Modeling Scientific Research Articles – Shifting Perspectives and Persistent Issues

Anita de Waard[1,2]; Joost Kircz[3, 4]

[1] Elsevier Labs, Radarweg 29, 1043 NX,
Amsterdam, The Netherlands
e-mail: a.dewaard@elsevier.com
[2] Department of Information and Computing Sciences,
Universiteit Utrecht, The Netherlands
[3]Institute for Media and Information Management (MIM),
Hogeschool van Amsterdam, The Netherlands
e-mail: j.g.kircz@hva.nl
[4]Kircz Research Amsterdam, http://www.kra.nl

Abstract

We review over 10 years of research at Elsevier and various Dutch academic institutions on establishing a new format for the scientific research article. Our work rests on two main theoretical principles: the concept of modular documents, consisting of content elements that can exist and be published independently and are linked by meaningful relations, and the use of semantic data standards allowing access to heterogeneous data. We discuss the application of these concepts in five different projects: a modular format for physics articles, an XML encyclopedia in pharmacology, a semantic data integration project, a modular format for computer science proceedings papers, and our current work on research articles in cell biology.

Keywords: Scientific publishing models; new scholarly constructs and discourse methods; metadata creation and usage; pragmatic and semantic web technologies.

1. Introduction

The objective of our work is, one the one hand, to analyze and investigate what role the research article plays in the connected world that scientists live in today, and on the other hand to propose and experiment with new forms of publication, which contain the knowledge traditionally transferred by 'papers', but are better suited to an online environment. Our research is driven both by an analytical approach stemming from the humanities, including argumentation theory, discourse modeling, and sociology of science, and a knowledge engineering approach from the computer science end, using semantic web technologies, argumentation visualization, and authoring and annotation tools.

We present five examples of our work, in roughly chronological order [1]. We have been driven by two main theoretical concepts: firstly, the concept of modularity: the idea that a scientific text can consist of a set of self-contained and reusable content elements that are strung together to form one or more variants of an evolving series of documents. To explore this concept, Kircz and Harmsze have developed a modular format of the research article in physics [2] this work was extended to create a modular format for Major Reference Work in Pharmacology, which can be used as a database or a linear text [3].

The other main theoretical driver for our research is the use of semantic technologies to access scientific content. In the DOPE project [4], we developed an RDF (Resource Description Framework, [5])-based architecture to access a diverse content set through a thesaurus. This project included the RDF formatting

of Elsevier's EMTREE thesaurus [6] and development of an explorative user interface [7] to access ta heterogeneous dataset.

Lastly, we discuss two projects where we combine the concept of modularity with that of semantic tools and standards. We first identified a simple modular structure for articles in computer science that can be created using LaTeX and converted to semantic formats, entitled the ABCDE Format [8]. Our current work delves more deeply into the text of research articles. We are currently investigating a discourse modeling approach to develop a theoretical framework for a pragmatic model of research articles, linked through a network of argumentational relations. We probe of the pragmatic roles which various discourse elements provide, and modeling the way in which textual coherence and argumentative roles of textural elements are expressed, through an analysis of the linguistic forms used in various parts of a biology text.

2. Modular Documents in Physics

Kircz and Roosendaal [9] summarized the communications needs in the scientific community as follows:

- Awareness of knowledge about the body of knowledge of one's own or related research domains;
- Awareness of new research outcomes, needed for one's current research program;
- Specific information on relevant theories, detailed information on design, methodologies etc.;
- Scientific standards on research approaches and reporting, that develop in the process of a certain research program and shapes the social structure of a field;
- Platform for communication as a tool that enables formal and informal exchange of idea's opinions, results and (dis)agreements between peers;
- Ownership protection on the intellectual results and possible commercial applications.

All of these roles demand different ways of identifying pertinent information units that together compose the paper as we know it today. In early period of transition of the scientific paper to electronic media, proposals for new formats remained traditional, without taking into account the extent to which electronic media change the whole spectrum of dissemination and reading. In a critique of this, we explored the functions of the article and discussed changes in form due to the fact that in an electronic medium, text and non-textual material obtain a different relationship than in the paper world [10, 11]. In our proposal the essay format, typical for a paper product that is meant to be read as an individual information object, is replaced by a mode of communication that is an intrinsic fit for reading electronically. Specifically, this will allow the reader to only read those parts that really serve an information need at a particular place and time. In other words with proper tools we will see a change from *read and locate situation* where first a document is identified and then it is read to identify the needed information, to a *locate and read situation* where we start with a relevant passage of text and from that as starting point decide to read other parts or to skip on.

Such organized browsing, by immediately skipping to determined parts of the text, demands changes in the way research reports are structured and represented on paper and electronic media. This aspect suggests an intrinsically modular structure for electronic publications, first explored in [2]. The PhD research project on modularity of scientific information by Harmsze [12] focused on the dissection of the research paper into different types of information that are conveyed by the structure of the research paper.

