Development of a Semantic Structure for Scientific Articles

A. de Waard, Advanced Technology Group, Elsevier, Amsterdam
H. van Oostendorp, Centre for Content and Knowledge Engineering, University of Utrecht

We propose the development of a semantic structure for scientific papers, for use by the author. This structure should enhance the integration of knowledge contained within the paper, and improve the usability of scientific articles in a computer-centered environment.

Introduction

Scientists are increasingly unable to process the ever-increasing flood of scientific literature that surrounds them. Biomedical literature, for instance, grows by over 500,000 publications each year (Cohen, 2005). In a recent study on user needs among British archaeologists, 71% of the respondents felt that information was produced of which they were unaware (Jones, 2001). Next to problems in accessing one's own field, it becomes more and more difficult to access adjacent domains of science. Furthermore, scientists do not only want to know what publications contain specific words, and how to rank them by relevance, but what *knowledge* is contained within the papers, and how it relates to their existing knowledge. For example, cell biologists might want to know: "What functions of this gene are known?" Astronomers might ask "What radiation patterns have we seen in red-dwarf stars?" or "What theories does this new observation support?" Ideally, a new publication should situate itself within the existing knowledge context of the reader, and show how it affects or alters this context.

There have been many efforts to combat this information overload in science. Abstracts have been developed in the sixties and seventies. Although they are shorter to read, abstracts do not provide a full summary of the work described in the document, nor do they offer any way to integrate the document into the existing knowledge. Metadata is a broad term covering many different types of information, but generally includes the bibliographic reference to a document, and descriptors such as keywords[1]. Metadata helps retrieve an article when descriptive elements (author, title) are known. The main function of a keyword list is to classify the article in a category. But neither provides any direct insight in the *knowledge* conveyed within the body of a scientific paper.

Text mining and information extraction are methods specifically developed to find relevant information in unstructured texts and encode the information in a structured form, like a database record (Couto, 2003). In theory, text mining is the perfect solution to transforming factual knowledge from publications into database entries. However, automatically identifying concepts such as genes and proteins poses many problems; see e.g. Mons (2005) and Cohen (2005). Moreover, computational linguists have not yet developed tools that can analyse more than 30% of English sentences correctly and transform them into a structured formal representation. For this, the papers still need to be handled by a curator (Rebholz-Schuhmann, 2005).

The main problem with automatically extracting information from scientific articles is that the genre of the scientific publication has developed to be an indivisible information unit (see e.g. Bazerman, 1988). The scientific paper is a self-contained narrative, created anew in each iteration, with specific genre characteristics that minimize the potential of identification, content reuse and knowledge integration. All this rhetorical freedom comes at the expense of usability in a computer-centered environment. The linear narrative was fine when we still read and wrote on paper, but the changing (digital) environment in which scientists live and work calls for a changing fundamental unit of communication.

[1] See de Waard & Kircz (2003) for a more extensive discussion of metadata

Our Approach: Semantic Structuring

We believe that the best way to present a narrative to a computer is to let the author explicitly create a rich semantic structure for the article during writing (see also de Waard, 2005). At a high level, this structure will consist of self-contained modular elements or entities, and discourse relationships between such elements (within a text, and between texts). The tension between these self-contained 'knowledge elements' or conceptual structures, and the meaning conveyed in the conventional narrative of the document as a whole, poses an interesting topic of study in terms of both knowledge modeling and rhetoric/discourse studies.

As conceptual structures become the central bearer of information, a set of structured documents can be integrated to form a 'knowledge network', or structured package of related knowledge regarding a topic. This can be envisaged (and modeled) as a network of nodes and relationships, and can be seen to form an incarnation of the 'intelligent data' ideal, which the Semantic Web is meant to enable (Berners-Lee, 2001)[2].

