, both for implementers and the community. The full range of software must include authoring tools, editors, validators, in-memory data models, renderers, and conversion to and from legacy. We have to accept that legacy will be common for at least a decade. In chemistry the commercial software producers have no interest in XML and virtually all the CML-aware software been produced in the Blue Obelisk, which now covers almost all the requirements. Advocacy is through viral implementation, not politics, and a particularly effective tool is the FoX library which outputs CML from many widely used computational codes (it does require that source code is accessible and editable, which is often not the case). There will almost certainly be a need for intelligent client-side software - current browsers are unsympathetic to domain XML (MathML and SVG are rare exceptions and even then give problems). Our community is therefore developing Bioclipse (based on Eclipse) as a universal scientific client, with emphasis on chemistry, bioscience and XML infrastructure.
- Repositories. Most repository software has little support for compound documents and we are using the METS protocol to package ours. We must not assume that legacy data, even if not binary, will be readable in the future. It is also critical that all software is Open Source so it can be maintained independently of manufacturers.
- Metadata and Discovery. Formal metadata are only useful if entirely populated by machine (as human crafting is both expensive and variable). This is possible for provenance, rights, formats but is harder for discovery and navigation. We are committed to exploring free-text as the primary source of discovery metadata along with automatic generation of chemistry-specific indexes (such as the IUPAC InChI for molecular structure). We have recently developed OSCAR3 to discover and retrieve chemical entities in text and its recall/precision can be as high as 90%. It is being extended to other chemical concepts and contexts and in the new JISC-funded SPECTRa-T project will examine chemistry theses and extract metadata as RDF-based SKOS (Simple Knowledge Object System). In this way the documents themselves will define the metadata most valuable for discovery.
- Quality and validation. The quality of hypopublished datuments can vary enormously. As an example the CIFs published by Acta Crystallographica go through a thorough machine validation (CheckCIF), are reviewed by humans, and generally have no syntactic errors. There are a few semantic errors and omissions. By contrast many other journals publish the authors' CIFs without revision. These are unfortunately sometimes syntactically invalid and often have inconsistent semantics (e.g. using non-existent dictionary entries). It is clear that many are hand-edited.

We stress that no data set is 100% correct. It is often assumed that commercially aggregated resources add and ensure quality, but there is much variation. Our OSCAR1 tool checks and extracts data from synthetic chemistry papers and we have found that almost all contain errors of some sort, many serious. Moreover annotation is not absolute - our human-human survey shows that 90% agreement on chemical entity identification is probably the maximum achievable.

Therefore data-driven science must accept the likelihood of undetectable error. In a study of 250,000 compounds from the National Cancer Institute we found that ca 1-2% did not make chemical sense as represented, i.e. they could not be computed in theoretical

chemistry codes which effectively acted as a high-throughput validator. In one study we have compared theoretical and experimental geometries and found that experimental errors were more prevalent than we had assumed. We could devise a protocol, based on the CIF metadata, which allowed us to remove likely problems from the dataset. After this was done, the agreement between theory and experiment was excellent and has started to highlight real (but small) problems with the theory. Where data contains internal constraints or well-defined ranges of values (as in crystallography) it is often possible to detect and remove suspect datasets.

- Scale. This inevitably brought us problems. Each successful data filter tends to reveal problems of lower frequency. With large data sets it is impossible to eyeball all likely problems (a 0.1% error on 250,000 entries gives 250 problems to inspect). Systematic computation (as on 250K entries) is likely to throw up new program bugs - we reported more bugs in MOPAC than the rest of the world had found in 5 years. If the workflow is not well designed it will certainly throw up problems of scale. We had several - the most serious being that we did not have separate identifiers for molecules and jobs.
- Complexity. All our studies have been on well-understood ("homogeneous") data with consistent metadata between all members. In some cases we have tackled mashups between different homogeneous data types. Thus we have used text-mining (OSCAR3) to extract chemical identities from Pubmed abstracts and link them to other data sets (e.g. Pubchem). At this stage we would not wish to tackle problems where machines had to interpret relationships between objects of variable types in datuments of variable structure. Until there is greater experience on machine analysis of homogeneous data the problems listed above will intrude into the analysis of complex systems.

