User-centric design and evaluation of a semantic annotation model for scientific documents

de Ribaupierre, Hélène; Falquet, Gilles

User-centric design and evaluation of a semantic annotation model for scientific documents

Hélène de Ribaupierre
CUI, Université de Genvève
Battelle, 7 rte de Drize
Carouge, Geneva, Switerland
helene.deribaupierre@unige.ch

Gilles Falquet
CUI, Université de Genvève
Battelle, 7 rte de Drize
Carouge, Geneva, Switerland
gilles.falquet@unige.ch

ABSTRACT

When performing document search, scientists have specific goals in mind. We conducted interviews with scientists to understand exactly how they were looking for information and working with documents. We found that scientists are generally searching specific discourse elements, not the entire document. Therefore, we created an annotation model that can represent the different types of discourse elements contained in documents. We have implemented this model in the form of an OWL ontology and a semantic indexing and retrieval tool. The experiments we have conducted (in the gender studies field) show that the model is sufficient to represent a large part of the document contents and that it is possible to automatically annote documents according to this model. We also showed that this model can be used to answer specific and complex queries on a corpus of scientific documents.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Indexing methods; H.3.3 [**Information Search and Retrieval**]: Retrieval models; H.3.7 [**Digital Libraries**]: User issues

General Terms

Human Factors, Experimentation

Keywords

Ontologies; Academic publishing; users model, semantic publishing

1. INTRODUCTION

Hannay [7] wrote that scientists have better tools to manage their personal data (photos and video) than to manage or search their professional data. This observation is still valid, it is always difficult for a scientist to find the right documents that actually correspond to an information need. Some of these difficulties are of a general nature: the size of the document corpora, synonymy and homonymy, natural language variability, etc.. However, some problems are specific to scientific texts and to the information needs of scientists. Traditional search engines can find documents by their metadata (title, author, year of publication, etc..) or by the words they contains (full-text search). The first search mode is effective only when the user knows at least one metadata element, such as the title or author. Full-text indexing and search are effective to find documents *about* some topic but they do not take into account the discursive or rhetorical context in which the terms are set. Therefore, it is not possible to know whether a term appears, for example, in a definition, or in the description of a methodology, or in the statement of a problem, etc.. Semantic indexing, while solving some homonymy or synonymy problems and detecting some semantic relations among terms, is not sufficient for scientific search. In order to answer specific and complex queries such as "find all the research results that show that girls are better at reading tasks than boys and that uses a quantitative methodology" a system must be able to detect if the required terms effectively appear in sentences or paragraphs that describe research result or methodologies.

In addition, conventional information retrieval systems are based on the idea that each document has a degree of relevance to the user query, and only the most relevant documents must be selected. For scientists, the goal is often to find *all* the documents that deal with a very specific issue of interest. Moreover, scientists do not have time to read the retrieved documents in their entirety. Therefore a search system must provide methods and tools to strategically read these documents, i.e. to highlight or select the passages that actually contribute to answering the user query.

2. USER STUDIES AND DISCOURSE ELEMENTS

In this paper, we propose a model and a system for scientific document annotation, that take into account the needs of scientists. To understand these needs, we have conducted two studies, one quantitative and the other qualitative, with scientists from different communities [3]. In the qualitative study we interviewed scientists. We asked them what were the questions they had in mind just before turning them into keywords submitted to search engines. The collected questions (see Table 1) helped us, along with the others questions asked, define the annotation model, and they constitute a set of use cases to evaluate our system.

i-KNOW '14 Sep 16-19 2014, Graz, Austria
ACM 978-1-4503-2769-5/14/09.
http://dx.doi.org/10.1145/2637748.2638446

Table 1: Sample user questions

User questions	Induced generic use case
Find all the definitions of the concept "semantic homogeneity"and if it possible to calculated it.	Find the different definitions of a term, and the different facets of the term
Do Christine Delphy argues against Patricia Roux in a paper?	Find author X that agree or disagree with author Y
Find all authors working on intra-individual variability in terms of behavior	Find authors in my field of research

Our model is also based on results provided by studies in the fields of behavioral research, information retrieval, and reading science. Among other things, Bishop [1] showed that when writing a scientific article, scientists aggregate and re-aggregate information in an iterative knowledge construction process. She also showed that indexing specific components in a digital library (figures, conclusions. references, title, title of figures / tables, authors, etc..) help scientists formulate more relevant search queries. Tenopir et al. [13] showed that scientists read for different reasons, such as teaching, article writing, project proposals, etc.

