**Use Case: On the need for embedding markup at authoring time**

**Key Concepts:**

* Automatic or semi-automatic creation of semantic metadata from literature versus author enablement of embedding and marking semantic metadata. I believe that the first is un-attainable, at least in the near-term (see Jon Bosak article below), whereas much benefit can be gained from the second.
* “The current scientific literature, were it to be presented in semantically accessible form, contains huge amounts of undiscovered science,” Peter Murray-Rust, Data-driven science: A Scientists’s view: NSF/JISC Repositories Workshop, 2007.
* The focus should be on presenting science literature in semantically accessible form versus trying to extract semantic data from traditional papers which are generally inaccessible (or very inconvenient).

[**http://www.nature.com/nature/debates/e-access/Articles/bosak.html**](http://www.nature.com/nature/debates/e-access/Articles/bosak.html)

**Text markup and the cost of access**, by Jon Bosak, Sun Microsystems

Jon Bosak organized and led the working group that created XML. He subsequently served for two years as chair of the XML Coordination Group of the World Wide Web Consortium. He is a founding member of OASIS, the Organization for the Advancement of Structured Information Standards, and he chaired the committee that developed the OASIS process for the definition of industry-specific XML markup standards.

The following are excerpts from the above article in Nature. Essentially Jon advocates investing in human insertion and identification of key semantic details. Specifically, he is recommending a two-pronged approach:

1. Form a wikipeidia-like volunteer staff that reviews and adds markup to published literature. Jon says: “… but the notion that we can noticeably improve upon full-text searching without thorough standardization and a continuing investment of consistent, expert labour is, in my opinion, a chimaera.”
2. Enable author-entered metadata. “Anyone capable of writing an acceptable scientific paper can master a consistent metadata format and can learn the subject-matter classification of their specialty.”

The following are tools and techniques to assist with the insertion, identification and extraction of semantic metadata.

<http://ucsdbiolit.codeplex.com/>

**Word Add-in For Ontology Recognition**

Microsoft External Research’s goal with this project is to enable communities who maintain ontologies to more easily experiment and to enhance the experience of authors who use Microsoft Word for content creation, incorporating semantic knowledge into the content. This add-in should simplify the development and validation of ontologies, by making ontologies more accessible to a wide audience of authors and by enabling semantic content to be integrated in the authoring experience, capturing the author’s intent and knowledge at the source, and facilitating downstream discoverability.

The goal of the add-in is to assist scientists in writing a manuscript that is easily integrated with existing and pending electronic resources. The major aims of this project are to add semantic information as XML mark-up to the manuscript using ontologies and controlled vocabularies (from the National Center for Biomedical Ontology) and identifiers from major biological databases, and to integrate manuscript content with existing public data repositories.

<http://broadcast.oreilly.com/2009/03/microsoft-and-science-commons.html>

Microsoft and Science Commons Team Up To Add Semantic Content to Online Science

This article refers to a Microsoft Word plug-in “… that will allow scientists to mark up their papers with scientific entities directly.” The quotes are from John Wilbanks, VP of Science for Creative Commons.

"The scientific culture is not one, traditionally, where you have hyperlinks," Wilbanks told us. "You have citations. And you don't want to do cross-references of hyperlinks between papers, you want to do links directly to the gene sequences in the database."

Wilbanks says that Science Commons has been working for several years to build up a library of these scientific entities. "What Microsoft has done is to build plugins that work essentially the same way you'd use spell check, they can check for the words in their paper that have hyperlinks in our open knowledge base, and then mark them up."

**<http://sifnos.ilsp.gr/Poeticon/ilsp_talks/liakata-present.pdf>**

Towards automated understanding of scientiﬁc papers

Dr Maria Liakata, Department of Computer Science, Aberystwyth University, proposes a tool, SAPIENT, used to auto-separate sentences and allow experts to mark up the document at the sentence level. References: “Paper zoning and sentence classiﬁcation in papers,“ ([Teufel and Moens 2002], [Wilbur et al.2006],

**References**

[**http://www.nature.com/nature/debates/e-access/Articles/bosak.html**](http://www.nature.com/nature/debates/e-access/Articles/bosak.html)

**Text markup and the cost of access**, by Jon Bosak, Sun Microsystems

Additional Excerpts from Jon Bosak.

In industrial contexts, it is commonly estimated that marking up technical material adds about 40% to the work of writing the text.