This approach leads to a modular model of scientific information in physics, which contains two elements:

- **Modules**: information elements such as positioning (introduction), methods, results, interpretation, outcome and their subdivisions, and
- **Relations** between these elements, both to non-textual elements in the paper as to external relations to (parts of) other works

In Figure 1, we show the modular system developed by Harmsze [12] to model a set of papers in physics, where each module contained a unique type of content, focusing on e.g. the experimental setup or the central problem of a piece of research. Core to the use of modular elements is the concept of reusability: when a paper is updated, it might not need a new Positioning module, but merely provide e.g. new Methods and Results sections.

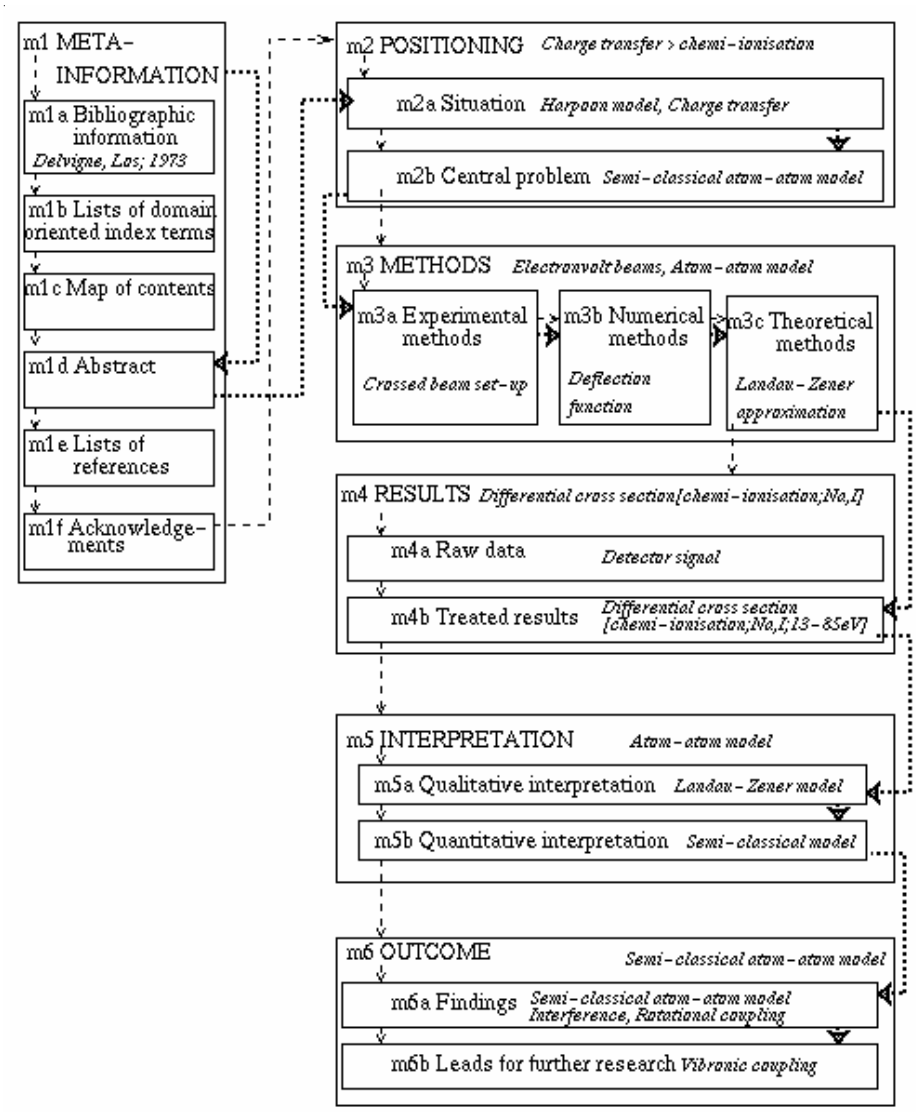

Figure 1: The module meta-information and the modules follow the conceptual function of the information, and the sequential paths leading through the article. The dashed line indicates the complete sequential path, and the dotted line the essay-type sequential path [12].

The other pillar of the modular model is the concept of 'meaningful links' (as the jargon was, at the time), Although the ubiquity of the <A HREF="link.htm"> typical html link </A> has meant a great triumph of the simple hyperlink – one-to-one, mono-directional, not containing information about the link type – a body of research in hypertext has been going on for decades that identifies many different types of links and roles that can play in connecting pieces of data. In thinking about relations in this way, it becomes clear that a relation is not simply a pointer to another piece of data. The fact that the relation exists, and the relationship it expresses between the linking and linked data, provides information in itself, which can be made explicit (e.g. visible and/or searchable) for the reader. We therefore explicitly considered the relation and the information presented within it, or the relation *type*, as separate entities.