Renear [11], showed that scientists read and extract specific information such as findings, equations, experimental protocols, or data. They also found that scientists read to monitor the progress of their research peers and competitors and to extract facts and evidences to build their knowledge.

By aggregating these studies and our own results, we found, among other things, that scientists focus on specific elements of the documents they read, depending on the task they have to perform. The five main types of document elements that scientists are looking at (not counting the abstract) are those describing *findings*, *methodologies*, *hypothesis*, *definitions*[1] , and *background* (knowledge obtained from referenced work). The interviews also confirmed that elements of a given element type may appear almost anywhere in a document, not necessarily in the section or subsection whose title matches the element type. For instance, a methodology element may appear in the introduction or background. Thus, element types do not correspond to structural parts of the documents.

3. AN ANNOTATION MODEL FOR SCIENTIFIC ARTILES

There are a number of annotation models for scientific documents. Some authors [6, 8, 14, 9, 4] suggest using the rhetorical structures or elements of discourse document to annotate, either manually or automatically, the documents to produce better systems for information retrieval or create summarizers automatic document. These studies generally use documents from the so-called "hard" sciences such as biology, medicine or physics, where documents are highly structured, and therefore the way to describe the results, assumptions or methods may be more formalized. Furthermore, only [5], taking into account the definition as the element type of discourse. In addition, to our knowledge, only [14] and [5] automatically annotate documents, other models are used for manual annotation .

The construction of our annotation model (see figure 1) is also based on the results of these studies, it includes and aggregates some concepts of these models. Our model is based on four axis: a taxonomy of discourse elements; the semantic indexing of the element contents (texts); explicit relationships between elements; and standard metadata.

Discourse elements. The discourse elements (findings, definition, hypothesis, methodology and related works) are the central part of the first axis of our annotation model, the structure of these elements is formally defined in the SciDocAnnotation OWL ontology[2], according to the following principles.

A *definition* is decomposed into a *definiens* (the sentence(s) that provides the meaning of the definition) and the *definiendum* (the defined term). This decomposition allows for specific queries such as: *Find all the definitions of the term 'gender' where the definiens contains 'social construction'*. In addition, the definition is connected to the domain concept that has a label equal to the definiendum. This concept is a member of an auxiliary high-level domain ontology that represent a consensual view of the scientific domain of the document corpus.

Findings include all research results, observations, discussions, and conclusions of a document. They are subdivided into raw results, which are results not yet analyzed or discussed, and already analyzed and discussed results.

Methodology elements describe, using the concepts of a external ontology of methods, everything about the research methodology: techniques, equipment, variables, etc.

Hypothesis elements typically propose answers to open questions. They do not necessarily exist in all scientific documents. There are a lot of research that do not describe or do not have a research hypothesis, especially in research using and inductive approach.

Related work (or background) elements are definition, findings, methodologies, or hypothesis that come from previous works.

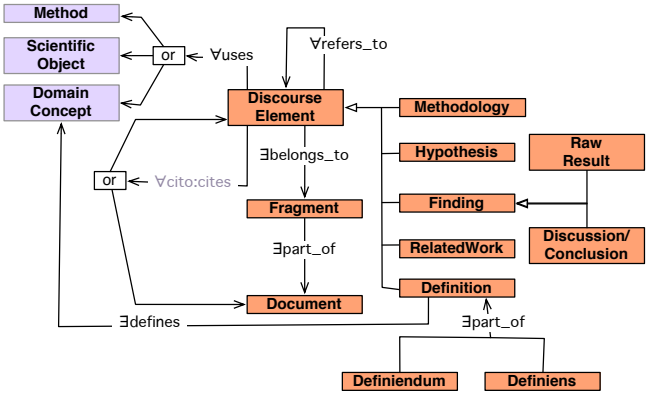
In this model we assume that the annotations of scientific papers must be done in a "universal knowledge" perspective and not centered on the author, i.e. the interesting definitions, findings, etc. from a reader point of view are not