Our starting point will be use to create a working model of a scientific document in two domains: Cell Biology and Archeology. The domain choice is motivated by several factors:

- Cell biology is the single area where most research and development is taking place in terms of text mining, data mining and entity definition. The community is very large, very distributed and the number of publications enormous. The information needs can be clearly defined and the subject matter at hand is well defined – basic entities such as genes, proteins and organisms are all well-catalogued, identified and freely available in electronic format.
- Archeology is chosen partly to be as different from Cell Biology as any field can come up with. It is a relatively small, close-knit field, where data mining is still in its infancy; however, it poses interesting issues of using spatially (GIS) based information and providing electronic access to the objects of study.

The motivation of both domains is further supported by Nentwich (2001), who investigated the degree of "cyberness" of various scientific (sub)disciplines. Since these are two such diverse domains, we hope to be able to generalize our findings to extend to science as a whole.

For the working model, we will create a schema of a scientific document in either RDF or OWL. RDF (Resource Description Framework) is developed by the World Wide Web consortium to be the *lingua franca* for entity-relationship triples on the Semantic Web (W3C, 1999). RDF documents contain information in the form of statements, consisting of a subject, predicate and an object[3]. OWL (the Web Ontology Language) is a language for defining and instantiating Web ontologies (W3C, 2004). Given such an ontology, the OWL formal semantics[4] specifies how to derive its logical consequences, i.e. facts not literally present in the ontology, but entailed by the semantics. This enables the construction of reasoning over distributed documents – and thus allows the construction of knowledge spaces as we envisage them.

To begin with, we must identify entities and relationships that can be reasoned about (and form the basis of the knowledge space, which can be called an ontology[5]). In our model, the entities can correspond to either identifiable objects in the real or virtual world (such as genes, proteins, stars etc.)

[2] 'The Semantic Web is not so much about intelligent agents, but more about stupid agents and intelligent data', Berners-Lee WWW4, Boston, 1995 http://www.w3.org/Conferences/WWW4/Program_Full.htm, personal record.

[3] For instance, a statement could be 'Anita de Waard is the author of this document', where "Anita de Waard" is the subject, "author" is the predicate (relationship) and "this document" is the object; all have URIs uniquely identifying them in cyberspace.

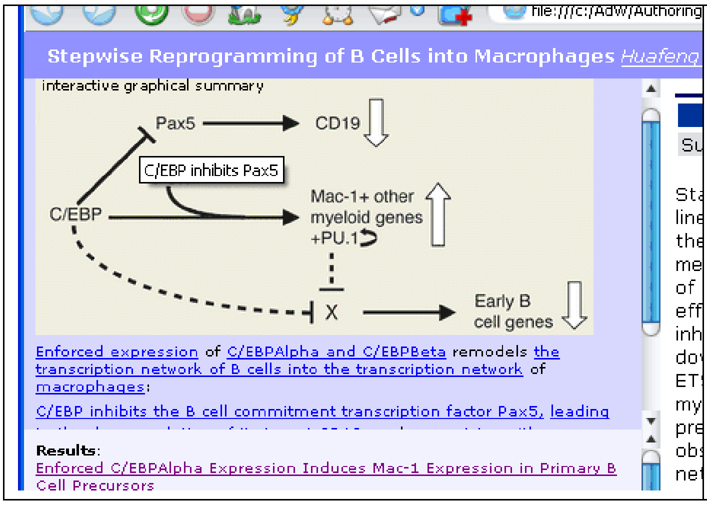
[4] OWL formal semantics: http://www.w3.org/TR/2004/REC-owl-semantics-20040210/

[5] Ontology is not used here as a fancy word for index, but the more classically philosophical sense of 'worldview', or 'describing the kinds of entities in the world and how they are related' (W3C, 2004).

or to discourse elements such as claims, theories, statements etc. The relationships can be thought to represent discourse relations (Uren, 2004).

An Example

For an example in Cell Biology, we can envisage the creation of a “semantic skeleton” of concepts and relationships to form the framework for an article, in this case from the journal Cell[6]:



Stepwise Reprogramming of B Cells into Macrophages *Huafeng*

interactive graphical summary

Enforced expression of C/EBPAlpha and C/EBPBeta remodels the transcription network of B cells into the transcription network of macrophages:

C/EBP inhibits the B cell commitment transcription factor Pax5, leading

Results:
Enforced C/EBPAlpha Expression Induces Mac-1 Expression in Primary B Cell Precursors

Example from Cell Biology. The concepts “C/EBP” and “Pax 5” are identified to be genetic loci, which have a virtual representation in Genbank. The paper postulates an inhibitory relationship between these two. This is given in the graphical summary as a line with a bar at the end, to denote inhibition. The relationship is linked into the article, where a paragraph of text describes the relationship “C/EPB inhibits Pax5”.