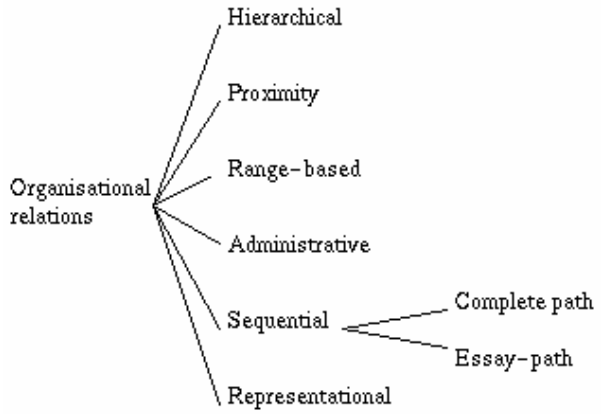

Figure 2: Different types of organisational relations distinguished in the modular model by Harmsze [12]

For physics articles, Harmsze identifies the following detailed taxonomy of relations between modules:

- *Organisational relations* that are based on the structure of the information. These dovetail with the structural XML information we discussed above – see figure 2 for a detailed subdivision, and
- *Discourse relations,* which define the reasoning of the argument. In the model of Harmsze an elaborate skeleton has been worked out. Based on the systematic pragma-dialectical categorization of Garssen [13], these can further be subdivided as:
  - **Causal relations:** relations where there is a causal connection between premise and conclusion (or between *explanans* and *explanandum*). This kind of relation exists between a statement or a formula and an elaborate mathematical derivation. Obviously, the usage of the causal relation as an argument and as an explanation, lie close together.
  - **Comparison relations:** relations where the relation is one of resemblance, contradiction or similarity. The analogue is a typical subtype. Comparisons used as argument are well-known phenomena, such as with the comparison of measured data from, e.g., the module Treated Results with theoretical predictions that fit within certain acceptable boundaries. We can also think of similarity relations, where results of others on similar systems are compared to emphasize agreement or disagreement. In the case of an elucidation, we can think of the relation between the description of a phenomenon and a known mechanical toy model. A link between a text and an image that illustrates the reasoning or results belong to this category. Another example is the suggestion that a drug that is effective in curing a particular ailment might also help against similar symptoms.

- **Symptomatic relations**, which are of a more complicated nature. Here we deal with relations where a concomitance exists between the two poles. This category is more heterogeneous than the other two. This kind of relation can be based on a definition or a value judgment such as the role of a specific feature that serves as a sufficiently discriminatory value to warrant a conclusion. We can think of a relation between the textually described results and a picture in which a specific feature, like a discontinuity in a graph, is used to declare a particular physical effect present or not.

3. A Modular Major Reference Work

The main drawback of the model developed for physics articles was that it is very demanding to the author to adhere to the proposed structure and the model presupposes strong editorial assistance in the form of advanced XML-based text processing software. This could be enforced within the context of a reference work, where a) the content elements are commissioned, and therefore a writing template can be proscribed and b) the main rhetorical purpose of the goal is to inform the reader of existing knowledge, rather than convince him or her of the validity of a specific claim (for more on this, see below). Therefore, we adapted Harmsze's model to use it for XPharm, a state of the art, online, comprehensive pharmacology reference work. XPharm contains information on agents (drugs), targets, disorders and principles of pharmacology [3]. The 4,000 XPharm entries are authored by a group of 600 contributors who write in a very modular format. The idea for four databases was driven by the fact that Agents, including drugs, which are the core of pharmacology, act at molecular Targets to treat Disorders. The Principles database is included as a repository of information fundamental to the discipline but generally independent of the chemical entity,