[1]Note that the use of the Google "define"option is far from satisfactory. In fact, Google looks up definitions in well known glossaries and repositories, such as Wikipedia, that represent consensual knowledge. The goal of a scientist, in this case, is instead to find definitions proposed by scientists in the articles of a given corpus

[2]`http://cui.unige.ch/isi/onto/sdl/SciDocAnnotation.owl`

necessarily those created by the author. For example, when an interviewee mentioned the question "Find the different articles that deal with the evaluation of surgical simulators ", this scientist was interested in finding all documents on this subject, irrespective of the author. In her practice she starts by looking for surveys on techniques for assessing surgical simulators, if none exists, then she starts reading various articles on this subject, looking for passages that deal with these techniques. Theses passages may be original works by the author or references to other works.

Figure 1: Sientific document annotation model (the classes Methods, Scientific Object and Domain Concept are imported from other ontologies)



Textual contents. The second axis consists of the representation of the element contents. The content of each discourse element is semantically indexed by means of concepts from three auxiliary ontologies: an ontology of the studied domain; an ontology of scientific objects (equations, models, algorithms, theorem, etc..); and an ontology of methods (types of methods, types of variables, tools, material, etc..).

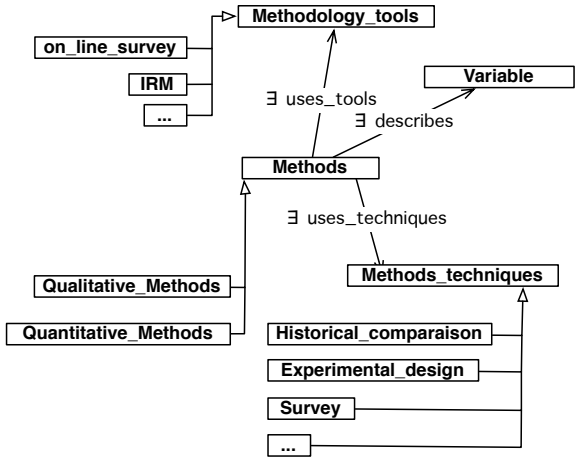
The ontology of methodologies (see Figure 2) and the ontology of scientific objects are generic while the domain ontology is of course domain dependent (e.g. an ontology of gender studies, an ontology of particle physics, etc.).

The ontology of scientific objects describes the various artefacts that are used in the expression of discourse elements. Typical concepts of this ontology are: test, example, equation, model, algorithm, data, table, diagram, drawing, etc.

These ontologies are kept separate from the annotation ontology so they can be easily interchanged and there is a clear distinction between the categorization of discourse elements on the one hand and their content on the other.

Element Relationships. The third axis consists of all explicit references from a document or discourse element to another document or discourse element. We re-used the CiTO ontology [12], and extended it to represent citations between documents but also between discourse elements since a large proportion of the citations do not refer to a whole document but to some, often very restricted part (an equation, a sentence, a definition, etc..). This is necessary to answer

Figure 2: Extract from the methodology research ontology



precise queries such as "Find all the paragraphs containing an outcome ("finding ") about the difference between girls and boys in school and referencing a *result* of Zazzo." It also becomes possible to perform more detailed analysis of the network of citations, depending on the types of citing or cited elements. We kept the numerous types of citations defined in CiTO, but we grouped them in three upper-level citation types: positive (agreesWith, confirms, ...), neutral (discusses, extends, reviews, ...), and negative (corrects, critiques, disputes, ...).

Metadata. The fourth axis consists of current meta-data in the field of scientific documents (bibliographic data), as the names of authors, title of the article, the journal name or publisher, date of publication etc..