The semantic features included here enrich the conventional structure of a scientific article on several levels. They enable the user to integrate the knowledge from this article by, for example, allowing them access to, for instance:

- What does C/EBP Expression lead to,
- Mechanisms of Pax5 Inhibition,

where the statements made here are compared and connected to statements made in other publications.

But defining entities and relationships is not enough. We need to identify other aspects of content which contribute to the knowledge conveyed on a scientific paper. A full study of these aspects is planned for the near future. Some likely aspects to be taken into account include:

- Knowledge contained within the narrative itself, as defined in the field of “narratology”(see Tuffield, 2005 for a series of references)
- Knowledge networks, as defined in discourse theory. This subfield of cognitive psychology concerns ‘processes and strategies involved in constructing representations from textual input’ – see van Oostendorp and Goldman (1999). Kintsch (1998) proposed that ‘knowledge […] is represented in the form of associative networks. The nodes in these networks correspond to concepts or propositions‘ (Ferstl & Kintsch, 1999).
- Rhetorical Structure Theory, which ‘identifies hierarchical structures in text’, ‘describes the relations between text parts in functional terms, and ‘provides a general way to describe the relations among clauses in a text’ (Mann & Thompson, 1988). RST describes ‘spans’ of text as ‘satellites’ and ‘nuclei’, which are related by named relations (RST Website).
- Intertextual relationships between the existing paper and other papers – including discourse relations (such as ‘support’, ‘oppose’, ‘in the spirit of’, etc). For this aspect, we can investigate the practicality of the intertext model model, which represents documents as ‘document nodes’ and ‘intertext predicates’ as links between documents or parts of it (Perfetti, Rouet & Britt, 1999).

[6] This prototype can be viewed at. http://labs.elsevier.com/resources/adw/changingdoc/CellDemo/index.htm

All four fields concern themselves with textual analysis and comprehension, in a psychological, social and linguistic context. And interestingly enough, they all consider texts to contain modular elements that are more or less self-contained (although they bear different names), and have inter-and intra-textual relationships. Support for this model is given by the work of Kircz (1998) on modular structures. Studying and comparing the results of these approaches to textual comprehension and creation seems to offer a very unique opportunity to develop a new model for s scientific article, which facilitates knowledge creation and cross-fertilization.

Next steps

Our project has recently started, and the development of an appropriate structure will be the first priority. To achieve a corpus of articles to test this structure on, we need to select an appropriate authoring and editing environment for scientists to work in. We aim to work with and expand an existing authoring tool (such as e.g. developed by Van Zwol and Callista (2005)) to create an online environment for content authoring. An explicit goal here is also to examine whether authors can create structured submissions, and whether the narrative freedom they need to express their research can be expressed in these explicit structures.

Once a corpus has been created, a series of user tests will be performed to examine whether scientists indeed retrieve more relevant knowledge packages with these newly structured documents, and can make available information relevant to the questions mentioned above.

If this format is indeed successful, it can lead to a new type of publishing, where the end goal of information seeking is not to find a document, but actual *knowledge* on topics defined by the user. It will be interesting to explore business models that follow from this paradigm shift: one can think of selling subscription to information on an entity such as a gene, or a star system, or a field of thought, rather than to a journal.