SPECTRa, CrystaLEye and eCrystals

The JISC-funded [SPECTRa](#) project evaluated the feasibility of capturing experimental data directly into repositories. We confirmed the high loss of data and discovered the relatively high ignorance of repositories and open access in chemistry. The project then created implementations for repositing crystallographic data, computational chemistry and analytical data (spectra). In close cooperation with Southampton we hoped to adapt their eCrystals software to a portable generic workflow for departmental crystallographic services. This proved impossible - the crystallographers had their own institution-specific paper-based workflow and were not willing to change. Moreover the concept of "ownership" of data was critical. In some places it "belonged" to the chemist, in others to the crystallographer.

They were keen to have an archive but not prepared to reposit directly into an Open archive. We therefore created a separate "dark archive" into which the data could be deposited in escrow. At a later stage data might be transferred to an open archive either automatically or through triggered human actions. We recognised a "golden moment" when the experiment was essentially complete (and the raw data unlikely to need revision) when data could be easily reposit in a few minutes. The same philosophy was adopted for computational chemistry and spectroscopy. In all cases we encountered problems with legacy file types and this aspect should never be underestimated.

We do not believe that human aggregation and curation of data will remain viable. Nick Day in our group has built an automatic aggregator ([CrystaLEye \(CMLCrystBase\)](#)) for published crystal structures which scrapes all legally allowed CIFs from journal sites at time of publication. We now

have 60-100 K structures from 5+ years of the major journals (except those like ACS, Wiley and probably Elsevier which hide or copyright factual information). CrystalEye automatically validates all data, creates CML, generates metadata and creates a wide range of derived data (such as molecular fragments). Moreover it generates several thousand RSS feeds which can alert subscribers (Open, of course) to a great variety of precise chemical information. For example the Cu-N feed, coupled with the Acta Cryst feed would alert a subscriber whenever a new Cu-N bond appeared in a structure published in Acta Crystallographica. We are now adding social computing facilities. (e.g. Wikis and annotation) so that the world can help clean and critique the data.

As 80% of crystallography (and 100% of comp chem) is never published the combination of SPECTRA reposition and re-use through CrystalEye offers a major opportunity to increase the quantity and quality of chemical data. Leading crystallographic departments are starting to adopt SPECTRA and will expose their open repositories - it is then straightforward to aggregate these (e.g. through RSS or OAI-PMH) into larger knowledgebases. This technology will be available to the partners in the JISC-funded eCrystals program led by Southampton and currently exploring the possibility of a global federation of repositories.

Chemistry forms an ideal discipline for exploring new science repositories. It is well understood, homogeneous across the world and of universal importance and impact. I shall be delighted to discuss how the technology and practices can be disseminated.

Summary

The major problem is people - the restrictive and destructive practices of publishers and the ignorance or apathy of funders, learned societies and academia (at all levels) to the re-use of Open Data. Until this is addressed, data-driven science will be restricted to a few enlightened subjects such as physics and astronomy and parts of bioscience. If that can be changed, and control is wrested back by the scientists, semantic publishing will be technically possible in a number of subjects. This could then generate an understanding of the role of ontologies which will be needed to address complex objects.

Notes

1. this lecture contains a number of key features of modern data-driven science including data cleaning (assessment of quality by adherence to patterns), the need to try out different mathematical formalisms and the need for updating.
2. The publishers' archive will not allow you to actually read the important part of the paper unless you are a subscriber. I do not, of course, have the original manuscript and I can't reproduce it without violating copyright. But I did draw the diagram by hand and surely I have the right to reproduce it
3. see e.g. http://www.wiley-vch.de/contents/jc_2111/2005/f400616_s.pdf for an example of "copyrighted facts".
4. There is no use of the CCDC data in the work reported here.
5. I particularly thank my friends Henry Rzepa (CML), Joe Townsend (OSCAR1), Nick Day (CrystalEye), Andrew Walkingshaw (Golem), Toby White (FoX), Peter Corbett (OSCAR3), Jim Downing, Peter Morgan, Alan Tonge (all SPECTRA), many in Earth Sciences (Cambridge) and the Blue Obelisk ("Open Data, Open Standards, Open Source").