# Crowdsourced Semantic Annotation of Scientific Publications

Master Thesis

Jaana Takis
Matr. no.: 2424318
takis@informatik.uni-bonn.de

11th November 2014

Rheinische Friedrich-Wilhelms-Universität Bonn
Institut für Informatik III
Professor Dr. Sören Auer

## Declaration of Independent Work

I herewith declare that all the work described within this Master thesis is the original work of the author. Any published (or unpublished) ideas or techniques from the work of others are fully acknowledged in accordance with the standard referencing practices.

Jaana Takis
11th November 2014
Signature:

# Abstract

This work presents the architecture of a semantic annotation tool for scientific publications and implements a prototype. This work also advances the state of the art in the semantic publishing field by enabling semantic tagging of scholarly papers as a crowdsourced effort of interested parties. Earlier semantic publishing research has not sufficiently addressed the semantic annotation of papers published already in PDF format. The architecture presented in this work supports the semantic annotation of PDF documents with any available annotation vocabulary and with resources from the linked open dataset DBpedia derived from Wikipedia. It also includes the implementation of a functionality to recommend similar papers. A novel approach here is the ability to take into account the structural context of annotations with a special focus on the discourse elements of scholarly papers. A prototype was developed that implements this architecture – a web-based system based on PDF.js that communicates with a SPARQL endpoint. The usability of the user interface and the usefulness of recommendations were evaluated with 10 test users and by interviewing an expert. Evaluation results include an overview how the number of previously annotated papers in the database influenced the number of recommendations of related papers. An interesting point from the evaluation was the discovery of semantic matches between seemingly unrelated papers and how a very good coverage of related papers was achieved with very few annotations. The architecture of the tool was designed to be extensible and one of the extensions discussed is the integration with the Annotopia Open Annotation Server, a universal hub for storing and publishing of annotations for the common benefit of all of us.

# Index of Figures

# Index of Tables

# Index

# 1 Introduction and Motivation

Each year a growing number of research papers are published and researchers find it increasingly hard to keep track of relevant research in their field. Many have been trying to find solutions to alleviate this problem, giving rise to the semantic publishing field, yet scholarly knowledge is still not efficiently shared. Not all the parties that are expected to partake in the semantic publishing effort have the motivation, knowledge or access to tools to do this. Many have difficulties adapting to the new roles that the semantic publishing envisions for them. As a result, there is a certain status quo in the field of semantic publishing, aggravated further by the belief that semantic publishing cannot be implemented without the full participation and timely contribution of all the parties involved. This thesis addresses some of these obstacles. It approaches semantic publishing as something that does not necessarily need to depend on the timely contribution of all parties but rather as something that can also be achieved as a crowdsourced effort of interested parties – whatever their role or involvement in the field.

Other motivations for this work are also influenced by very common problems in scholarly publishing. Namely, the vast majority of research papers are published in the PDF format, a format that has not sufficiently been supported by available semantic annotation tools. A prototype of such a tool that solves this issue is presented in this thesis. The vision for this tool was to not be just another tool for applying classification vocabularies but also allow the users to create connections between annotations and support more complex annotation scenarios that are much more useful in the scholarly context.

Another goal is to design such an architecture that is capable of supporting the recognition of the more relevant semantic connections between papers. This is done with the help of an additional ontology, which allows the user to define the context of annotations through the typical elements of a scientific discourse like abstract, motivation, problem statement etc. This novel approach enables the system to view a scholarly paper as a hierarchical tree of annotations, its hierarchy determined with respect to the physical position and length of each annotation. This information is very useful because as the evaluation section of the thesis shows, surprising semantic connections can be discovered between papers that are only vaguely related or even assumed to be unrelated. Information of this kind helps us to better evaluate the relevancy of each paper and allow us to ask questions that are very common in the field of scholarly publication, but not supported by the traditional text-based search - e.g. “find papers that are motivated by …”. The similar paper recommendation functionality of the prototype demonstrates that approach in practice.

In conclusion, this work implements a universal and flexible prototype solution motivated by some common problems in the field of semantic publishing. The crowdsourced RDF data that can be gathered with this tool also has a strong potential to contribute to further research in semantic publishing, ontology development and linked data effort.

In the rest of the thesis, a review of the state of the art is given to better identify the topical issues, followed by an introduction to the architecture and implementation of the developed prototype. Evaluation of the prototype includes interesting findings on surprising semantic connections between seemingly unrelated papers and how 97% of related papers in the dataset were found on average per user with only 5 annotations per paper. Also an overview is given of how the number of previously annotated papers in the database influenced the number of recommendations of related papers.

# 2 State of the Art in Semantic Publishing

This section reviews current approaches in the semantic publishing of scholarly articles. The review focuses on the semantic enrichment and publishing stages of this process, any useful services that might be built upon the published data remains outside the scope of this overview.

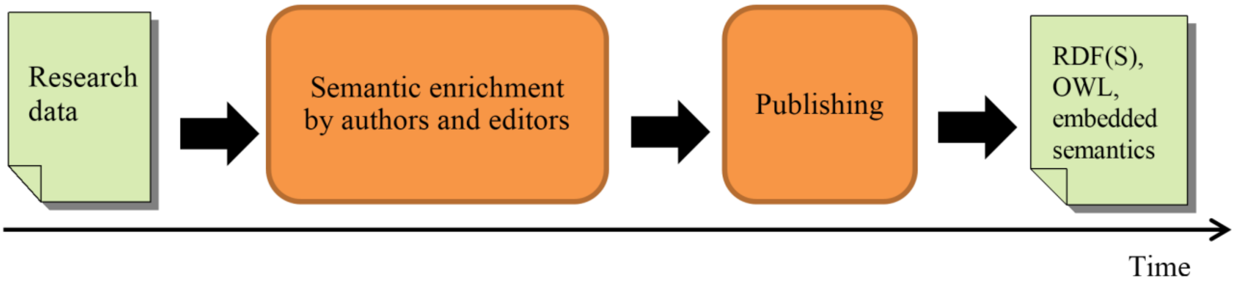

Research data
Semantic enrichment by authors and editors
Publishing
RDF(S), OWL, embedded semantics
Time

**Figure 1 - Semantic publishing as a process**

Semantic publishing can be viewed as a process which consists of the semantic enrichment of scholarly data (publications, datasets, etc.), with the aim to improve openness and usability, including the semantic enrichment in the form of machine-readable metadata.

### Semantic enrichment

Semantic enrichment can happen at various stages during the lifecycle of research. Rather than seeing it as the responsibility of the author only, semantic enrichment can be distributed between various roles, each concentrating on the content they are best equipped to annotate (Shotton, 2009). E.g. authors know the discourse of their work best. As such they are best equipped to semantically annotate the motivation, hypothesis, claims, citation contexts and various other discourse elements of their research. On the other hand, technical editors at publishing houses are experts within their respective domain and are best equipped to annotate domain specific concepts within the publication. In addition, they are also the most likely to have access to automatic processing tools that would facilitate such semantic annotation, whilst freeing the authors of the need to have familiarity with the subject. It is the role of the publishers to later make this data available in a machine-readable format.

According to the above mentioned model, it can be the role of the authors to semantically annotate the rhetorical discourse of their research which has its own particular characteristics compared to other document structures. Several ontologies have been developed for this purpose (Iniesta & Corcho, 2014) with various levels of granularity – from the identification of the main rhetorical blocks within the publication to the fine-level semantic modelling at a sentence level with its own subgoals and claims. Some of these ontologies are more suitable for manual annotation than others that are very detailed and better suited for application in natural language processing. It has been assumed in this work that any higher level automatic semantic tagging is something that if it is done is unlikely to happen at the author level and falls more into the domain of technical editors. Core Scientific Concepts (Liakata, Saha, Dobnik, Batchelor, & Rebholz-Schuhmann, 2012) (CoreSC) is an example of such an ontology, in addition it has been successfully trained and used in the automatic annotation of a corpus of papers from Biochemistry and Chemistry. The SWAN ontology[1] is an example of scientific discourse

[1] http://www.w3.org/TR/hcls-swan

modelling in neuromedicine that also includes means for relationship modelling between discourse elements. The latter is also supported via the positive example of a Domeo[2] annotation toolkit – a web application for creating and sharing annotations in HTML and XML format. The Discourse Elements Ontology[3] also describes the main rhetorical elements of a scholarly publication (some borrowed from SALT), e.g. evaluation, methods, results, conclusions, etc. One way of making use of such annotations is the ability to query publications based on similar motivations etc.

Separate ontologies are dedicated to citations. The Citation Typing Ontology[4] (CiTO) allows the authors to put their citations into context. The ontology has also been integrated into the SWAN Discourse Relationships Ontology and allows one to express what the context and motivation of the citation is – whether it agrees with, contradicts with, reviews etc. with the cited source.

However, despite the variety of ontologies developed that are suitable for scientific publishing, bottom-up mark-up is costly and time consuming and is therefore unlikely to be embraced by the authors who lack the means and skills to do this properly. This is where the input of the editors is so important – it is through their semantic annotation that a research publication would get more coverage beyond the mere basics of citations and rhetoric discourse, a task better suited for authors. Whilst intelligent text mining and natural language processing tools exist that help to recognise concepts for semantic mark-up, some human supervision is still needed to ensure the accuracy of automatic processing, including the further training of learning algorithms. It is up to the technical editors of publishing houses to decide what kind of ontologies to apply, e.g. with the help of the GATE framework[5].

Another important side of the editors' work would be the automatic classification of research papers - one of the challenges of the scientific community. Even though various manual classification systems exist in different academic disciplines, their use is incoherent and nor are they revised frequently enough to reflect new trends and concepts in science. Fixed classification systems need to be replaced by automatic ones. Luckily, there is an increased trend to move towards automatic semantic categorisation, the most interesting approaches of which are also able to describe semantic relationships between research fields beyond the classic subsumption and equivalency. Current approaches to solving this problem via automatic classification include co-occurrence analysis, keyword analysis and use of domain models. In order to understand what is going on at a given moment in time within a specific research discipline, one must have in place some ways of identifying important events and entities within the research area, including relationships between them. Osborne et al. (Osborne & Motta, Mining Semantic Relations between Research Areas, 2012) claims that keywords associated with academic publications lack structure and are often noisy. This means e.g. that mere keyword based categorisation of research is insufficient in the context of semantic publishing.

