Use Case 4 in Media - Research
Science Publishing and The Semantic Web, Or why are you reading this on paper?

KW Partner: Elsevier

1. Overview

Challenge
The exploration of entity-relationship-based authoring tools

Solution
investigation of more sophisticated knowledge models and the development of a platform for scientific discourse that better suits the connected, virtual environment in which science is done today.

Why a Semantic solution
To host our XML, we have been working towards a Web Services architecture, which allows for more distributed development of our online products, and compatibility with outside systems.

Key Business Benefits
It would be interesting to allow authors to directly create and manipulate entities and relationships to represent the point they are making. These entity-relationship-entity (or entity-property-value) triples could make up the backbone of an argument.

Business Partners

Keys components

Existing Software
RDF

Research and development

Technology locks

1. Step One: Getting Journals Online.

Electronic science publishing started with the widespread use of SGML, which increased production efficiency, improved the quality assessment process, and allowed the reuse of content for different delivery platforms [1]. In the early nineties, Elsevier developed one of the main DTDs for science [2] which helped shape and standardize the scientific article format. In the beginning, the creation of SGML and the creation of print files for publishing were independent processes, and print publication had precedence. At this time, online journals were created in a variety of ways, by a plethora of different tools and at different locations. The SGML was not always used; in some cases html files for articles were more or less handcrafted and posted individually. However, over the course of the nineties this started to change, and Science Direct, Elsevier's main online journal platform, was launched in 1997.

Together with the DOI [3] and Crossref [4] foundations, Elsevier helped pioneer a system by which every paper has a unique identifier. Today, most article references can be resolved to connect directly to the document cited.

Currently, our journals are produced directly from XML using a DTD that includes XML-derived standards such as MathML and XLink [5]. Multimedia components are accepted in a wide range of standards, including MPEG, QuickTime, and SVG [6].

Of course, "the great thing about standards is that there are so many to choose from".

To learn about and help establish data standards, (Reed-)Elsevier actively participates in several key standards bodies, including the W3C, OASIS, (N)ISO, and other, more domain-focused efforts such as the Medbiquitous consortium [7]. To host our XML, we have been working towards a Web Services architecture, which allows for more distributed development of our online products, and compatibility with outside systems.

But once all the data is available electronically, the trouble is not over: in a way, this is where it begins. First of all we face the "information infarct": too much time is spent on searching data, and still relevant information is missed. Secondly, each content source has its own interface, and users in information-dependent professions (such as scientists) quickly tire of using many different search systems. Therefore, data integration is a very important topic for science publishers. As an example, in the DOPE project an RDF architecture was built to disclose heterogeneous data sources via an ontology based on the EMTREE thesaurus [8]. But even when content is available through a single interface, it remains impossible to oversee developments in an area of science by doing a search and looking at the results list. To make sense of large volumes of data, Elsevier is involved in various data mining projects, which aim to extract entities and/or facts from large collections of data. However, there are limits to the accuracy with which entities and relationships can be detected. Which begs the question: if in the end we want to *mine* the data, why do we *bury* it in the first place?

2. So Now What? Entities, Relationships and Discourse.

To date, the fundamental unit of information exchange in scientific communication remains the *paper*, which, as the word indicates, is a document format deeply grounded in a paper world, optimized for paper consumption, and written in ways that were developed when we still actually wrote on paper. The dominant mode of scientific communication is still the production and consumption of articles (such as this one). Apart from reference linking (to more papers!), there is no relationship between the knowledge posited in the paper, and that already in existence. So how can we make scientific publications better suited for the online, connected world in which scientists work and live?

It seems that the authoring phase of a document is a unique opportunity to graft the author's new insights onto the existing information space. It would be interesting to allow authors to directly create and manipulate entities and relationships to represent the point they are making. These entity-relationship-entity (or entity-property-value) triples could make up the backbone of an argument. The publication could further be augmented, for instance, by modular text elements such as proposed by Kircz et al [9].

It would also be interesting to investigate whether these triples can be supported by other triples representing subunits of discussion. Can existing rhetorical theories help us to represent a publication as sets of nesting triples? Could usable authoring tools be constructed from these models? If we had such tools, a new publication would not stand on its own, needing software and human brainpower to relate it to existing facts or entities. Instead, new work would be directly connected to the existing entity relationship space of a research field by the author during writing. It would be available to the reader as a related set of relationships, connected to existing knowledge.

As an example, molecular biologists work with biological entities, such as cell lines, species, proteins, genes, etc. They are used to access virtual representations of these objects online through sites such as Genbank [10] and Swissprot [11].

Ontologies of genes and genetic functions are provided by the Gene Ontology project [12]. The development of such knowledge banks and other registries form a very active field of research, and initiatives such as the Life Science ID project [13] make these fully Semantic-Web compliant. An authoring tool could use these developments by offering authors an environment to forge well-defined biological relationships between uniquely identified biological entities. Such tools can be envisioned in other fields as well: chemical, geographical, and astronomical objects could be linked, discussed and connected in an authoring environment. More generically, elements of scientific discourse such as proofs, statements, disagreements etc. could be manipulated by the author to make a point, and represented in context to the reader.