4. IMPLEMENTATION AND EVALUATION OF THE MODEL ON A USE CASE IN GENDER STUDIES

To evaluate the relevance of the proposed model we must evaluate 1) to what level is it possible to *automatically* annotate documents according to this model and 2) the benefits for the users in terms of search precision and recall. This is why we developed an annotation and retrieval system based on this model. The heart of the annotation system is the SciDocAnnotation ontology, it provides a reference to categorize and describe the discourse elements of each scientific document. The ontology contains 69 classes, 137 object properties and 13 datatype properties (counting those imported from CiTO). An annotated document is represented by interconnected individuals belonging to the DiscourseElement class or one of its subclasses. Thus the assertion level (ABox) of the ontology stores the semantic index of the document corpus.

The methodology ontology has 36 classes and 8 object properties. We also created a domain ontology for our use case domain, namely, gender studies. It contains 365 classes, 10 object properties and 4 properties datatype.

4.1 Coverage evaluation

At first, we manually annotated 1127 sentences drawn from four articles in the field of gender studies (in this case study we equated sentences and discourse elements, but the model supports larger or smaller size elements). We chose this domain because it gave rise to very heterogeneous written document, ranging from highly empirical studies to "philosophical" texts, and these documents rarely use the IMRaD model (introduction, methods, results and discussion). Among the sentences we found 29 definitions, 497 findings, 56 assumptions, 128 methodologies, and 154 reference to other works (background). Sentences that could be annotated with one of our five discourse element types represent between 16.6 % and 64.23% (49.3% on the average) of the article sentences (see Table 2) (Document Doc2 contains a large number of interview extracts, which have so far not yet been annotated).

The (human) annotators observed that these element types could be found in different places in the text, for example, the following finding of an the article by Correll [2, p.2]: "For example, human capital theorists have argued that women choose jobs with flatter rates of wage growth, because these jobs, which are primarily in female-dominated occupations, have smaller wage penalties for sustained periods of absence from the paid labor force and have higher starting wages (Polachek 1976, 1981; Zellner 1975)." is located on the second page of the article in a section entitled *Human Capital Explanations*. It is therefore impossible to rely on the section title, or on the sentence location in the document to determine its discourse element type. This first test has allowed us to see that the model adequately covers the main elements discourse elements of a scientific paper. We also verified that the different use cases (see Table 1) can be expressed as formal queries (in SPARQL) over the individuals stored in the ontology. We took advantage of this manual test to produce a corpus of 555 annotated sentences that can serve as a reference (golden standard) to evaluate the performance of automated annotation tools. Indeed, this experiment also clearly showed that manual annotation is extremely time consuming and therefore is not a realistic approach, even for a small corpus.

4.2 Automated annotation

In a second step, we have implemented an automated annotator based on the GATE platform[3] with ANNIE[4], JAPE syntactic rules, and the ontology management modules. In this implementation we have considered sentences as discourse elements and paragraphs as document fragments. We have created specific JAPE rules to recognize the different types of discourse elements (20 rules to recognize findings, 34 for definitions, 11 for hypothesis, and 19 for methodologies). To create these rules, we started from the manually annotated sentences and have analyzed the different patterns of grammatical structures produced by the ANNIE parser. We also added typical terms that appear in each type of discourse elements. For example, the term *paper* followed, at a short distance, by the the verb *show* probably indicates a finding.

Below are some examples of sentences together with the induced JAPE rules.

"This result would be consistent with research showing that individuals are more prone to cognitive biases that are self-serving (Markus and Wurf 1987)."[2]

```
((RESULT)
({Token})[0,2]
(CONSISTENT)
{Token.kind==word,Token.category==IN}
(NOUN|ADJECTIVE))
```

"On this usage, gender is typically thought to refer to personality traits and behavior in distinction from the body."[10]

```
(NOUN)
(VERBbe)
{Token.kind==word,Token.category==RB}
{Token.kind==word,Token.category==VBN}
({Token})?
{Token.kind==word,Token.category==TO}
(REFER)
{Token.kind==word,Token.category==TO}
)
```

To test the quality of the automatic annotation process we run it on the 555 manually annotated sentences of our golden standard. We performed measurements of precision / recall on these sentences (see Table 3) which show good precision, but low recall.