The Klink algorithm is trying to fill this semantic gap in research by connecting scholarly publications tagged with keywords to a structured ontology of research topics that it relates to. Three types of semantic relationships are detected within the ontology and made available for exploitation: *skos:broaderGeneric* (topic T1 is a subtopic of T2), *relatedEquivalent* (identifies alternative names for the same topic), *contributesTo* (research topic T1 contributes to topic T2 whilst T1 is not a subtopic of T2). Exploitation of this ontology can make a dramatic difference

[2] http://swan.mindinformatics.org/
[3] http://purl.org/spar/deo
[4] http://purl.org/spar/cito
[5] https://gate.ac.uk/

in the quantity of research that becomes visible to semantic search and recommendation engines. Also, the fact that the algorithm is capable of adapting to new trends in research topics means that less manual effort goes into the maintenance of this ontology. A number of useful services can be implemented that exploits such knowledge as demonstrated by the tool Rexplore (Osborne & Motta, Understanding Research Dynamics, 2014), that allows the investigation of research trends along the temporal axis, clustering of authors based on common interests and shared academic trajectories. However, the fact that Klink uses statistical approach in its algorithm, means that in order to get unbiased inferences, one needs to have a very large corpus of scholarly documents. Authors are unlikely to have access to such data and hence it is something that is best cured by publishers themselves.

### Publication of Scholarly Literature

An increasing number of publishers are making their publications available in a better machine-readable format than the static PDF. However, overall only a few publishers make use of semantics. E.g. the Public Library of Science (PLoS), a project dedicated to open access, provides access to raw XML versions of their publications conforming to the U.S. National Library of Medicine Document Type Definition (NLM DTD) that includes metadata about basic bibliographic fields and citations. Also the HTML versions of their publications include the embedded semantic metadata, a result of applying a style sheet to the aforementioned XML. Another example is The Royal Society of Chemistry whose journals include links to semantic concept definitions and chemical compounds in the HTML and PDF versions of some of their articles and is thought to be the first major application of Semantic Web technologies in science publishing (Shotton, 2009). This initiative, originally known as the RSC Project Prospect, won the 2007 ALPSP/Charlesworth Award for pioneering work in semantic publishing and is now integrated within their routine publication processes (RSC Semantic Publishing). Even though publishers themselves still have a long way to go in terms of making metadata available via SPARQL endpoints, basic bibliographic data is starting to become available[6]. As of yet RSC has issued a press release (RSC opens up data for global scientific community via TSO's OpenUp® Linked Data platform, 2013) dating back to 2013 about their plans to open a dedicated RSC Linked Data Portal with a SPARQL endpoint but work is still in progress and outcome yet to be seen.

The reality is that semantic publishing is still in its pioneering status. Part of this is down to publishers still looking to define their role in the semantic publishing world, whilst coming up with models that would allow them to keep their operation cost-effective. Another aspect is the lack of solutions – whilst there is a clear push towards the openness and public availability of research data in order to promote transparency and reproducibility of research data, there are no clear guidelines or standards as to who exactly is responsible for this and how it is supposed to be done (International Association of Scientific, Technical and Medical Publishers, 2007). Till certain standards have been formed, some scientist who sympathise more with the cause, publish their research data in digital repositories like FigShare[7]. There is still a need for solutions to support the semantic publishing of research, with the initiative either coming from the publishers themselves or some regulative organisation. As proof of more concrete efforts the term “Science 2.0” emerged within the research community with emphasis on open data and open access. These trends have been recognised e.g. by the European Commission and its policy makers and there are active discussions and workshops relating to open access happening in the autumn of 2014

[6] http://dblp.rkbexplorer.com/

[7] http://figshare.com/

with direct impact on policies due to be published by the end of 2014 (Geoghegan-Quinn, 2014). The European Commission has applied some of the “Science 2.0” concepts into its future programmes, e.g. open access to scientific publications is mandatory for Horizon 2020 programme (European Commission). Though the latter does not specify in what format the data should be made available, a new element called Data Management Plans (DMP) expects scientists to name what data the project generates and how it is made accessible (European Commission, 2013). This acknowledgment of the existing drive for openness and accessibility within the research community by the European Commission policy makers could be considered as paving the road for semantic publishing in the future.

# 3 Problem Description

Various aspects of semantic publishing are currently difficult to implement. There is a lack of availability of generic tools, knowledge and cooperation between the various parties that should be participating in the effort according to the vision of semantic publishing.

This section thereby presents an overview of problems that have not been sufficiently addressed so far in the context of scholarly publishing and which offer the motivation for the current work, namely the lack of support for the PDF format, limited tool support for multiple vocabularies or support beyond typed annotations. Below paragraphs will detail each problem further

**Lack of Support for PDF Format**

Earlier research has mainly concentrated on the semantic annotation of information in XML[8] or HTML[9] format, the semantic annotation of PDF documents has not been sufficiently addressed. As explained by (Shotton, 2009): "[PDF document] is antithetical to the spirit of the Web, being static rather than interactive, and difficult for machines to read, thus inhibiting the development of services that can link information between articles." As a result, there are currently only a few publicly available tools for this task, yet the majority of scholarly papers are still published in the PDF format. On the positive side, there is a tool for displaying semantic content in PDF documents, namely the Utopia Documents[10] (Attwood, Kell, McDermott, Marsh, Pettifer, & Thorne, Utopia documents: Linking Scholarly Literature With Research Data, 2010), however it does not support creating semantic annotations and is limited to private notes in free text form only. Of the tools that do support semantic annotations, namely GoNTogle (Bikakis, Giannopoulos, Dalamagas, & Sellis, 2010) and PDFTab (Eriksson, 2007), none are fully suitable for the semantic annotation of scholarly papers in the scope defined, yet it is clear that there is a need for such a tool in the field of semantic publishing. GoNTogle's suitability in this context is hindered by its limited support of other ontologies besides that of categorisation vocabularies like ACM[11]. PDFTab on the other hand is a desktop solution and a Protégé[12] plugin that stores semantic data within the PDF itself. However storing annotations within the PDF documents is not in line with the Linked Data approach, according to which such data should be publicly made available in a standard format through a SPARQL endpoint. Hence the latter requires that annotations must be kept separate from PDFs on a central server where everyone can have access to it.

Till the majority of publishing houses and authors publish their research in the PDF format, the researchers miss out on a lot of scholarly information by excluding support for it in the annotation tools. As Pettifer et al. argue (Pettifer, McDermott, Marsh, Thorne, Villeger, & Attwood, 2011), one should not be so fixated on the representation format of a scholarly article. The fact cannot be ignored that the PDF format has been around since the early 1990s and a vast amount of already published scholarly papers can only be accessed in that format. At the current state when PDF is still the standard format of publishing papers, it is unrealistic to expect that authors spend effort to learn how to semantically annotate their works in more portable formats

[8] Extensible Markup Language, defined by the W3C's XML 1.0 Specification.
[9] HyperText Markup Language
[10] http://utopiadocs.com/
[11] ACM Computing Classification System at http://www.acm.org/about/class/
[12] Open source ontology editor, http://protege.stanford.edu/

like the XML when it is not even clear where and how that meta-information will be published by the publishing house.

### Limited Availability of Tools that Support Multiple Vocabularies

There is a general lack of freely available simple annotation tools that is not of specialised use and that does not limit the user to some specific ontology of a limited domain. Even better, if such a tool would not need to be installed or configured or only meant for the internal creation and consumption of semantic data. The more obstacles one puts on the way of end users before they can create semantic annotations, the less participation one will have. An architecture that supports the semantic annotation of scholarly articles should be able to support the application of any online annotation vocabulary. Science is a field that often requires the use of very concise vocabularies and new ones get constantly developed. A general purpose annotation tool that can be used for any scholarly paper, has a higher likelihood of making a valuable contribution to the semantic publishing field. E.g. new links could be discovered between ontologies that were not linked before. Domeo, a web-based annotation framework for online HTML and XML documents is a good example of such an effort (Ciccarese, Ocana, & Clark, Open Semantic Annotation of Scientific Publications Using DOMEO, 2012) that started off with a focus on biomedicine. Its new version v.2 is planned to be released in January 2015 so it is currently unclear how flexible it will be in its support of other ontologies due to currently ongoing major changes. This is an important aspect if one were to consider how different science disciplines tend to have their own vocabulary even if they mean the same thing. Neither is there any future for a single consistent Ontology of Everything, an approach that does not scale and is in deep contrast to how people actually use ontologies (Shadbolt, Hall, & Berners-Lee, 2006). Increasing the visibility of information across scientific disciplines can only result in more innovation and that is best supported by tools that support a wide range of ontologies so that previously undiscovered links between ontologies can be easily identified.

### Support beyond Typed Annotations

Simple classification vocabularies are not enough to represent more complex knowledge or relationships between them. Examples of such use cases are citation links between papers, citation contexts (CiTO), modelling arguments (Argument Model Ontology[13]), etc. Very valuable information for scientists can be formulated with such ontologies, information that would then become available for inferencing. When searching for scholarly papers via keywords, the real question that the researcher might be asking is “find papers motivated by […]”, “find papers that refer to the problem of […]”. Keyword based searches do not perform well on such questions but with the help of the correct ontology, such queries could be performed on semantically annotated papers. Hence this work argues that an annotation tool for scientific papers must be able to support more complex vocabularies such as those with properties in order to allow users to ask questions typical to the research field. The problem with the current state of art is that most tools are designed for the application of classification vocabularies or only support a small subsection vocabularies developed for expressing relationships between concepts.

[13] See http://www.essepuntato.it/2011/02/argumentmodel

## Conclusion

As emphasised by Attwood et al. (Attwood, Kell, McDermott, Marsh, Pettifer, & Thorne, Calling International Rescue: Knowledge Lost in Literature and Data Landslide, 2009), as long as we continue to fail in getting the most from scholarly literature, “we will continue to fail to know what we already know” and do science a disservice. This work contributes to the effort of making the most use of the scholarly literature.