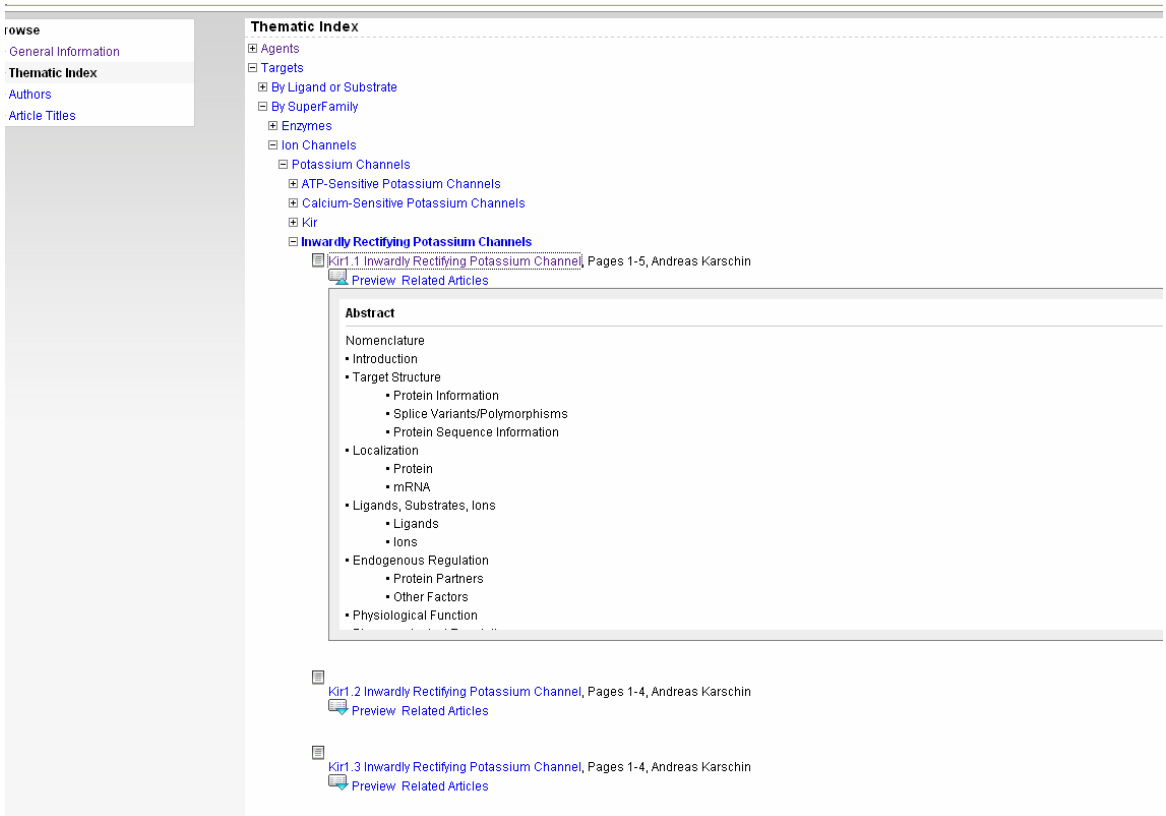

Figure 3: Outline of an XPharm Target Record, showing the modular structure; all topic headings are the same for each Target.

site of action, or clinical use. Each XPharm record can be rendered in a customizable way, and the interface allows for the rendering of modular content elements within different user-defined contexts [3]. XPharm uses the concept of modularity by proscribing a rigid format for each type of entry: for example, all target entries follow the format shown in Figure 3:

The XML of each record is highly granular, for example physical constants are individually marked up so they can, in principle, be extracted and compared to create tables of data, thus enabling the XML to function either as a text (in the html instantiation) or as a database. Relations exist between records to these modular headings, so that the text can be interlinked in a very granular way. Also, this system enables detailed updates of only specific parts of the texts, e.g. if a new antibody is found for a specific target, only that module can be updated. As a conclusion, the system of modular authoring can work quite well for texts in which structures can be mandated and which are more like a 'dressed-up database' than like a persuasive text. As a conclusion, we believe that the difference between informative content sources (such as textbooks and databases) and persuasive texts (such as primary research articles) needs to be taken into greater account when modeling scientific information.

In XPharm, a set of content relations was also proposed, which specifically hold between different elements; these are based on the specific biological rules that govern the interactions between content elements. For example, a disorder can be related to a drug (or Agent, in XPharm terms), by either the Treats relation (Aspirin treats Headache) or the Side Effect relation (Stomach Ache is a Side Effect of Aspirin). A system of 13 such relations was proposed, but because of technical issues (most notably, the lack of ability of current browsers to render relationship types) has not yet been implemented.

4. Semantic Access to Heterogeneous Data: The DOPE Project

The technologies used in the work described above predate (our knowledge of) the semantic web (XPharm was designed in 1998), and the lack of interoperable standards were partly what prevented us from scaling up or connecting to other projects. A next project focused on the use of such standards in the context of pharmacology research. DOPE, the Drug Ontology Project for Elsevier, focused on allowing access via a multifaceted thesaurus, EMTREE, to a large set of data: five million abstracts from the Medline database and about 500,000 full-text articles from Elsevier's ScienceDirect [7]. At the time (2003), no open architecture existed to support using thesauri for querying data sources. To provide this functionality, we needed a technical infrastructure to mediate between the information sources, thesaurus representation, and document metadata stored on the Collexis fingerprint server [14]. We implemented this mediation in our DOPE prototype using the RDF repository Sesame [15].

The records were first indexed to Elsevier's Proprietary thesaurus EMTREE [6]. The version we used, EMTREE 2003, contained about 45,000 preferred terms and 190,000 synonyms organized in a multilevel hierarchy, and currently contained the following information types:

- Facets: broad topic areas that divide the thesaurus into independent hierarchies.
- Each facet consists of a hierarchy of preferred terms used as index keywords to describe a resource's information content. Facet names are not themselves preferred terms, and they cannot be used as index keywords. A term can occur in more than one facet; that is, EMTREE is poly-hierarchical.
- Preferred terms are enriched by a set of synonyms—alternative terms that can be used to refer to the corresponding preferred term. A person can use synonyms to index or query information, but they will be normalized to the preferred term internally.
- Links, a subclass of the preferred terms, serve as subheadings for other index keywords.