4.3 Retrieval performance

To perform comparative tests with users we automatically annotated 903 articles in English, from various journals in gender and sociological studies. The full process is shown on Figure 3. It consists in transforming the original PDF files into text files; applying the GATE pipeline we defined (with ANNIE and JAPE rules) to produce a list of discourse elements (in XML); transforming this file into an RDF graph and loading it into an Allegrograph triple store. We chose Allegrograph because it supports $RDFS++$ reasoning in addition to SPARQL query execution.

To increase recall we added heuristics such as: If a fragment (paragraph) contains unrecognized elements (sentences) and at least three elements with the same type T then assign type T to the unrecognized elements. With these rules, we created 341 findings, 130 methodologies, 29 hypothesis and 6 additional definitions. Nevertheless, we observed that the coverage is significantly lower than with manual annotation. This is certainly due to a very conservative automatic annotation.

Since the Allegrograph system provides a full-text search operator we were able to analyze the difference between traditional full-text search and querying with our model (by specifying the element types). For example let's see what happens with the query "Find definitions that refer to the *gender* concept" on the gender studies corpus. The following SPARQL query return 143 definitions.

[3] `http ://gate.ac.uk/`
[4] `http://gate.ac.uk/ie/annie.html`

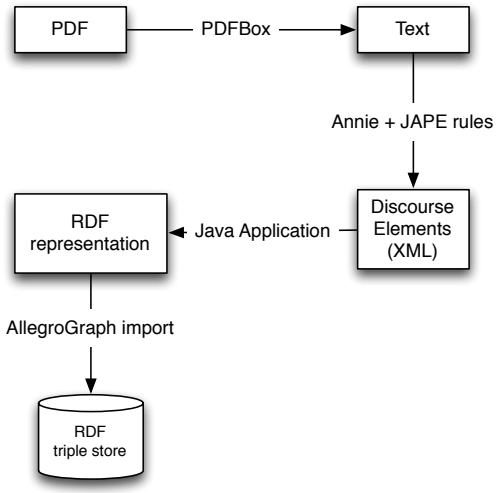
Table 2: Coverage for manually annotated documents

NoDoc	Nb de phrase	Findings	Definition	Methodologie	Hypothese	Total
Doc1	97	58.76%	-	-	1.03%	60.82%
Doc2	288	13.41%	1.77%	-	-	16.61%
Doc3	387	27.91%	2.33%	1.81%	-	55.56%
Doc4	355	41.97%	-	1.13%	1.41%	64.23%

Table 3: Precision/recall values

Discourse element type	Nb of sentences	Precision	Recall	F1.0s
findings	168	0.82	0.39	0.53
hypothesis	104	0.62	0.29	0.39
definitions	111	0.80	0.32	0.46
methodology	172	0.83	0.46	0.59

Figure 3: Automated annotation process




```
select DISTINCT ?p ?com ?term {
?p rdf:type DocuScientific:Definition>.
?p rdfs:comment ?com.
?f annotDocuScientific:has_discourse_element> ?p.
?p annotDocuScientific:describe> ?concept .
?concept genderStudies:term> ?term.
FILTER regex(str(?term), "gender", "i") }
```

A full-text search at the document level (that is how usual search engine operate) returns the contents of all the documents that contain the word *gender* (13'210 sentences). Even though the documents are ranked, the elements are not and the user must read a large part of the returned sentences to find the ones that are definitions.

```
SELECT ?s ?o WHERE
{?p annotDocuScientific:has_discourse_element ?s.
?s rdfs:comment ?o.
?s fti:match "gender". }
```

With a more precise query, using typical keywords that appear in definitions we obtain much fewer elements (68) but only 30 of them are definitions.

```
SELECT ?s ?o WHERE
{?p annotDocuScientific:has_discourse_element ?s.
?s rdfs:comment ?o.
?s fti:matchExpression
'(and (or "definition" "define") "gender")'.
}
```