# 4 Description of the Solution

A prototype[14] has been created that addresses the problems outlined in chapter 3 - issues that need addressing within the semantic publishing community. This prototype is the further development of the SemAnn project, a semantic annotation tool for PDF documents. The current work is not the only development within the SemAnn project – others have also built on it, creating another extension[15] independent from this development that focuses on the semantic annotation of tables[16], whilst this prototype focuses on the semantic annotation of text. This solution will be described and discussed in the following subsections in relation to how it addresses the requirements outlined in chapter 4.1.

## 4.1 Requirements

The requirements for the tool have been derived from the identification of some of the problem areas in the field of semantic publishing (see chapter 3). A few additional requirements have been added that either helped to better focus the work or were considered to be important with respect to the principles of Linked Data.

The architecture for the semantic annotation of scientific papers implemented with this thesis must hence satisfy the below requirements.

**Base requirements are:**

(a) ability to semantically annotate text in PDF documents
(b) support for multiple ontologies
(c) support for complex vocabularies with properties

**Additional requirements include:**

(d) an implementation of a recommender functionality of similar papers
(e) support for multiple users
(f) adherence to the requirements of Linked Data

The reasons for including these additional requirements are outlined below.

Requirement (d). The recommendation functionality is to be implemented for the following reasons:

- It helps to focus the development on the usefulness of the architecture created. The implemented architecture needs to be flexible and support a lot of different semantic annotation use cases as part of base requirements. In order to not get carried away with designing overly complex data models, a clear focus on how the data model will be later on used in the semantic querying process was considered to be necessary.
- In order to test some interesting hypotheses outlined in chapter 5.2.1.
- In order to observe the correlation between the number of pre-annotated papers in the database and the number of recommendations returned (see chapter 5.2.2).

Requirement (e). Bottom-up semantic tagging of information is a very laborious process. Entity recognition tools exist but human involvement is still necessary. Accuracy is important in

[14] See https://github.com/AKSW/semann
[15] To be merged with the SemAnn project, see details: https://github.com/saifulnipo/eis-semantic-annotation
[16] See https://github.com/saifulnipo/eis-semantic-annotation/wiki

scholarly papers and automated semantic tagging without human involvement is not yet of the maturity needed. An alternative approach to help with the semantic annotations would hence be crowdsourcing that data. It is therefore important that the tool can be used by multiple users and that the data is collected into a central database where it can benefit everyone.

Requirement (f). The principles of linked data should be followed as a fundamental part of what enables semantic publishing. Hence, the use of URIs, RDF and interlinking with other data are fundamental to the architecture.

## 4.2 Overview of the Functionality of the Tool

A high-level overview of the prototype and its main use cases is now presented for the reader's convenience.

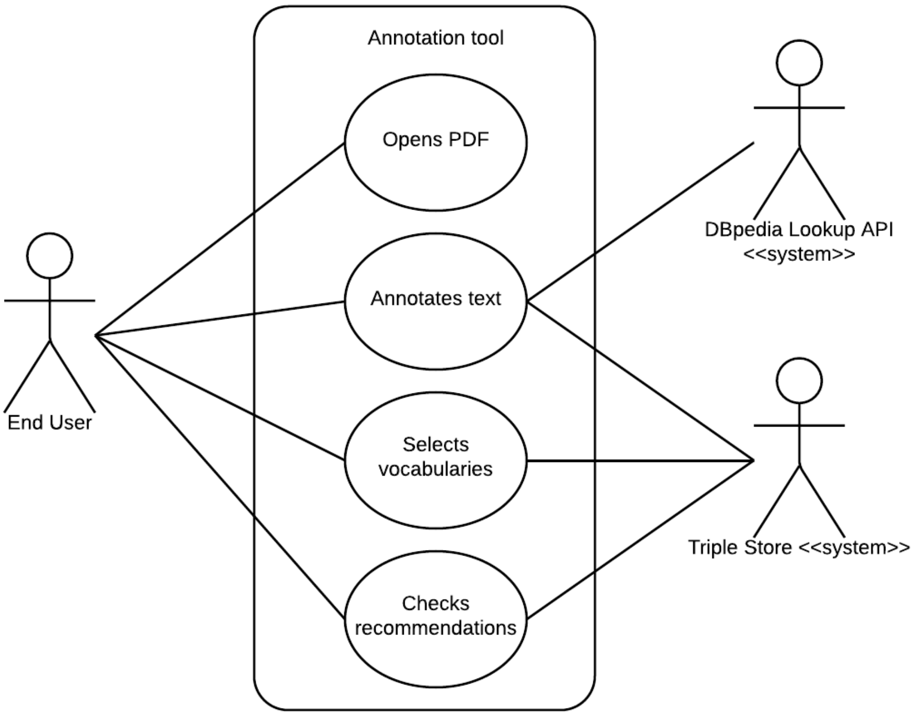
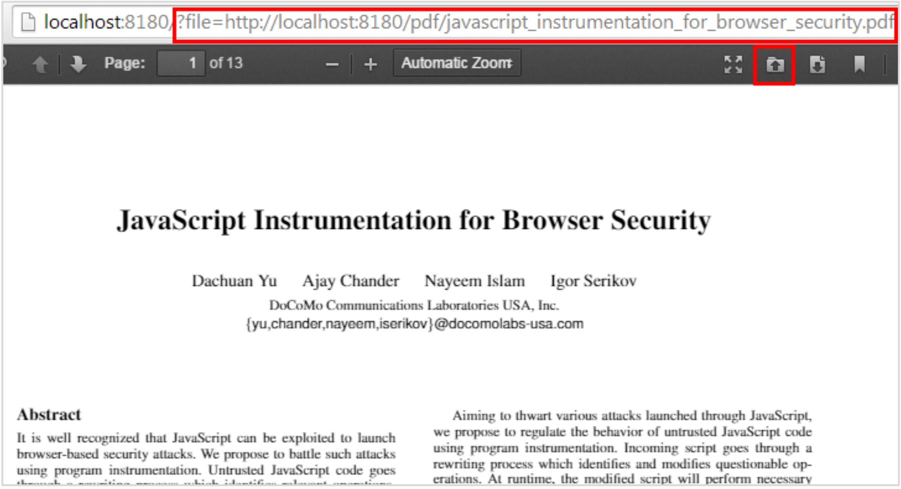


**Figure 2 - Main use cases of the prototype**

**Use case "Opens PDF"**
Opening of a PDF file in the tool can either be done via the "Open File" menu or by appending the URL of the file to the "file" parameter of the tool.


localhost:8180/?file=http://localhost:8180/pdf/javascript_instrumentation_for_browser_security.pdf
Page: 1 of 13
Automatic Zoom

JavaScript Instrumentation for Browser Security

Dachuan Yu Ajay Chander Nayeem Islam Igor Serikov
DoCoMo Communications Laboratories USA, Inc.
{yu,chander,nayeem,iserikov}@docomolabs-usa.com

Abstract
It is well recognized that JavaScript can be exploited to launch browser-based security attacks. We propose to battle such attacks using program instrumentation. Untrusted JavaScript code goes

Aiming to thwart various attacks launched through JavaScript, we propose to regulate the behavior of untrusted JavaScript code using program instrumentation. Incoming script goes through a rewriting process which identifies and modifies questionable operations. At runtime, the modified script will perform necessary

**Figure 3 - Opening of a PDF document with the tool**

## Use Case "Annotates text"

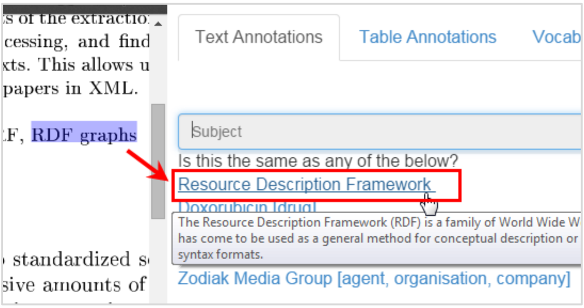
There are different ways one can annotate text based on the type of annotation one wants to create. The general workflow of semantic annotating is the following:

1. User selects some text from the PDF document
2. User adds some semantic enrichment to the annotation
3. User clicks on "Add annotation"

Variations in the semantic enrichment step (2) are dependent on:

- The type of annotations the user wants to insert.
  - (a) Simple annotations, i.e. annotations that are instances of some class. This is the main use case for annotating for classification purposes. See example in Figure 5.
  - (b) Complex annotations, i.e. an annotation that is part of a relationship (see example in Figure 8).
- What kind of semantic tagging the user applies to the annotation:
  - (a) DBpedia resources
  - (b) classes and / or properties from a selected vocabulary (see example in Figure 8)

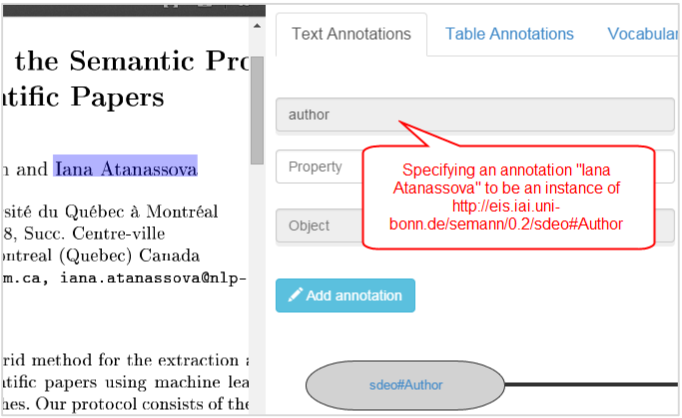
(a) annotation of type DBpedia resource



The Resource Description Framework (RDF) is a family of World Wide W has come to be used as a general method for conceptual description or syntax formats.

**Figure 4 - Specifying selected text to be an instance of a DBpedia resource**

(b) annotation of type ontology class



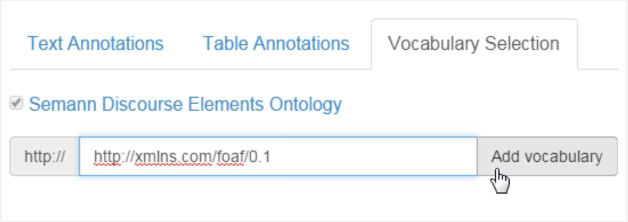
Specifying an annotation "Iana Atanassova" to be an instance of http://eis.iai.uni-bonn.de/semann/0.2/sdeo#Author

**Figure 5 - Specifying selected text to be an instance of some class**

**Use Case "Selects Vocabularies"**

By default, the SemAnn Discourse Elements Ontology (see chapter 4.4.1.2) is loaded into the tool, but users can also specify any other vocabularies they might want to use.