They denote a context or aspect for the main term to which they are linked. Two kinds of link terms, drug-links and disease-links, can be used as subheadings for a term denoting a drug or a disease.

The indexing process was done by the Collexis Indexing Engine using a technique called fingerprinting [14], which assigns a list of weighted thesaurus keywords assigned to a document. Next to the document fingerprints, the Collexis server housed bibliographic metadata about the document such as authors and document location. The DOPE architecture (see Figure 4) then dynamically mapped the Collexis metadata to an RDF model. An RDF database, using the SOAP protocol, communicated with both the fingerprint server and the RDF version of EMTREE. A client application interface, based on Aduna's Spectacle Cluster Map [7], let users interact with the document sets indexed by the thesaurus keywords using SeRQL queries, an RDF rule language sent by HTTP [16]. The system design permits the addition of new data sources, which are mapped to their own RDF data source models and communicate with Sesame. It also allows the addition of add new ontologies or thesauri, which can be converted into RDF schema and communicate with the Sesame RDF server [15].

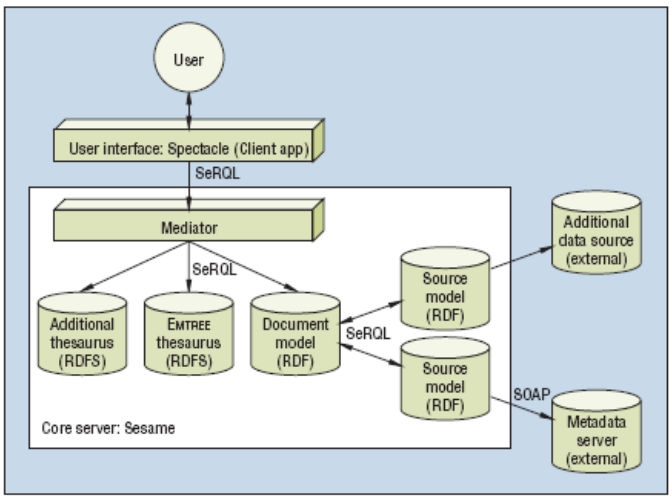

Figure 4: Basic components of the DOPE architecture (technologies are given in brackets)

We performed a small user study with 10 potential end users, including six academic and four industrial users [17]. These users found the tool useful for the exploration of a large information space, for tasks such as filtering information when preparing lectures on a certain topic and doing literature surveys (for example, using a "shopping basket" to collect search results). A more advanced potential application mentioned was to monitor changes in the research community's focus. This, however, would require extending the current system with mechanisms for filtering documents based on publication date, as well as advanced visualization strategies for changes that happen over time, which were not part of the project scope.

Overall, the DOPE system was a useful, working implementation of Semantic Web technologies that allowed for the inclusion of new distributed data sources and ontologies using the RDF data standard. In juxtaposing this project with the experiments in modularity discussed above, we note that a complex representation of the EMTREE thesaurus in RDF was constructed, using historically meaningful relationships between thesaurus elements. The use of semantic standards enables easy scaling of the system with new thesauri, or new relationships. However, of course, within DOPE the documents accessed were not modular, and they could only be related using overlapping or related thesaurus entries. Combining these two concepts, modular documents with meaningful relations and semantic technologies, led to our next series of investigations.

5. Semantic Modular Publishing: The ABCDE Format

Our current research focuses on developing a new format for publications that combines the concepts of modularity with semantic technologies. Our first foray into this area was to develop a simple modular format for structuring conference contributions in computer science, and the authoring, editing and retrieval processes needed to use them. Specifically, this format was meant as a way to allow the use of conference papers by Semantic Browsers such as PiggyBank [18] and semantic collaborative authoring tools such as Semantic Wikis [19]. The ABCDE Format (ABCDEF) for proceedings and workshop contributions is an open-standard, widely (re)useable format, that can be easily mined, integrated and consumed by semantic browsers and wiki's [8]. The format can be created in several interoperable data types, including LaTeX and XML, or a simple text file.

It is characterized by the following elements:

- **A - Annotation.** Each record contains a set of metadata that follows the Dublin Core standard. Minimal required fields are Title, Creator, Identifier and Dat*e.*
- **B, C, D - Background, Contribution, Discussion.** The main body of text consists of three sections:
  Background, describing the positioning of the research, ongoing issues and the central research question;
  Contribution, describing the work the authors have done: any concrete things created, programmed, or investigated;
  Discussion, contains a discussion of the work done, comparison with other work, and implications and next steps.
  These section headings need to exist somewhere in the metadata of the article - but they can be hidden markup; also, each of the sections can have different, and differently named, subheadings.
- **E- Entities.** Throughout the text, entities such as references, personal names, project websites, etc. are identified by:
  - The text linking to an entity
  - The type of link (reference, footnote, website, etc.)
  - The linking URI, if present
  - The text for the link
  In other words, the entity link can be described as an RDF statement [5].
- There is no abstract in an ABCDE document - instead, within the B, C and D paragraphs the author denotes 'core' sentences. Upon retrieval or rendering of the article, these can be extracted to form a structured abstract of the article - where one can jump directly to the core of the Background, Contribution or Discussion. This allows the author to create and modify statements summarizing the article only once, which prevents a misrepresentation in the abstract of the paper, which, in fact, occurs quite often [20].