A body of information inconsistently provided with metadata will return dangerously misleading results to a system expecting that metadata. For example, a sophisticated query keyed to search for papers in which a certain term appears tagged as a chemical reagent will find papers in which that term has been given the standardized tag for chemical reagents but not those whose authors failed, due either to ignorance or to a lack of resources, to embed this additional information in the text. Partial adoption of the tagging scheme means that an intelligent search engine asked to aggregate papers on this basis will return a number of papers answering the query but will have no way to list the papers that actually contain the target information that has not been tagged appropriately. Such a system is much worse than a simple full-text search, because it gives a user the false impression that a precise and comprehensive report has been generated. In a field such as medical research, misleadingly incomplete query results of this kind could have disastrous consequences.

So although XML markup, properly applied, can greatly increase the utility of scientific texts, the cost of using it effectively is not trivial.

The principle here is quite simple: the more information we add to a document, the better use we can make of it. But the information has to come from us; it is not going to come from the computer. XML does nothing to change this basic fact. The potential for increased access to information made possible by the new XML-based technologies is indeed spectacular, but the notion that we can noticeably improve upon full-text searching without thorough standardization and a continuing investment of consistent, expert labour is, in my opinion, a chimaera. I believe that the big question in making scientific texts more accessible is how to provide the additional expert labour.

Anyone capable of writing an acceptable scientific paper can master a consistent metadata format and can learn the subject-matter classification of their specialty. This indicates that we should look to the authors of scientific articles for a considerable portion of the extra work needed to enable better online access. In particular, we should be able to expect authors to use a standardized XML markup for scientific articles and to expect authors or their editors to classify the work properly in one or more standard registries. But this would still leave the question of how to define and implement an online catalogue system that would allow humans and search tools to assemble variously related texts.

One cost-effective way to establish and maintain online catalogues is suggested by the Open Directory Project. This self-sustaining directory of the World Wide Web is driven by classification data entered by the creator of each web page. Management of the registry is accomplished through a system of volunteer area editors, each responsible for the organization and maintenance of some small part of the overall taxonomy. One can imagine a directory such as this for scientific literature that would be created and maintained by a distributed network of domain experts.

1. Adopt a standard XML markup for scientific texts within each specialty and require all authors working in that specialty to adhere to it.

2. Adopt a standard format for bibliographic data and require all authors and publishers to provide such data in texts made available in electronic form.

3. Institute a collaborative project to catalogue scientific papers using a distributed system based on the labour of volunteer editors.

[**http://text0.mib.man.ac.uk/software/facta/main.html**](http://text0.mib.man.ac.uk/software/facta/main.html)

**FACTA - Finding Associated Concepts with Text Analysis**

**http://bioinformatics.oxfordjournals.org/cgi/reprint/btn469?ijkey=g6i4qMXGGdyGkEc&keytype=ref**

**FACTA: a text search engine for finding associated biomedical concepts, 2008**

**Thursday, March 12, 2009**

**Semantic markup tool for Microsoft Word uses Science Commons ontologies**

Word Add-in For Ontology Recognition is a new, free/open source add-in for Microsoft Word 2007 to add semantic information as XML mark-up using ontologies and controlled vocabularies, and to integrate manuscript content with existing public data repositories. See the Microsoft press release or coverage at O'Reilly; from the latter:

... [John] Wilbanks says that Science Commons has been working for several years to build up a library of these scientific entities. "What Microsoft has done is to build plugins that work essentially the same way you'd use spell check, they can check for the words in their paper that have hyperlinks in our open knowledge base, and then mark them up." ...

[**http://ucsdbiolit.codeplex.com/**](http://ucsdbiolit.codeplex.com/)

**Word Add-in For Ontology Recognition**

**Summary**

Microsoft External Research’s goal with this project is to enable communities who maintain ontologies to more easily experiment and to enhance the experience of authors who use Microsoft Word for content creation, incorporating semantic knowledge into the content. This add-in should simplify the development and validation of ontologies, by making ontologies more accessible to a wide audience of authors and by enabling semantic content to be integrated in the authoring experience, capturing the author’s intent and knowledge at the source, and facilitating downstream discoverability.

The goal of the add-in is to assist scientists in writing a manuscript that is easily integrated with existing and pending electronic resources. The major aims of this project are to add semantic information as XML mark-up to the manuscript using ontologies and controlled vocabularies (from the National Center for Biomedical Ontology) and identifiers from major biological databases, and to integrate manuscript content with existing public data repositories.

**http://sifnos.ilsp.gr/Poeticon/ilsp\_talks/liakata-present.pdf**

Towards automated understanding of scientiﬁc papers

Dr Maria Liakata Department of Computer Science, Aberystwyth University

mal@aber.ac.uk

She proposes a tool, SAPIENT, used to auto-separate sentences and allow experts to mark up the document at the sentence level. References: “Paper zoning and sentence classiﬁcation in papers,“ ([Teufel and Moens 2002], [Wilbur et al.2006],

[Shatkay et al. 2008],

[Hirohata et al. 2008],[Medlock and Briscoe 2007])