(a) Type vocabulary URL

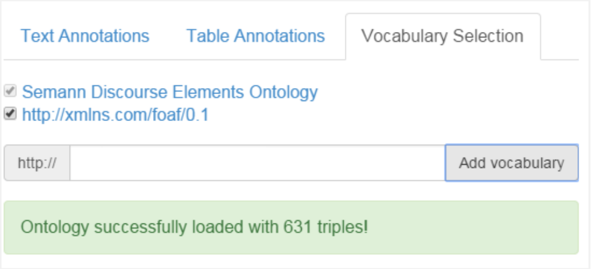


☑ Semann Discourse Elements Ontology

http:// http://xmlns.com/foaf/0.1

Figure 6 - Add vocabulary functionality

(b) Click on "Add vocabulary"



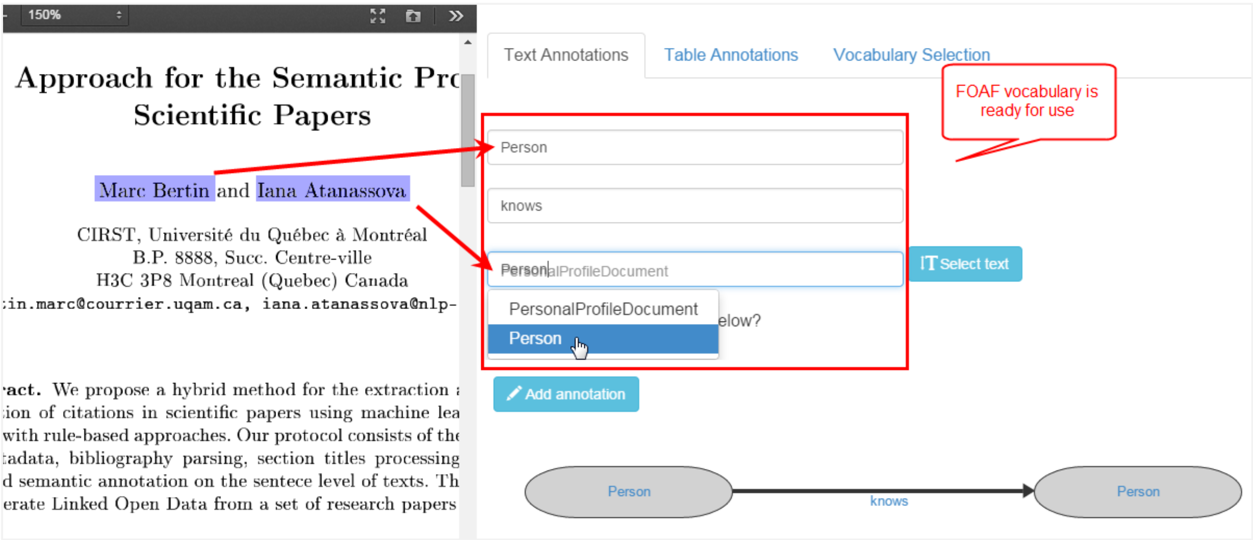
☑ Semann Discourse Elements Ontology
☑ http://xmlns.com/foaf/0.1

http://

Ontology successfully loaded with 631 triples!

Figure 7 - Active vocabulary selection

(c) Apply classes and properties from the vocabulary to selected text



Approach for the Semantic Pro
Scientific Papers

Marc Bertin and Iana Atanassova

CIRST, Université du Québec à Montréal
B.P. 8888, Succ. Centre-ville
H3C 3P8 Montreal (Quebec) Canada
:in.marc@courrier.uqam.ca, iana.atanassova@nlp-

:act. We propose a hybrid method for the extraction a
:ion of citations in scientific papers using machine lea
with rule-based approaches. Our protocol consists of the
:adata, bibliography parsing, section titles processing
d semantic annotation on the sentece level of texts. Th
erate Linked Open Data from a set of research papers

Person

knows

PersonalProfileDocument

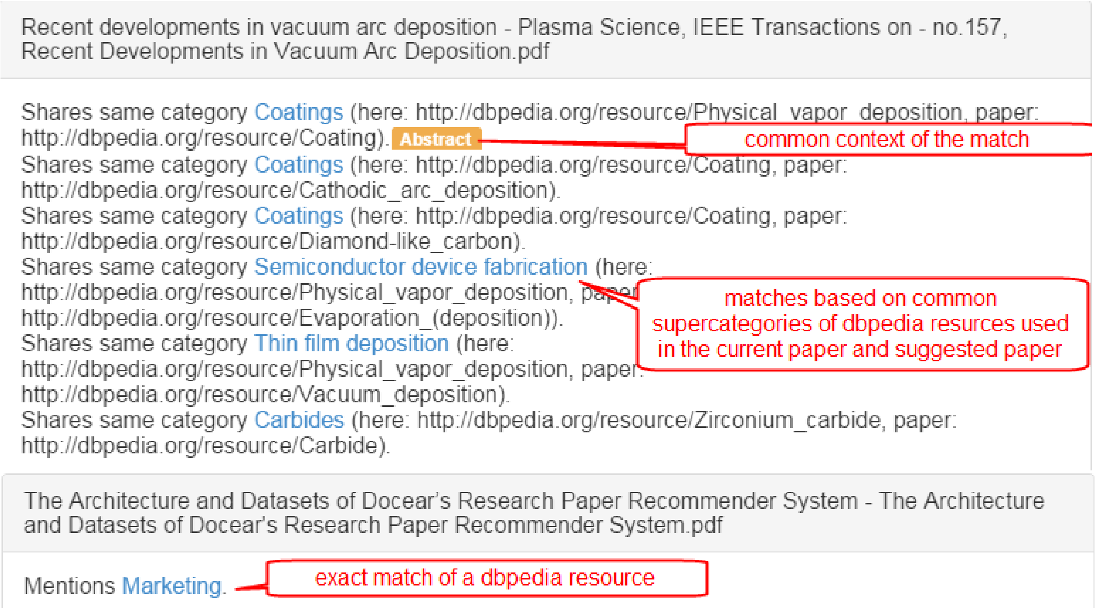
elow?

Figure 8 - Applying classes and properties from the loaded vocabulary

**Use case "Checks recommendations"**

Recommendations for similar papers are displayed as a result of semantically comparing annotations of the currently loaded document to those of all the other papers in the database (see chapter 4.4.3).

(a) User clicks on the "Find Similar" button
(b) Recommendations for other similar papers are displayed with explanations.



Recent developments in vacuum arc deposition - Plasma Science, IEEE Transactions on - no.157, Recent Developments in Vacuum Arc Deposition.pdf

Shares same category Coatings (here: http://dbpedia.org/resource/Physical_vapor_deposition, paper: http://dbpedia.org/resource/Coating). Abstract
Shares same category Coatings (here: http://dbpedia.org/resource/Coating, paper: http://dbpedia.org/resource/Cathodic_arc_deposition).
Shares same category Coatings (here: http://dbpedia.org/resource/Coating, paper: http://dbpedia.org/resource/Diamond-like_carbon).
Shares same category Semiconductor device fabrication (here: http://dbpedia.org/resource/Physical_vapor_deposition, paper: http://dbpedia.org/resource/Evaporation_(deposition)).
Shares same category Thin film deposition (here: http://dbpedia.org/resource/Physical_vapor_deposition, paper: http://dbpedia.org/resource/Vacuum_deposition).
Shares same category Carbides (here: http://dbpedia.org/resource/Zirconium_carbide, paper: http://dbpedia.org/resource/Carbide).

common context of the match

matches based on common supercategories of dbpedia resurces used in the current paper and suggested paper

The Architecture and Datasets of Docear's Research Paper Recommender System - The Architecture and Datasets of Docear's Research Paper Recommender System.pdf

Mentions Marketing.

exact match of a dbpedia resource

**Figure 9 - Recommendation functionality example**

## 4.3 Technical Limitations of the Solution

The current prototype has the following technical limitations:

- It runs on a local Virtuoso backend and not on a central server. There are ongoing efforts to set up a demo site[17] for this tool that would solve this issue. During the development phase, in order to simulate the use case of multiple users using this tool, the approach was to enable access to the relevant port of localhost over LAN.
- The prototype does not currently support loading of PDF files from another server[18]. This is due to PDF.js, a JavaScript platform for parsing and rendering PDF files, enforcing the same origin policy. There are ways around it so it is not a fundamental limitation of the architecture, but it has not been implemented in the prototype.
- The rendering of PDF files is implemented in JavaScript in PDF.js and runs in a browser. Not every feature of PDF files is supported or working correctly, but documents containing text and images work very well.
- The annotations' start and end positions are currently serialised through Rangy API[19] in relation to the PDF.js viewer window. In theory it is possible that a different version of the PDF.js might give different results for serialisations causing potential backwards incompatibility issues with the data collected so far (but this has not been verified).
- The inferencing capability of the tool is currently limited by Openlink Virtuoso's inference rules and property paths. No external reasoner is used.

## 4.4 Architecture

The general architecture of the prototype is presented in Figure 10. There is a web-based user interface in which the end-user can open a PDF document and add semantic annotations. Semantic enrichment of annotations is done through ontology selection and application or by taking advantage of the suggested resources from the linked open dataset DBpedia derived from Wikipedia. User created annotations are then stored in a RDF triplestore.

