How do we make Scholarship Semantic?

Peter Murray-Rust,

Unilever Centre for Molecular Sciences Informatics, Department of Chemistry, University of Cambridge, CB2 1EW, UK

This paper is an account of an invited keynote to the SePublica Workshop of the ESWC 2013 meeting.

I am grateful for the opportunity to comment on the opportunities and challenges of using semantic approaches in scholarly publishing. ("Scholarship" covers many fields of endeavour and should extend beyond the ivory tower of academia.) We were asked to provide crazy ideas and a polemic, where arguments are backed by personal conviction as much as proven experience. The primary aim of a polemic is to galvanize people into action and the ideas in this paper were aired as several blog posts shortly before the SePublica meeting.

There are very few examples of semantic publishing because the present culture militates against it. If an author sends a semantic manuscript to a journal it will almost certainly be rejected or dumbed down to a more primitive format. To be fair, proper publication requires considerable work from the journal editors and to get semantic benefit, the reader probably has to have special functionality installed. There have been a number of one-off attempts to publish semantically (including some of my own) but they haven't "caught on".

So for most of us, including the readers of this article, semantic publications are an act of faith. We believe them to be valuable, and that when the literature is semantic a brave new world will emerge. I had the privilege of hearing TimBL at CERN at WWW1 in 1994 and he changed my life. The picture of a semantic world mirroring the human and physical world was immediately obvious and imperative. The current problem is that we know it takes a revolution which is not only technical, but cultural, and it is surprising how slow it is proving.

In this polemic I suggest that we limit the level of semantics to simple concepts, (some similarities to TimBL's 5-stars of data):

- We have **to be a community**.
- We have to **identify things** that can be described and on which we are prepared to agree.
- We have to **describe things**
- We have to **name things**
- We have to be able to **find things** (addressing)

All of this is sellable to those who use the web – we don't need formal logic and Aristotelian ontologies. We need identifier systems, ideas of objects and classes, and widely distributed tools. DBpedia and Schema.org are good enough to start with. If this is all that we manage to introduce to scholarly publishing that will be a major success.

So why are semantics important for scholarly publishing? At the most basic level it is about control. The people who control our semantics will control our lives. Semantics constrain the formal language we use and that will constrain our natural language. We humans may not yet be in danger of Orwell's Newspeak but our machines will be. And therefore we have to assert rights to have our say over our machines' semantics.

With a few exceptions I have lost faith in the ability of scholarly societies to act as leads in information development. The problem is that many of them are also major publishers and if they, not us, decide how we are able to express ourselves then it will be based on cost-effectiveness and control, not on what we believe is correct.

The major task for SePublica is to devise a strategy for bottom-up Open semantics. That's what Gene Ontology did for bioscience. We need to identify the common tools and the common sources of semantic material. And it will be slow – it took crystallography 15 years to create their dictionaries and system and although we are speeding up we'll need several years even when the community is well inclined. Semantics have to be Open, and we have to convince important players that it matters to them. Each area will be different. But here are some components that are likely to be common to almost all fields:

- Tools for creating and maintaining dictionaries
- Ways to extract information from raw sources (articles, papers, etc.) – that's why after the SePublica meeting we are "Jailbreaking the PDF".
- Getting authorities involved
- Tools to build and encourage communities
- Demonstrators and evangelists
- Stores for our semantic resources
- Working with funders

A small number of publishers do adopt this approach. I single out the International Union of Crystallography, which for many years has developed machine-understandable dictionaries for its discipline. Anyone publishing in their journals must submit in CIF format (Crystallographic Information Framework) which uses a name-value approach for data and a LaTeX-inspired approach for text. Papers are reviewed both by humans and machines, and the machines very frequently discover poor, bad, and sometimes fraudulent science. The final paper is automatically typeset from the (human-reviewed CIF). This is sufficiently compelling that a forward-looking publisher or society should surely be impressed.

So, apart from the political backdrop, why are semantics important?