5. DISCUSSION AND CONCLUSION

The current automated annotation process does not take into account all the model features. In the case of definitions, it globally annotate the definitions, but it is not able to recognize the definiendum and definiens. The model supports references at the document *and* at the discourse element level. However, the annotator does not cover this level of detail. In addition, it is not able to distinguish the different types of CiTO reference relationships, so we use the generic *cites*. Despite these inaccuracies and these simplifications, we were able to build a query system that already outperforms keyword search in many cases. We implemented two interactive search interfaces: a classic keyword based search (with a TF * IDF based weighting scheme) and a faceted interface based on our model (facets correspond to the types of discourse elements). The first pre-tests we conducted with users effectively show that they are able to use the model and obtain much better results than with the keyword search. We are currently conducting usability tests and collecting data to scientifically assess the quality of the system and to determine the influence of the precision/recall of the automated annotation process on the system performance.

The main contribution of this work is the creation a new annotation model constructed from interviews and questionnaires to scientists. This model is focused on the needs of the user, it is built around discourse elements most frequently used by the scientists. We can consider that the model is realistic insofar as automatic annotation of documents is possible with conventional natural language processing tools. The results we obtained on qualitative tests clearly show the contribution of such a model compared to the conventional keyword search.

6. REFERENCES

[1] A. P. Bishop. Document structure and digital libraries: how researchers mobilize information in journal articles. *Information Processing and Management*, 35(3):255 – 279, 1999.

[2] S. J. Correll. Constraints into preferences: Gender, status, and emerging career aspirations. *American Sociological Review*, 69(1):93–113, 2004.

[3] H. de Ribaupierre and G. Falquet. New trends for reading scientific documents. In *Proceedings of the 4th ACM workshop on Online books, complementary social media and crowdsourcing*, BooksOnline '11, pages 19–24, New York, NY, USA, 2011. ACM.

[4] A. de Waard, S. B. Shum, A. Carusi, J. Park, M. Samwald, and Á. Sándor. Hypotheses, evidence and relationships: The hyper approach for representing scientific knowledge claims. In *Proceedings 8th International Semantic Web Conference, Workshop on Semantic Web Applications in Scientific Discourse. Lecture Notes in Computer Science, Springer Verlag: Berlin*, October 2009.

[5] B. Djioua and J. Descles. *Indexing documents by discourse and semantic contents from automatic annotations of texts*. 2007.

[6] T. Groza, K. Muller, S. Handschuh, D. Trif, and S. Decker. Salt: Weaving the claim web. In *Proceedings of the Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007), Busan, South Korea (Berlin, Heidelberg*. 2007.

[7] T. Hannay. What can the web do for science? *Computer*, 43(11):84–87, 2010.

[8] F. Harmsze. *A modular structure for scientific articles in an electronic environment*. PhD thesis, Jan. 2000.

[9] F. Ibekwe-Sanjuan, F. Silvia, S. Eric, and C. Eric. Annotation of Scientific Summaries for Information Retrieval. In O. A. . H. Zaragoza, editor, *ECIR'08 Workshop on: Exploiting Semantic Annotations for Information Retrieval*, pages 70–83, Glasgow, Royaume-Uni, Mar. 2008.

[10] L. Nicholson. Interpreting gender. *Signs*, 20(1):pp. 79–105, 1994.

[11] A. H. Renear and C. L. Palmer. Strategic reading, ontologies, and the future of scientific publishing (vol 325, pg 828, 2009). *Science*, 326(5950):230–230, Oct. 2009.

[12] D. Shotton. Cito, the citation typing ontology, and its use for annotation of reference lists and visualization of citation networks. *The 12th Annual BioOntologies Meeting*, pages 1–4, 2009.

[13] C. Tenopir, D. King, and S. Edwards. *Electronic journals and changes in scholarly article seeking and reading patterns*. 2009.

[14] S. Teufel and M. Moens. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics 28*, 4:409–445, 2002.