ABCDEF allows an extensible set of relations to work on documents with a (simple) modular structure, and enables the use of open semantic standards. This format has been described and a LaTeX stylesheet has been published [8]; as a test, a small set of documents for the Semantic Wiki conference was converted to this format. The ABCDE format is a quite simple intermediary step towards creating a reusable, modular, semantic format for research articles. The relations between the modules are quite simple: the sequentiality is obvious (first B, then C, then D for the sections); 'elaboration' relations exist between core sentences in the abstract and their locations in the text; and the entities are related to their links by a link type which the user is free to name. Although this format allows access to the content by various semantic

tools, it still does not do a very good job of marking up the knowledge or argumentation in the text. An attempt at this is made in the currently ongoing project, discussed in the next section.

6. Semantic Modular Publishing: Rhetoric in Biology.

At present, we are developing a more integrated approach, where we look more closely at the way in which rhetoric and persuasion are expressed. The main goal of our research is not a linguistic analysis of a research paper in a field, but the creation of a model that will enable faster browsing through a single paper as well as a collection of related papers. The most important observation from the work done on the modular physics articles was if when you break up the essay-type of article in well-defined units that can be retrieved and re-used, the units of information never become fully independent. A research paper is an attempt to convey meaning, and convince peers of the validity of a specific claim, using research data: therefore, to optimally represent a scientific paper, we should model how it aims to convince. To use a chemical metaphor: breaking up a molecule into its constituent atoms immediately confronts you with the various aspects of chemical binding. In the same way, parts of a scientific text are glued together with arguments, which cannot be disconnected without a loss of meaning to the overall structure.

As a knowledge transmission tool, the research article offers an amalgamate of pragmatic, rhetorical and simply informative functions. Our modularity experiments led us to understand that although certain parts of the paper can be made into database-like elements, other parts are quite complex to modularize, and their format plays a critical role in transferring knowledge and convincing peers of the correctness of a statement. Our current efforts focus on obtaining a better understanding of the sociology and linguistic expressions of scientific truth creation in science. We are using a corpus of full-text articles in the field of cell biology, partly because it is a vast field, where presentations are already quite standardized, and partly because the role of research results vs. theoretical descriptions is very clear-cut. In modeling these articles, we are staying close to the traditional 'IMRaD' (Introduction, Methods, Results and Discussion) format, since first of all the field has consistently adopted this format [21]; an additional motivation for this format can be found by looking at models from classical rhetoric and story grammar models [22].

Therefore, to optimize granularity but still enable the rhetorical narrative flow, our current model in biology has three elements [23]:

I: Content Modules:

- Front matter/metadata
  - Introduction, containing the following subsections:
  - Positioning
  - Central Problem
  - Hypothesis
  - Summary of Results
- Experiments, containing the following discourse unit types:
  - Fact, Goal, Problem, Method, Result, Implication, Experimental Hypothesis
- Discussion, containing the following subsections:
  - Evaluation
  - Comparison
  - Implications
  - Next Steps