[17] Registered in https://github.com/AKSW/semann/issues/9
[18] See https://github.com/mozilla/pdf.js/wiki/Frequently-Asked-Questions#faq-xhr
[19] https://code.google.com/p/rangy/

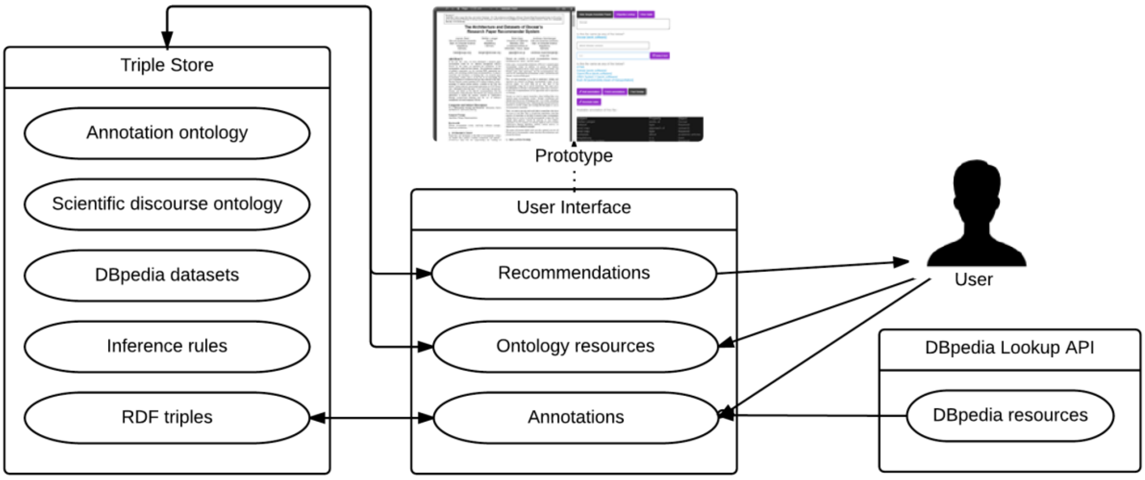


**Figure 10 - Architecture of the prototype**

Components in the architecture:

- **User interface** – the following tasks are supported:
  - creating new semantic annotations for a PDF document
  - viewing of existing annotations for a PDF document
  - viewing of recommendations for similar papers to the currently open PDF document
  - selecting and applying of vocabulary terms and properties
- **Triple Store** – annotations are stored as triples in Openlink Virtuoso triple store[20]. Virtuoso also mediates ontologies that the user might want to activate in the user interface and performs reasoning for retrieving recommendations of similar papers to the user.
- **DBpedia Lookup API** - the prototype uses the Keyword Search API of the DBpedia Lookup in the additional semantic enrichment of user annotations.

The SemAnn project extends PDF.js[21], a JavaScript platform for parsing and rendering PDF files. This design choice eliminated platform dependence and compatibility issues caused by different PDF readers that the users might have installed on their computers. This means that the end user can use the prototype via JavaScript supported browsers without the need to install additional software and get the same user experience as others. Additional information about the various libraries used in the implementation of the SemAnn project can be found in the project's wiki[22]. During the implementation of the prototype various libraries have been updated to newer versions.

### 4.4.1 Ontologies

One of the architectural goals in the development of this functionality was to enable the end-users to have maximum freedom in the type of annotations they might want to create. That means both support for simple annotations that are mere classifications and more complex annotations that describe annotations in a relationship.

[20] http://www.openlinksw.com/

[21] http://mozilla.github.io/pdf.js

[22] https://github.com/AKSW/semann/wiki/Libraries-Used

The following approaches were adopted when developing ontologies:

- flexibility of the solution in supporting various types of annotations
- minimalistic annotation ontology
- usefulness of the RDF triples in supporting the intended flexibility

The use of the ontologies mentioned in this chapter makes it possible to not only model the information about annotations themselves but also the structural context in which the annotations appear. The modelling of the structural context is a novel approach which enables us to view annotations of a paper as a hierarchy of annotations. The latter in return allows one to perform a lot more interesting queries with the SPARQL endpoint.

Two ontologies were used in this prototype as a result of the requirements outlined in chapter 4.1. In this chapter the design choices for both of these ontologies are explained in detail with examples.

- SemAnn Annotation Ontology – this lightweight ontology was specifically developed for the given prototype to model information related to annotations.
- SemAnn Discourse Elements Ontology (SDEO) – this ontology extends the existing Discourse Elements Ontology[23] (DEO), an ontology for describing the major rhetorical elements of a scholarly paper, itself part of SPAR, the Semantic Publishing and Referencing Ontologies[24]. It is used for the identification of annotations of special interest to the scientific community and comes preloaded in the prototype tool. The properties in this ontology serve a special purpose within the architecture in the creation of the hierarchical model.

Whilst the above ontologies are incorporated into the architecture of the system, end-users are not limited to the use of these ontologies only. In fact, additional ontologies can be loaded by the end-user via the tool's user-interface which calls upon Virtuoso's sponger service for extracting the triples from the specified ontology and makes them available for use in the user interface (see Figure 8).

#### 4.4.1.1 SemAnn Annotation Ontology

The design goal for this ontology was to model the information about annotations and represent this in a compact format. The emphasis on the compactness was driven by the aim of keeping SPARQL queries simple, an aspect that is relevant when making the SPARQL endpoint publicly available.

Annotation Ontology (AO) (Ciccarese, Ocana, Castro, Das, & Clark, 2011), an open ontology in OWL-DL for annotating scientific documents, was also considered for modelling the annotations, however for the purposes of this prototype it was discarded for being too heavy-weight. In order to express the same information that is currently encoded into the URI of an annotation in the SemAnn annotation ontology, one needs 5 triples in the AO. A simpler and minimalistic approach was hence favoured for the implementation of this specific architecture instance. However, potential integration with AO is something that is not excluded from future work.

[23] http://www.essepuntato.it/lode/http://purl.org/spar/deo
[24] http://sempublishing.sourceforge.net/

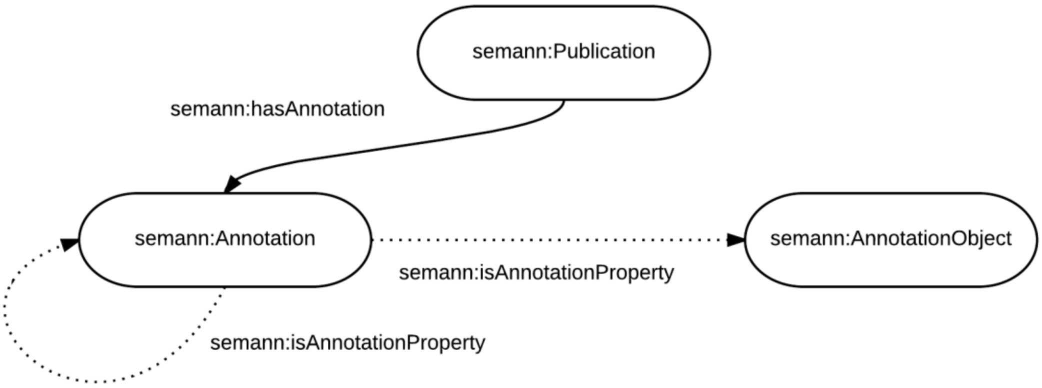
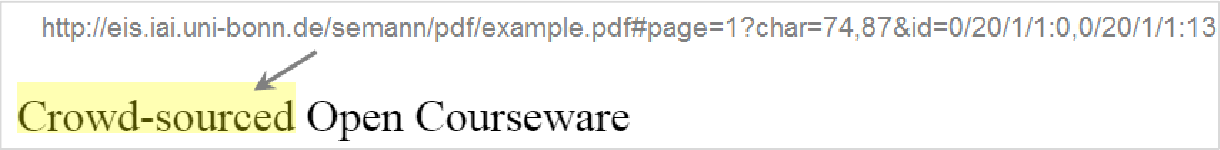


**Figure 11 - SemAnn Annotation Ontology**

The reason why it was possible to keep the SemAnn Annotation Ontology[25] so simple is largely down to a clever selection of the annotation URI (instance identifier for semann:Annotation).



http://eis.iai.uni-bonn.de/semann/pdf/example.pdf#page=1?char=74,87&id=0/20/1/1:0,0/20/1/1:13

Crowd-sourced Open Courseware

**Figure 12 - URI composition of the annotation**

Each annotation instance URI is composed so that it can easily be used as a shareable public link. A link that could open the relevant PDF document in the SemAnn tool, on the relevant page and highlight the annotation in question. In the given example, the path to the PDF file does not necessarily need to reflect the physical location of the PDF file but rather act as an alias for redirection to the physical location of the file. This is necessary in order to make the annotation link open up within the SemAnn tool.

- The “page” reference is a standard parameter for opening a PDF file on a specific page[26].
- The “char” parameter refers to the start and end positions of the annotation with respect to the start of the document. This information is crucial for allowing us to view annotations as a hierarchy from which additional contextual information can be deduced for reasoning purposes.
- The “id” parameter is a convenience parameter used by the Rangy API[27] library, that is used in the identification and highlighting of text fragments.

An instance of the annotation can inherit from multiple classes and this approach is used in the semantic enrichment of annotations. It also makes SPARQL queries simpler. E.g. the below triple statements state that the annotation in Figure 12 is also an instance of http://dbpedia.org/resource/Crowdsourcing:

```
<rdf:Description rdf:about="http://eis.iai.uni-
bonn.de/semann/pdf/exmple.pdf#page=1?char=74,87&amp;id=0/20/1/1:0,0/20/1/1:13">
    <rdf:type rdf:resource="http://dbpedia.org/resource/Crowdsourcing" />
    <rdf:type rdf:resource="http://eis.iai.uni-bonn.de/semann/0.2/owl#Annotation" />
    <rdfs:label xml:lang="en">Crowd-sourced</rdfs:label>
</rdf:Description>
```

[25] https://github.com/AKSW/semann/blob/mergebranch/ontologies/semann.0.2.owl.ttl
[26] http://www.adobe.com/content/dam/Adobe/en/devnet/acrobat/pdfs/pdf_open_parameters.pdf
[27] https://code.google.com/p/rangy/

```
<rdf:Description rdf:about="http://eis.iai.uni-bonn.de/semann/pdf/example.pdf">
    <rdf:type rdf:resource="http://eis.iai.uni-bonn.de/semann/0.2/owl#Publication" />
    <semann:hasAnnotation rdf:resource="http://eis.iai.uni-
bonn.de/semann/pdf/example.pdf#page=1?char=74,87&amp;id=0/20/1/1:0,0/20/1/1:13" />
    <rdfs:label xml:lang="en">Example Document</rdfs:label>
</rdf:Description>
```

**Figure 13 - RDF/XML representation of the annotation as an instance of multiple classes**

More complex constructs involving relationships are also possible. One of the requirements of the architecture was to allow the user to define relationships involving annotations. Here are some use cases for the more complex annotations that the tool supports, expressed in the triple format:

### Use case A

This use case represents a relationship between two annotation instances. Some examples of how it could be used in the context of scholarly papers:

- <A.1> <http://purl.org/spar/cito:disagrees_with> <B.1>
  <A.2> <http://www.essepuntato.it/2011/02/argumentmodel:proves> <B.2>
  <A.3> <http://purl.org/swan/1.2/swan-commons/citesAsSupportiveEvidence> <B.3>
  <A.4> <http://purl.org/spar/cito:plagiarizes> <B.4>

  This type of construct is well suited for describing the scientific discourse in a paper, building citation links, characterising citations (e.g. CiTO[28] ontology), describing experiments, etc. The annotation in the object position of the triple does not necessarily need to belong to the same paper. It can point to the URI of an annotation in another paper (see Figure 12). In this way one can easily create interesting relationships between specific text fragments of different papers. E.g. instead of creating citation links between papers which has been the approach used so far, one could be more specific and reference the specific block of text on which the citation was based on within the cited paper. That would reduce the amount of time needed to locate the necessary information.