References

Bazerman, C. (1988). Shaping written knowledge: The genre and activity of the experimental article in science. *Madison: University of Wisconsin Press.*
http://wac.colostate.edu/books/bazerman_shaping/shaping.pdf

Berners-Lee, T., J. Hendler and O. Lassila (2001). The Semantic Web, A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American, May 2001*
http://www.scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&pageNumber=1&catID=2

Cohen, A.M., et al (2005). Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts. *BMC Bioinformatics. 2005; 6: 103.*
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1090552

Couto, F.M., M. J. Silva, P. Coutinho (2003). Improving Information Extraction through Biological Correlation. *Proceedings of the European Workshop on Data Mining and Text Mining, 2003, Dubrovnik, Croatia.*
http://www.informatik.hu-berlin.de/%7Escheffer/publications/ws03proc/couto.pdf

De Waard, A. (2005). Science Publishing And The Semantic Web, Or: Why Are You Reading This On Paper? *European Conference on the Semantic Web 2005, Industry Forum, Alain Léger (ed.)*
http://labs.elsevier.com/resources/adw/changingdoc/papers/deWaardECSW2005.pdf

De Waard, A. and J.G. Kircz (2003). Metadata in Science Publishing. *In: P. de Bra (ed.), Proceedings Conferentie Informatiewetenschap 2003. TUE, 20 November 2003. CS-Report 03-11. pp.73-84.*
http://www.kra.nl/Website/Artikelen/Metadata-wi-2003.htm

Ferstl, E. C. and Kintsch, W. (1999). Learning From Text, in *H. van Oostendorp & S. Goldman (Eds.), The Construction of Mental Representations during Reading. Mahwah, NJ: L. Erlbaum Ass..*

Jones, S., et.al (2001) From The Ground Up: The Publication of Archaeological Projects, *Council for British Archaeology, May 2001*
http://www.britarch.ac.uk/pubs/puns/

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge, UK: Cambridge University Press.

Kircz , J. G. (1998). Modularity: the next form of scientific information presentation? *Journal of Documentation, 54(2):210–235, 1998.*

Mann, W.C. and S. A. Thompson (1988). Rhetorical Structure Theory: towards a functional theory of text organization. *Text, 8(3):243--281, 1988.*

Mons, B (2005), Which gene did you mean? *BMC Bioinformatics 2005 Jun 7, 6:142*

Nentwich, M. (2003), Cyberscience, Research in the Age Of the Internet. *Austrian Academy of Sciences, Vienna 2003.*

Perfetti, C.A., Rouet, J.F, & Britt, A. (1999)., Towards a Theory of Documents Representation, in H. v*an Oostendorp & S. Goldman (Eds.), The Construction of MentalRepresentations during Reading. Mahwah, NJ: L. Erlbaum Ass.*

Rebholz-Schuhmann, D., H.Kirsch, F. Couto (2005). Facts from Text—Is Text Mining Ready to Deliver? PLoS Biology, *www.plosbiology.org* February 2005, Volume 3, Issue 2, e65
http://xldb.fc.ul.pt/data/Publications_attach/dietrich2005.pdf

RST Website: Introduction to Rhetorical Structure Theory, http://www.sfu.ca/rst/05bibliographies/report.html

Tuffield, M. M., Shadbolt, N. R. and Millard, D. E. (2005) Narrative as a Form of Knowledge Transfer: Narrative Theory and Semantics. *In Proceedings of the First AKT DTA Colloquium.*
http://eprints.ecs.soton.ac.uk/11010/

Uren, V., Buckingham Shum, S., Li, G. and Bachler, M. (2004) Sensemaking Tools for Understanding Research Literatures (submitted*). PrePrint: Scholarly Ontologies Project, KMI, The Open University, UK*
www.kmi.open.ac.uk/projects/scholonto

Van Oostendorp, H. and S. R. Goldman (1999), The Construction of Mental Representations During Reading. Mahwah, NJ: *Lawrence Erlbaum Ass..*

Van Zwol, R. and Callista, A. (2005). Content Authoring in an XML-based and Author Friendly Environment, *Technical Report UU-CS-2005-019.*
http://www.cs.uu.nl/research/techreps/UU-CS-2005-019.html

W3C (World-Wide Web Consortium), (1999). Resource Description Framework (RDF) Model and Syntax Specification
http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/

W3C (World-Wide Web Consortium), (2004), OWL Web Ontology Language, Guide, W3C Recommendation 10 February 2004
http://www.w3.org/TR/2004/REC-owl-guide-20040210/