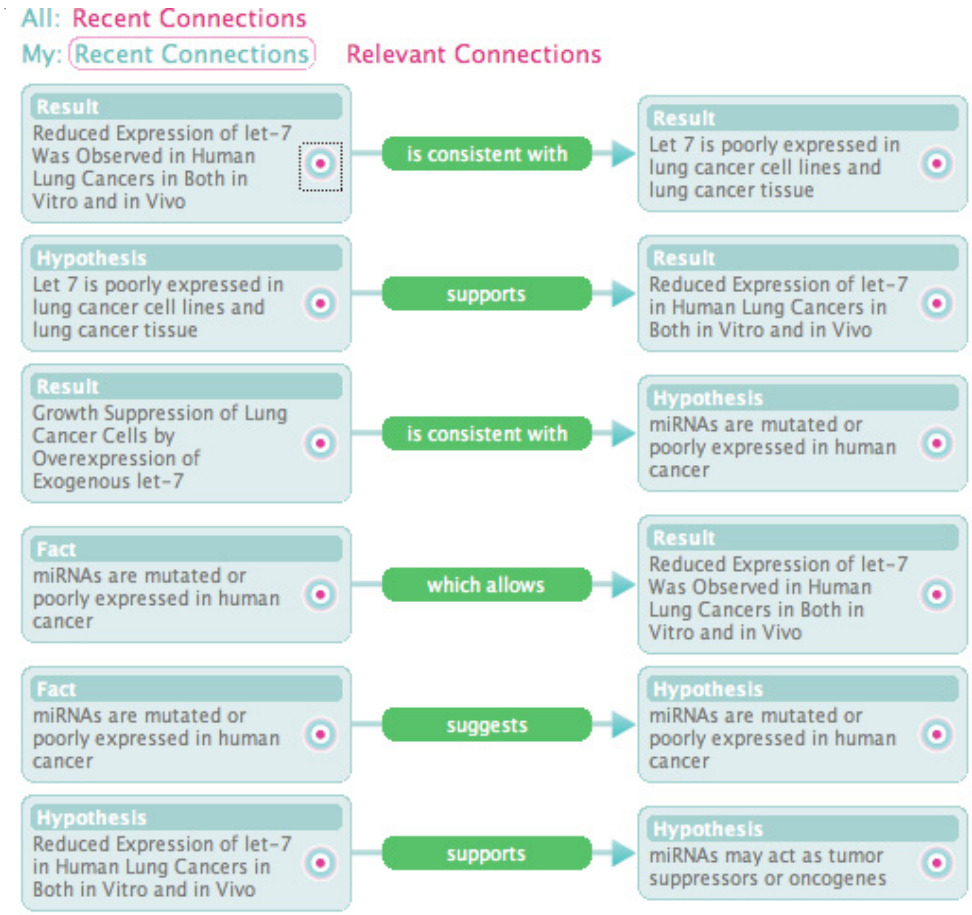

Figure 4: A Subset of Statements and relations from two biology texts, modeled in Cohere [25]; each 'target' is linked into the appropriate location in the underlying documents

II: Entities (contained within the (sub)sections described above), consisting of:

- Domain-specific entities, such as genes, proteins, anatomical locations, etc.
- Figures and tables
- Bibliographic references

III: Relations, consisting of:

- Relations within the document from entity representations to entities (figure, table and bibliographic references)
- Relations out of the document, from entities to external representation (e.g., from a protein name to it's unique identifier in a protein databank)
- Relations between moves within documents, e.g. elaboration, from a summary statement the Introduction section to a Result element within the Experiment section)
- Relations between moves between documents (e.g., agreement between a Result in one paper and that in another paper)

The modular division for the Introduction and the Discussion is based on Harmsze's model and our own empirical investigations (it was easy to fit a collection of 13 biology articles within this framework, and we

hope it will cover the needs of the corpus in general). The Experiments are subdivided in a different way, where smaller elements consisting of one or more phrases are identified using verb tense and cue phrases, as motivated in [24] (a preliminary computational assessment will be given in [23]). Currently, we have marked up a corpus of 13 documents in this format, and we are working on implementing these, linked by the relationships described, in the online argumentation environment Cohere [25], see figure 4 for a screenshot of some of our statements in this environment.

One of the main challenges is to represent the argumentation and the research data in a way which will allow a user to quickly oversee which claims are based on which experimental data, both within and between research articles. Our final goal is to develop a network that clearly differentiates claims from their validation, based on data, and enables insight into the quantitative motivation of a specific statement from its constituent experimental underpinnings. A further direction is to attempt automatic identification of the elements, specifically the moves within the experiment sections, which could enable a (semi) automatic representation of a paper as a set of claims and underlying data.

7. Conclusion

Each of these projects has provided us with insights that, in part, have led to the next experiment. In particular, we have explored various incarnations of modular content representations, linked by meaningful relations. In certain cases, this can be fruitful: for example, a modular structure for an encyclopedic work can allow certain user functions that a narrative, linear structure does not allow; the ABCDE format enables an accessible representation of a collection of research papers inside a semantic architecture. The next major issue is to see whether a partly modular, partly linear format, where content elements are at least identified by type (Method, Hypothesis etc.) can indeed replace the existing linear narrative. If it does turns out to enable more useable reading environments, we need to ensure that the creation of the format can be achieved, given current publishing practices. We hope that our current experiments can help provide a format that offers computational handholds to access the argumentative elements within a research paper. Lastly, we want to state our are interest in exploring collaborations on this subject with the myriad initiatives that are currently ongoing, since we firmly believe this complex problem can only be solved by collaborative effort. This issue does not have a purely technological solution; to truly improve the way in which science is communicated will require serious scrutiny by the scientific community of the social, political and psycholinguistic way in which it claims, confirms, and creates knowledge.

8. References

[1] This paper is aimed to describe previous and current projects, and does not contain a theoretical embedding or references to related work; these have been addressed in [9, 10, 22, 24] and will be addressed in a forthcoming [23].

[2] Kircz, J.G. and F.A.P. Harmsze, "Modular scenarios in the electronic age," *Conferentie Informatiewetenschap 2000. Doelen, Rotterdam 5 April 2000. In: P. van der Vet en P. de Bra (eds.) CS-Report 00-20. Proceedings Conferentie Informatiewetenschap 2000. De Doelen Utrecht, 5 April 2000*. pp. 31-43.