- **They unlock the value of the stuff already being published**. There is a great deal in an a current PDF (article or thesis) that would be useful if it were semantic. Diagrams and tables are frustrating shadows of Plato's cave. Mathematical equations could be brought alive and computed in real-time by the reader ("plot that data, integrate the areas under the curves and compare with the equations"). Chemical structures can be extracted and their properties computed using Schroedinger's equation. Even using what we have today converted into semantic form would add billions.
- **They make information and knowledge available to a wider range of people**. If I read a paper with a term I don't know then semantic annotation may make it immediately understandable. What's rhinovirus? It's not a virus of rhinoceroses - it's the common cold. Semantic resolution makes it accessible to many more people.
- **They highlight errors and inconsistencies**. Ranging from spelling errors to bad or missing units to incorrect values to stuff which doesn't agree with previous knowledge. And machines can do much of this. We cannot have reproducible science until we have semantics.
- **They allow the literature to be computed**. Many of the semantics define objects (such as molecules or phylogenetic trees) which are recomputable. Does the use of newer methods give the same answer?
- **They allow the literature to be aggregated**. This is one of the most obvious benefits. If I want all phylogenetic trees, I need semantics – I don't want shoe-trees or B-trees or beech trees. And many of these concepts are not in Google's public face – we have to collect them.
- **They allow the material to be searched**. How many chemists use halogenated solvents? (The word halogen will not occur in the paper so Google can't find it). With semantics this is a relatively easy thing to do. Can you find second-order differential equations? Or Fourier series? Or triclinic crystals?
- **They allow the material to linked into more complex concepts**. By creating a data base of species, a database of geolocations and links between them we start to generate an index of biodiversity. What species have been reported when and where? This can be used for longitudinal analyses – is X increasing/decreasing with time? Where is Y now being reported for the first time?
- **They allow humans to link up**. If A is working on *Puffinus Puffinus* in the northern hemisphere and B is working on *Puffinus tenuirostris* in Port Fairy Victoria AU then a shared knowledge base will help to bring the humans together. That also happens between disciplines – microscopy can link with molecular biology with climate with chemistry.

None of this requires inferential logic. A hybrid mixture of terminologies, identifiers, data structures can be glued together into domain-aware systems. Semantics allow smart humans to develop communal resources to develop new ideas faster, smarter and better.

How do we make this happen? What I suggest may seem daunting but it's a smaller scale than the already successful Wikipedia. It is critical we act now**,** because Semantics/ContentMining is now seen as an opportunity by some publishers to "add value" by building walled gardens. If semantic enhancement is done by publishers then it will be very small, heavily controlled and expensive. So we must build something better, fully Open (i.e. no restrictions on re-use), and demonstrably valuable. It took one person to launch Open Street Map – and for many of us it's the gold standard of modern semantic maps – we can do the same for semantic publications.

We must create coherent communities. In the past this would be based on learned societies, but that will no longer work – we need a bottom-up approach. DBpedia is a beacon of how to create a world semantic resource – we must find ways of scaling this to disciplines it doesn't currently serve. It's conceivable that a mixture of the Wikimedia culture with public organizations (e.g. Galleries, Libraries, Museums, Archives) becomes the semantic core of scholarly publications.

Some semantic visions we should now be able to sell:

- **Give power to authors**. Authors are frustrated –many understand the need for annotations and are disenfranchised. Tools are becoming easier to deploy and we can create a semantic symbiosis for authors. For example a "species-checker" or "chemical checker" could be built into an authoring tool, so that the information is captured but the best person to understand it – the author.
- **Discover, aggregate and search** ("Google for science"). Search engines do not and will not support scholarly semantics. I cannot search Google for the details of numerical quantities, chemicals or species. It's relatively cheap and simple to do much of this – we indexed 500,000 reactions from US patents to a higher semantic quality than elsewhere.
- **Make the literature computable**. If we can compute parts of a paper we read, or aggregate many papers and map-reduce them, huge visions open up. For example we could search for all compounds in the literature which might sequester Carbon dioxide and compute their properties. This is a well-defined task and relatively straightforward to do.
- **Smart "invisible" capture of information.** If we interact with information (by creating it or reading it) then machines can also read and compute it. We would use semantic because they help us, but their results would be useful to the world. We use Bitbucket/Git because it helps us produce better programs, but a by-product is the archival for the whole world. Tools to help authors can also capture information seamlessly.

A critically important thing we can do now is to create a single-stop location for tools. Many new tools are being created and libraries such as Apache, Guava or UIMA contain much of what we need for simple conversion of raw material to semantic form. Key aspects are

- Common approach to authoring
- Crawling tools for articles, theses.
- Converters of PDF and Word to XML or XHTML
- Classifiers
- NLP tools and examples
- Diagram interpretation (e.g. extraction of data from graphs or phylogenetic trees)
- Logfile hackers (much output is "FORTRAN"-like and semi-structured). We can convert this to semantic form or annotate it automatically
- Semantic repositories
- Abbreviations and glossaries
- Dictionaries and dictionary builders

Scholarly publishing must change dramatically if only because the world is changing so fast. The present "mainstream" (traditionally closed-access) publishers cannot continue here as their model is to possess and control re-use of information, not to enhance it. The new semantic world will not only be formally Open but will think that way. Among the organizations (deliberately unnamed) that I expect to be responsive to the ideas expressed here are:

- Funders of science

- major Open publishers
- Funders of social change
- Open publication advocacy organizations
- (Europe)PMC
- Wikipedia
- GLAM
- Governments and NGOs

Where can a reader start? If they are in an organization, they can examine how semantic publication can enhance its business. If they are individuals they can build semantic tools and semantic resources.

And a remarkable example of the possibility was given in the post-SePublica hackathon (“Jailbreaking the PDF”) – a collection of working tools and examples that show that current PDFs can often be transformed to semantic form.