There are a number of initiatives that could be fruitful starting points. The 1962 work of the visionary Doug Engelbart has still never seen a realization in today's sadly inadequate browser technology. Engelbart's Augment system was a browser *and* a text editor *and* a programming interface, and could do all these things collaboratively, as a matter of course [14]. There are a number of interesting projects ongoing to develop authoring and editing tools, including ClaiMaker and ClaiMFinder, a set of "sensemaking tools" that allow the creation, retrieval and browsing of scholarly debate based on claims [15]; TRELLIS, "an interactive environment that allows users to add their observations, opinions, and conclusions as they analyze information by making semantic annotations to documents and other online resources" [16]; S-CREAM, which "allows for creation of metadata and is trainable for a specific domain"[17]; the WiCK project, whose aim was "to produce a novel writing tool, which is underpinned by an enhanced knowledge structure and hypermedia design model." [18]; the semantic web authoring and annotation page [19] lists a number of other tools. But why are all the papers about these projects published as linear narratives, in pdf? What aspects of scientific writing cannot be modeled into sets of triples, or at least as annotated hypertext? Can we come up with knowledge models that do work for science writing, facing all the practical, political, emotional and technical issues that go into producing a scientific paper; - or are we doomed to mine narratives for ever?

There are several issues that need to be addressed if we do wish to move towards more connected method of scientific publishing:

Versioning. One of the main drivers behind modular publishing is to better deal with different versions of information. In the current system, papers pertaining to a piece of research are often largely overlapping; e.g. the description of methods, experimental setup, or well-known theories is repeatedly published in different articles ("salami-publishing"). If we move to a more elemental method of publication, each element can be independently reused or updated. Ideally, this helps structure the information space and identify what is new in a publication – but it does present us with issues regarding versioning that need to be addressed.

Identification. Once we have relationships between entities authored: how can we niquely identify them? Who keeps track of what has been argued, and how are intellectual property rights assigned to these links?

Validation. Published papers still form the main basis for validating scientific research, and more specifically, the researcher who writes them. Papers written for journals, edited by editors, and reviewed by peers still form the main way in which rank, tenure etc. are bestowed. Could a postdoc be judged on the merit of a set of RDF triples he or she has created? It would be very interesting if the Semantic Web community could take the lead in the discussion of alternative ways of judging and bestowing scientific credentials.

Non-textual elements. Mathematical, graphical, tabular, photographical and other representations are common elements of scientific publications. If we want to manipulate them, we will need identify, link and reproduce elements within tables and figures. Discussions with data standards bodies will need to take place to ensure future concordance.

If we are able to find at least partial solutions to these problems, the creation of tools to write, edit, referee and read articles that are compliant with Semantic Web standards and philosophies seem promising areas of development. Knowledge models need to be developed that allow us to represent the fullness of scientific discourse.

And we need to find the hidden drivers that will make not only to develop these tools, but actually *use* them for communicating research results. Collectively, we could make the way we publish science compatible with the way we practice science: in an interactive, connective setting, that evolves as we do.

References

1. See e.g.: J. G. Kircz and J. Bleeker. The use of relational databases for electronic and conventional scientific publishing. Journal of Information Science 13 (1987) 75-89.
2. See http://www.info.sciencedirect.com/librarian_help/dtds/index.shtml
3. http://www.doi.org/
4. http://www.crossref.org/
5. For DTD 5.0, see http://support.sciencedirect.com/sgml/dtd50/art501.dtd.txt
6. For instructions, see http://authors.elsevier.com/ArtworkInstructions.html?dc=AI1
7. http://www.medbiq.org/
8. H. Stuckenschmidt, F. van Harmelen, A. de Waard, T. Scerri, R. Bhogal, J. van Buel, I. Crowlesmith, Ch. Fluit, A. Kampman, J. Broekstra and E. van Mulligen, Exploring Large Document Repositories with RDF Technology: The DOPE Project, IEEE Intelligent Systems, 2004,19 (3), pp. 34—40.
9. F.A.P. Harmsze, M..C. van der Tol and J.G. Kircz, A modular structure for electronic scientific articles. In: P. de Bra and L. Hardman (eds), Computing Science Reports, TU Eindhoven (1999), Report 99-20. pp. 2-9.
10. http://www.ncbi.nlm.nih.gov/Genbank/
11. http://www.ebi.ac.uk/swissprot/
12. http://www.geneontology.org/
13. See http://xml.coverpages.org/lsid.html for a good introduction to LSID.
14. D.C. Engelbart, Augmenting Human Intellect: A Conceptual Framework, Stanford Research Institute, 1962 (!), http://www.histech.rwth-aachen.de/www/quellen/engelbart/
15. http://claimaker.open.ac.uk/
16. http://trellis.semanticweb.org/
17. See e.g. http://eprints.aktors.org/124/
18. http://wick.ecs.soton.ac.uk/
19. http://annotation.semanticweb.org/