  The following is an example of two annotations and a property connecting them (based on the example in Figure 8):

```
<rdf:Description rdf:about="http://eis.iai.uni-bonn.de/semann/pdf/example.pdf">
<rdf:type rdf:resource="http://eis.iai.uni-bonn.de/semann/0.2/owl#Publication" />
<semann:hasAnnotation rdf:resource="http://eis.iai.uni-
bonn.de/semann/pdf/example.pdf#page=1?char=158,186&amp;id=0/2/1/1:31,0/2/1/1:59" />
<semann:hasAnnotation rdf:resource="http://eis.iai.uni-
bonn.de/semann/pdf/example.pdf#page=1?char=127,149&amp;id=0/2/1/1:0,0/2/1/1:22" />
<rdfs:label xml:lang="en">Example Document</rdfs:label>
</rdf:Description>
<rdf:Description rdf:about="http://eis.iai.uni-
bonn.de/semann/pdf/example.pdf#page=1?char=127,149&amp;id=0/2/1/1:0,0/2/1/1:22">
<rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person" />
<rdf:type rdf:resource="http://eis.iai.uni-bonn.de/semann/0.2/owl#Annotation" />
<foaf:knows rdf:resource="http://eis.iai.uni-
bonn.de/semann/pdf/example.pdf#page=1?char=158,186&amp;id=0/2/1/1:31,0/2/1/1:59" />
<rdfs:label xml:lang="en">Marc Bertin</rdfs:label>
</rdf:Description>
<rdf:Description rdf:about="http://eis.iai.uni-
bonn.de/semann/pdf/example.pdf#page=1?char=158,186&amp;id=0/2/1/1:31,0/2/1/1:59">
<rdf:type rdf:resource="http://eis.iai.uni-bonn.de/semann/0.2/owl#Annotation" />
<rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person" />
<rdfs:label xml:lang="en">Iana Atanassova</rdfs:label>
</rdf:Description>
<rdf:Description rdf:about="http://xmlns.com/foaf/0.1/knows">
<rdf:type rdf:resource="http://eis.iai.uni-bonn.de/semann/0.2/owl#isAnnotationProperty" />
<rdfs:label xml:lang="en">knows</rdfs:label>
</rdf:Description>
```

**Figure 14 - RDF/XML representation of a relationship between two annotations**

[28] http://purl.org/spar/cito

**Use case B**
This use case represents a relationship between an annotation and some class. Some examples of how it could be used in the context of scholarly papers:

-
  ```
  <A.5> <http://purl.org/spar/cito:confirms>
  <http://projectX.org/owl#Experiment1>
  <A.6> <http://eis.iai.uni-bonn.de/semann/userprofileX/ontologyX#isBackgroundTo>
  <http://eis.iai.uni-bonn.de/semann/userprofileX/ontologyX#MyPhdTopic>
  ```

  This construct is highly suitable for flexible reference management (similar to Citavi[29] in concept but using semantics). E.g. a group of researchers might be collaborating on a crowdsourced project X. In that case it might be decided that a custom ontology for the project would help organise the research being done. One could then connect that research (as annotations in papers) to that public ontology. The second example is based on a hypothetical user profile of the tool (not currently implemented) in which case logged in users could create their own private ontologies on the fly to organise their research. Whatever the specific use case here, the major advantage will be the flexibility in querying offered by the SPARQL endpoint. E.g. one could compose SPARQL queries that retrieve annotations that are instances of specific classes (e.g. “MyClassA” and “MyClassB” and not “MyClassC”) or participate in a specific relationship. Such flexibility can only serve as extra motivation in using such a tool.

  The following is an example of a relation between an annotation and a class:

```
<rdf:Description rdf:about="http://eis.iai.uni-bonn.de/semann/pdf/example.pdf">
    <rdf:type rdf:resource="http://eis.iai.uni-bonn.de/semann/0.2/owl#Publication" />
    <semann:hasAnnotation rdf:resource="http://eis.iai.uni-
bonn.de/semann/pdf/example.pdf#page=13?char=10,50&amp;id=0/2/1:3,0/2/1:5" />
    <rdfs:label xml:lang="en">Example Document</rdfs:label>
</rdf:Description>
<rdf:Description rdf:about="http://eis.iai.uni-
bonn.de/semann/pdf/example.pdf#page=13?char=10,50&amp;id=0/2/1:3,0/2/1:5">
    <rdf:type rdf:resource="http://eis.iai.uni-bonn.de/semann/0.2/owl#Annotation" />
    <cito:confirms rdf:resource="http://projectX.org/owl#ExperimentResult" />
    <rdfs:label xml:lang="en">These results from our experiment are the
following...</rdfs:label>
</rdf:Description>
<rdf:Description rdf:about="http://purl.org/spar/cito/confirms">
    <rdf:type rdf:resource="http://eis.iai.uni-
bonn.de/semann/0.2/owl#isAnnotationProperty" />
    <rdfs:label xml:lang="en">confirms</rdfs:label>
</rdf:Description>
```

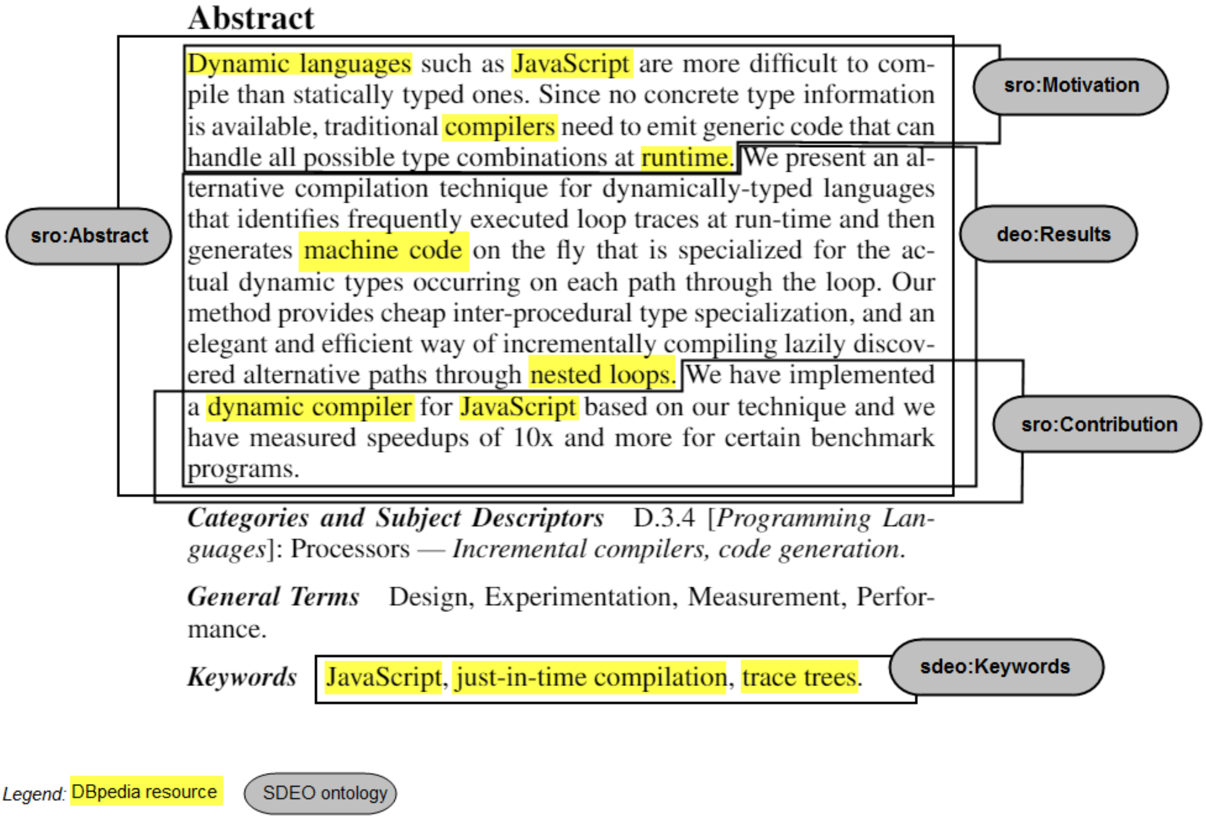
**Figure 15 - RDF/XML representation of a relationship between an annotation and a class**

### 4.4.1.2 SemAnn Discourse Elements Ontology

Results from the evaluation of the recommender functionality showed how only a few semantic annotations from DBpedia (similar to Figure 13) per paper can result in a surprisingly wide coverage of similar papers. This means that semantic searches based on DBpedia subject categories can return a lot of matches and it quickly becomes important to be able to differentiate which of these matches are likely to be more relevant to the user. Consider the example of one single semantic annotation of type http://dbpedia.org/resource/Marketing in the future work section of a paper from the engineering field. Recommendations that are returned based on only this single annotation that is not even representative of the paper in question are not likely to be useful. One way of resolving that issue is to consider the context of the annotation. When one

[29] https://www.citavi.com/

considers what is relevant context in scientific papers, then often it is the context that represents scientific discourse: motivation, claims, and problem statements. Hence, by encouraging the users to identify fragments of scientific discourse can put other annotations within it into much more useful context. The above reasoning led to incorporating the SDEO ontology into the architecture of the tool and preloading it in the user interface for convenient use.