[3] Enna, S..J., D. B. Bylund, Preface, *XPharm*, doi:10.1016/B978-008055232-3.09004-9

[4] Stuckenschmidt, H., F. van Harmelen, A. de Waard, et.al, , "Exploring Large Document Repositories with RDF Technology: The DOPE Project," *IEEE Intelligent Systems*, vol. 19, no. 3, pp. 34-40, May/Jun, 2004

[5] Brickley D. (ed.), *RDF Vocabulary Description Language 1.0: RDF Schema,* W3C Recommendation 10 February 2004, http://www.w3.org/TR/rdf-schema

[6] For more information, see http://www.info.embase.com/emtree/about/

[7] Fluit C., M. Sabou, and F. van Harmelen, "Ontology-Based Information Visualization," *Visualizing the Semantic Web*, V. Geroimenko and C. Chen, eds., Springer-Verlag, 2003, pp. 36-48.

[8] Waard, A. de and Tel, G., "The ABCDE Format: Enabling Semantic Conference Proceedings," In: *Proceedings of the First Workshop on Semantic Wikis, European Semantic Web Conference (ESWC 2006)*, Budva, Montenegro, 2006.

[9] Kircz, J. G. and Hans E. Roosendaal, "Understanding and shaping scientific information transfer,". *In: Dennis Shaw and Howard Moore (eds). Electronic publishing in science. Proceedings of the ICSU Press / UNESCO expert conference February 1996. Unesco Paris 1996*. pp. 106-116.

[10] Kircz, J.G., "New practices for electronic publishing 1: Will the scientific paper keep its form," *Learned Publishing*. Volume 14. Number 4, October 2001. pp. 265-272.

[11] Kircz,. J.G., "New practices for electronic publishing 2: New forms of the scientific paper*," Learned Publishing*. Volume 15. Number 1, January 2002. pp. 27-32

[12] Harmsze, F, *"*A modular structure for scientific articles in an electronic environment*,"* *PhD thesis, University of Amsterdam*, February 9, 2000.

[13] Garssen, B., "The nature of symptomatic argumentation," In: Frans H. van Eemeren, Rob Grootendorst, J Anthony Blair, Charles A. Wilards (eds*.). Proceedings of the 4th International Conference of the International Society for the Study of Argumentation*, Amsterdam, June 16-19 1998. Amsterdam: SICSAT, 1999.

[14] Van Mulligen, E.M. et al., "Research for Research: Tools for Knowledge Discovery and Visualization," *Proc. 2002 AMIA Ann Symp*., Am. Medical Informatics Assn., 2002, pp. 835–839.

[15] Broekstra, J., A. Kampman, and F. van Harmelen, "Sesame: An Architecture for Storing and Querying RDF and RDF Schema," *Proc. 1st Int'l Semantic Web Conf.*, LNCS 2342, Springer-Verlag, 2002, pp.54–68.

[16] Broekstra J. and A. Kampman, "SeRQL: Querying and Transformation with a Second- Generation Language*," technical white paper, Aduna/Vrije Universiteit Amsterdam,* Jan. 2004.

[17] Stuckenschmidt, H., A. de Waard, R. Bhogal et.al, "A Topic-Based Browser for Large Online Resources," *In: Proceedings of the Proceedings of the 14th International Conference on Knowledge Engineering and Knowledge Management ({EKAW}'04)*. Editors E. Motta and N. Shadbolt. Series Lecture Notes in Artificial Intelligence.

[18] Huynh, D., Stefano Mazzocchi, and David Karger. "Piggy Bank: Experience the Semantic Web Inside Your Web Browser", *Proceedings International Semantic Web Conference (ISWC)* 2005.

[19] For definitions and examples, see http://en.wikipedia.org/wiki/Semantic_wiki

[20] Pitkin, R.M., Branagan, M.A., Burmeister, L.F., "Accuracy of data in abstracts of published research articles," *JAMA* 281 (1999) 1110–1111

[21] See e.g., the International Committee of Medical Journal Editors, "Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publications," ***Updated October 2007***, available online at http://www.icmje.org/

[22] Waard, A. de, Breure, L., Kircz, J.G. & Oostendorp, H. van (2006), "Modeling Rhetoric in Scientific Publications," in: *Current Research in Information Sciences and Technologies* (pp. 352-356). Vicente P. Guerrero-Bote (Editor) (Ed.), Badajoz, Spain: Open Institute of Knowledge.

[23] Waard, A. de, "A Semantic Modular Structure for Biology Articles," forthcoming

[24] Waard, A. de, "A Pragmatic Structure for the Research Article," in: *Proceedings ICPW'07: 2nd International Conference on the Pragmatic Web, 22-23 Oct. 2007, Tilburg: NL*. (Eds.) Buckingham Shum, S., Lind, M. and Weigand, H. Published in: ACM Digital Library & Open University ePrint 9275.

[25] Buckingham Shum, S., "Cohere: Towards Web 2.0 Argumentation," *In: Proceedings, COMMA'08: 2nd International Conference on Computational Models of Argument, 28-30 May 2008, Toulouse*. IOS Press: Amsterdam.