**Abstract**

Dynamic languages such as JavaScript are more difficult to compile than statically typed ones. Since no concrete type information is available, traditional compilers need to emit generic code that can handle all possible type combinations at runtime. We present an alternative compilation technique for dynamically-typed languages that identifies frequently executed loop traces at run-time and then generates machine code on the fly that is specialized for the actual dynamic types occurring on each path through the loop. Our method provides cheap inter-procedural type specialization, and an elegant and efficient way of incrementally compiling lazily discovered alternative paths through nested loops. We have implemented a dynamic compiler for JavaScript based on our technique and we have measured speedups of 10x and more for certain benchmark programs.

***Categories and Subject Descriptors*** D.3.4 [*Programming Languages*]: Processors — *Incremental compilers, code generation.*

***General Terms*** Design, Experimentation, Measurement, Performance.

***Keywords*** JavaScript, just-in-time compilation, trace trees.

**Figure 16 - Annotations as instances of different vocabularies**

The annotations in the above example could have been added by different end-users at different times. The approach implemented within the prototype keeps track of the current hierarchical structure of the document in a separate graph in the triplestore:

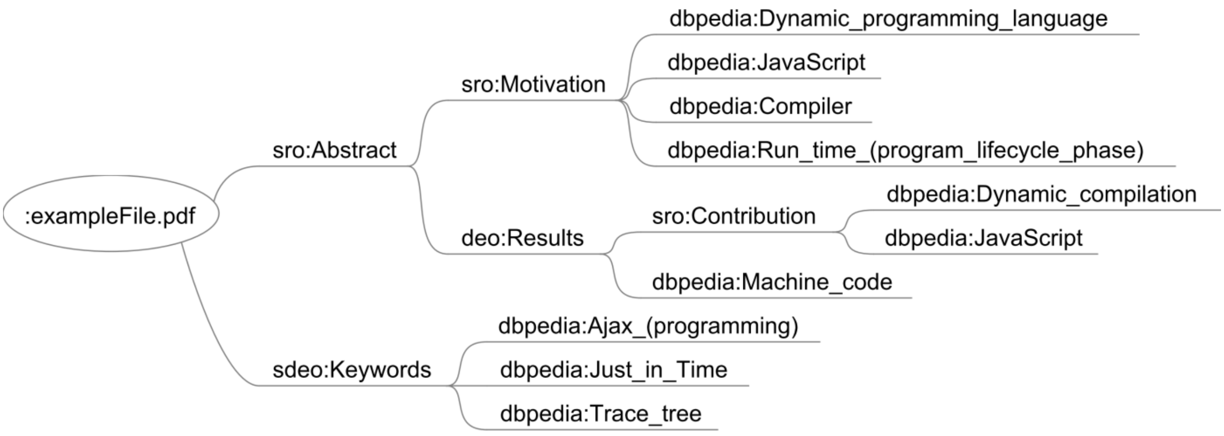


**Figure 17 - Hierarchical structure of a paper with annotations**

This hierarchical structure of a paper is kept in a separate graph in the triplestore and updated each time a new annotation is added. As a result, the implemented architecture is aware of the hierarchical structure of annotations in a publication and one can now claim that user annotations have context. Context in this case is the parent annotation. The above technique allows us to perform context specific semantic queries. With such a structure in place one could now answer the following queries:

- *“Which publications are motivated by dynamic programming languages?”*

  E.g. one could query for all scholarly papers, which contain the annotation http://dbpedia.org/resource/dbpedia:Dynamic_programming_language , its equivalent (by making use of linked data) or its subcategories in the context of a motivation of a scientific publication.

- *“What ontologies have been used the most in the annotation of publications from the computer science field?”*

  By taking advantage of the DBpedia resources that have been marked within the annotation of type “Keyword” and by applying the SKOS concepts, one could determine which publications are likely to belong to the same field, in this case – computer science. One could then query over those publications to see what concepts are used and to which ontologies they belong to. Since different ontologies have been developed for different fields, it can be confusing to learn what ontologies to use during annotation. Such a query could give an overview of popular ontologies that are used within a specific field.

- *“Which ontology concepts could potentially mean the same thing?”*

  Since the prototype is intended to be used as a crowdsource tool, it can easily happen that different users annotate the same text fragment. Such cases can be easily identified and the user can be asked to verify whether the same thing is meant. The value of this becomes easy to understand when two users from different scientific disciplines refer to the same thing with different concepts, the one that is commonly used in their own field. This provides an opportunity to find semantically equivalent concepts across various ontologies and make use of that in the inference rules. This would enable users from different scientific disciplines to better understand the research from other areas and above all, make them accessible to semantic search. There is also an added bonus for ontology developers who can incorporate that information into their ontology.

Whilst the hierarchical approach is easily implemented if the publication is in a hierarchical format itself like XML, this is not the case with PDF files. The architecture of the prototype overcomes this limitation by taking into account the end and start positions of the annotation within the file (see “char” parameter in Figure 12) in finding the best parent match and thus take the semantic querying of PDF files to a new level.

This was done with the help of the SemAnn Discourse Elements Ontology[30], an ontology extended from the Discourse Elements Ontology[31] with some minor adjustments. This ontology describes the major elements of a scientific publication, especially rhetorical elements. The purpose of this ontology is to provide a context reference for the annotation of a document within

[30] https://github.com/AKSW/semann/blob/mergebranch/ontologies/semann.0.2.sdeo.ttl

[31] http://www.essepuntato.it/lode/http://purl.org/spar/deo

the SemAnn project. This is done by preloading the ontology into the prototype at start-up and thereby encouraging the users to use it without additional effort.

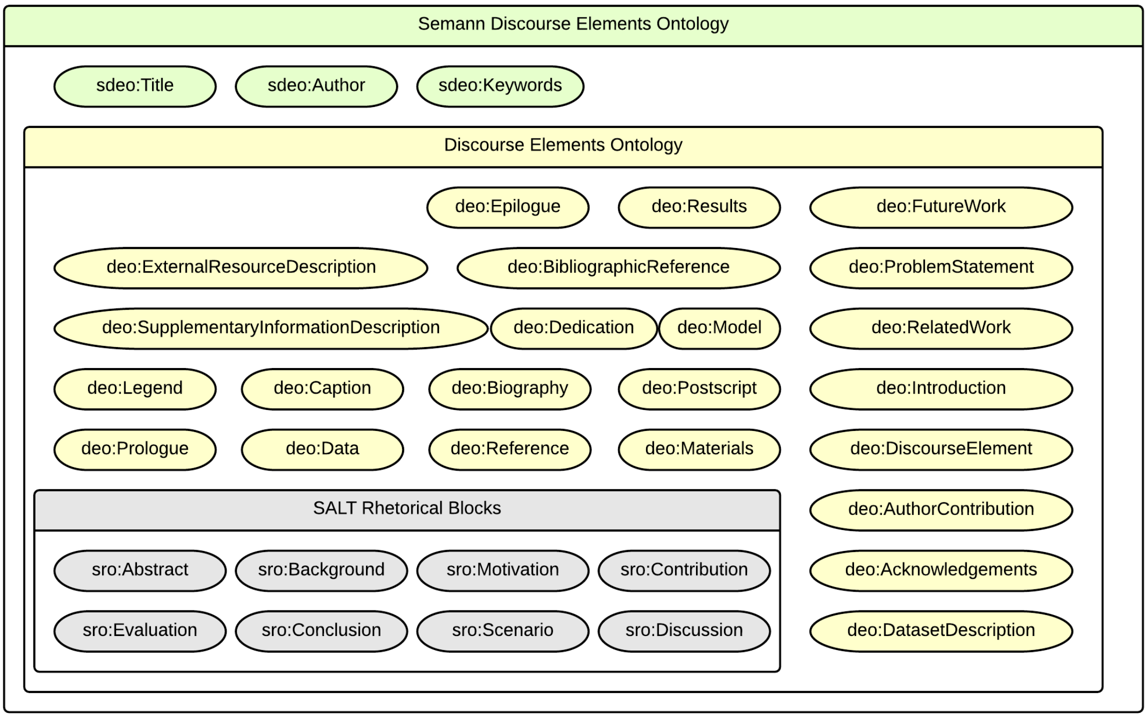


**Figure 18 - SemAnn Discourse Elements Ontology, a scientific discourse ontology**

DEO ontology was favoured due to its good coverage of the main structural elements that might be relevant in the context of semantic search of scientific papers, neither is it too detailed of an ontology to discourage users from using it. Hence, it seemed to be a good choice for the task envisioned. Some minor modifications were applied when extending the DEO ontology:

- Three new classes were added: “Title”, “Author” and “Keywords”. Similar constructs exist in other ontologies like the Dublin Core[32] and FaBiO[33] (also part of SPAR ontologies), but these terms were mostly modelled as properties in other ontologies, not classes. In order to not confuse the end-user too much, it was thought better to keep them as classes like the rest of the elements so they can be easily applied from the user interface, the same way as the rest of them. Dublin Core also has elements like “Agent” that could serve this purpose, however these classes are used very widely and if used elsewhere in the annotation of a document, can lead to a confusion when multiple matches per paper are found by queries. It was considered better to use separate unique classes for the task of identifying such key fields of a paper.
- An rdfs:isDefinedBy property was used to connect each resource to the SDEO ontology namespace so as to simplify the querying over the extended ontology.

Another advantage of the DEO ontology was the presence of the transitive properties “hasPart” and “isPartOf” which one could use within the prototype architecture for reasoning over hierarchical annotations. One could then determine all the parent annotations of an annotation or vice versa - a needed functionality in order to answer question like the following: “what type of annotations are included in the abstract of a paper?”

[32] http://dublincore.org/documents/dcmi-type-vocabulary/index.shtml
[33] http://purl.org/spar/fabio/

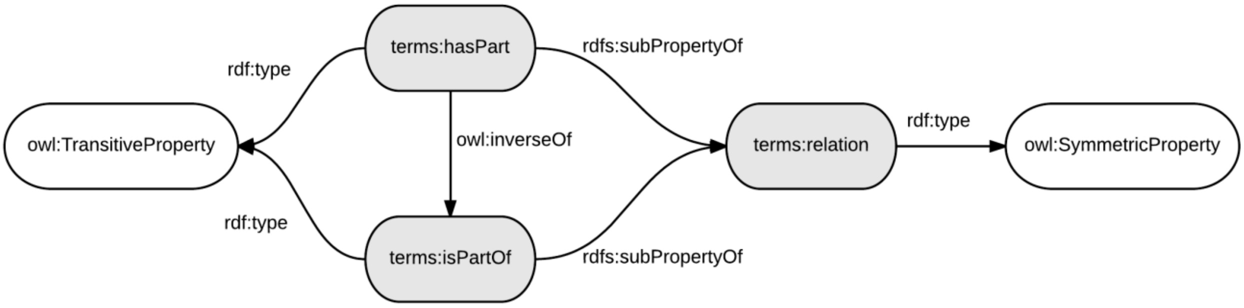
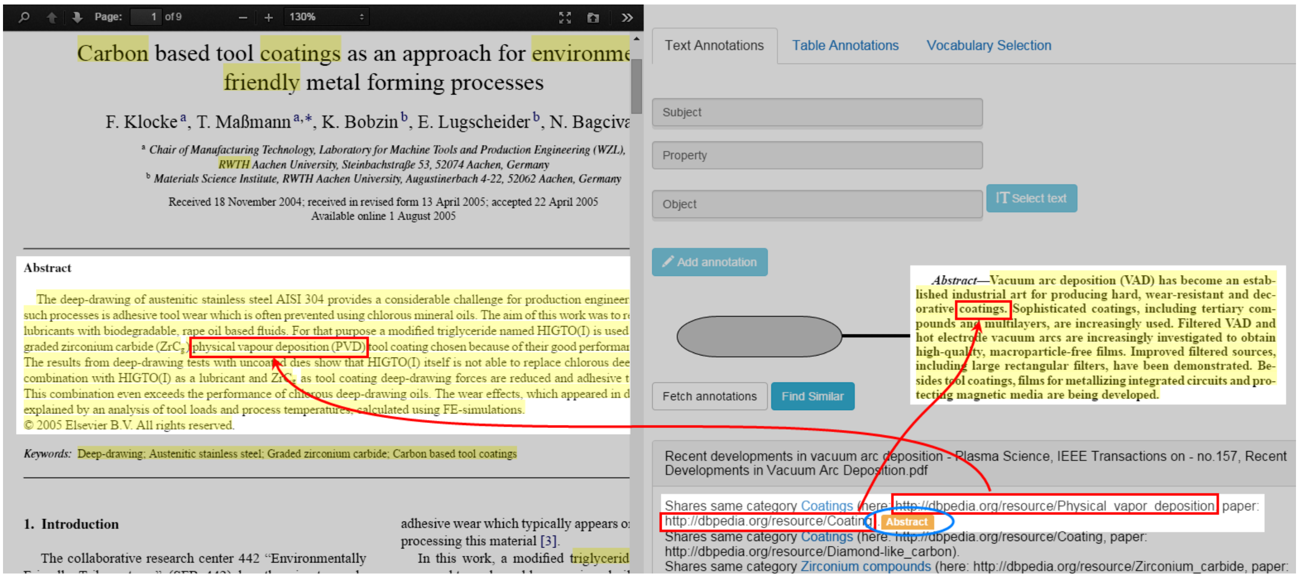


**Figure 19 - Properties of the SemAnn Discourse Elements Ontology**

As mentioned before, the prototype uses SDEO ontology in modelling the annotation contexts with the purpose of improving recommendation results from the similar paper search. An example of such a search result can be seen below where a match is made between papers within the context of an abstract (sro:Abstract) between two existing annotations from the same DBpedia subject category "Coatings" (http://dbpedia.org/resource/Category:Coatings).



Carbon based tool coatings as an approach for environme
friendly metal forming processes

F. Klocke[a], T. Maßmann[a,*], K. Bobzin[b], E. Lugscheider[b], N. Bagciva

[a] Chair of Manufacturing Technology, Laboratory for Machine Tools and Production Engineering (WZL), RWTH Aachen University, Steinbachstraße 53, 52074 Aachen, Germany
[b] Materials Science Institute, RWTH Aachen University, Augustinerbach 4-22, 52062 Aachen, Germany

Received 18 November 2004; received in revised form 13 April 2005; accepted 22 April 2005
Available online 1 August 2005

**Abstract**

The deep-drawing of austenitic stainless steel AISI 304 provides a considerable challenge for production engineer
such processes is adhesive tool wear which is often prevented using chlorous mineral oils. The aim of this work was to r
lubricants with biodegradable, rape oil based fluids. For that purpose a modified triglyceride named HIGTO(I) is used
graded zirconium carbide ($ZrC_g$) physical vapour deposition (PVD) tool coating chosen because of their good performa
The results from deep-drawing tests with uncoated dies show that HIGTO(I) itself is not able to replace chlorous dee
combination with HIGTO(I) as a lubricant and $ZrC_g$ as tool coating deep-drawing forces are reduced and adhesive t
This combination even exceeds the performance of chlorous deep-drawing oils. The wear effects, which appeared in d
explained by an analysis of tool loads and process temperatures, calculated using FE-simulations.
© 2005 Elsevier B.V. All rights reserved.

*Keywords:* Deep-drawing; Austenitic stainless steel; Graded zirconium carbide; Carbon based tool coatings

**1. Introduction**

The collaborative research center 442 "Environmentally

adhesive wear which typically appears o
processing this material [3].

In this work, a modified triglycerid

*Abstract*—Vacuum arc deposition (VAD) has become an established industrial art for producing hard, wear-resistant and decorative coatings. Sophisticated coatings, including tertiary compounds and multilayers, are increasingly used. Filtered VAD and hot electrode vacuum arcs are increasingly investigated to obtain high-quality, macroparticle-free films. Improved filtered sources, including large rectangular filters, have been demonstrated. Besides tool coatings, films for metallizing integrated circuits and protecting magnetic media are being developed.

Recent developments in vacuum arc deposition - Plasma Science, IEEE Transactions on - no.157, Recent Developments in Vacuum Arc Deposition.pdf

Shares same category Coatings (here: http://dbpedia.org/resource/Physical_vapor_deposition, paper: http://dbpedia.org/resource/Coating). Abstract
Shares same category Coatings (here: http://dbpedia.org/resource/Coating, paper: http://dbpedia.org/resource/Diamond-like_carbon).
Shares same category Zirconium compounds (here: http://dbpedia.org/resource/Zirconium_carbide, paper:

**Figure 20 - Example of the reasoning capabilities of the SDEO ontology**

Annotation hierarchies are kept in a separate graph[34] from the annotations themselves. See Appendix A for an overview of graphs and namespaces used in the prototype. The following example hierarchy shows how one annotation is within another:

```
  <rdf:Description rdf:about="http://eis.iai.uni-bonn.de/semann/pdf/example.pdf">
    <hasPart rdf:resource="http://eis.iai.uni-
bonn.de/semann/pdf/example.pdf#page=1?char=74,87&amp;id=0/20/1/1:0,0/20/1/1:13"/>
    <hasPart rdf:resource="http://eis.iai.uni-
bonn.de/semann/pdf/example.pdf#page=1?char=74,132&amp;id=0/20/1/1:0,0/21/1/1:13"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://eis.iai.uni-
bonn.de/semann/pdf/example.pdf#page=1?char=74,132&amp;id=0/20/1/1:0,0/21/1/1:13">
    <hasPart rdf:resource="http://eis.iai.uni-
bonn.de/semann/pdf/example.pdf#page=1?char=74,87&amp;id=0/20/1/1:0,0/20/1/1:13"/>
  </rdf:Description>
```

**Figure 21 - RDF/XML representation of annotation hierarchy**

[34] http://eis.iai.uni-bonn.de/semann/graph/meta

### 4.4.2 Annotations

One of the requirements for this prototype is the ability to annotate PDF documents. This requirement in its most basic implementation was already fulfilled in SemAnn[35] v.1.0, implemented as a lab project and part of preparatory work for this thesis. This allowed users to enter information about the annotations in a triple format but was of little use in reasoning. Namely, the information entered into the triplestore lacked connections to existing vocabularies. The prototype presented in this thesis implemented significant changes to the annotation functionality:

(a) Annotation URI was given a more compact format (see Figure 12)
(b) Ontologies were introduced which defined the format of annotations (see chapter 4.4.1).
(c) Annotations can now be linked to existing vocabularies. This was done by adding support for loading vocabularies into the tool (see Figure 5).
(d) Entity recognition[36] was introduced upon selection of text with suggestions from DBpedia.
(e) Users were given much more flexibility in their annotations. Support for four additional structures were added (see Table 1), whilst the user was previously restricted to only the last two structures (see example in Figure 22)

As a result, the implemented prototype supports the following kind of annotation instances which give very flexible means for formulating information about annotations:

**Table 1 - Annotation instances supported by the tool**

| | Various kinds of annotation instances supported by the tool | S | P | O |
|---|---|---|---|---|
| 1. | As an instance of selected class from user specified ontology | + | | + |
| 2. | As an instance of selected property from user specified ontology | | + | |
| 3. | As an instance of DBpedia resource | + | | + |
| 4. | As an instance of a property from DBpedia's mapping based properties dataset[37] | | + | |
| 5. | As an instance of a new user created class in the *publication* namespace[38] | + | | + |
| 6. | As an instance of a new user created property in the *publication* namespace | | + | |

This was a vast improvement from the original version of SemAnn, which only supported the last two cases from the above table:

```
<rdf:Description rdf:about="http://eis.iai.uni-bonn.de/semann/pdf/example.pdf">
    <semann:hasExcerpt rdf:resource="http://eis.iai.uni-
bonn.de/semann/pdf/example.pdf#page=1?char=2059,2378;length=319,UTF-
8&amp;rangyPage=1&amp;rangyFragment=0/187/1/1:0,0/208/1/1:14" />
</rdf:Description>
<rdf:Description rdf:about="http://eis.iai.uni-
bonn.de/semann/pdf/example.pdf#page=1?char=2059,2378;length=319,UTF-
8&amp;rangyPage=1&amp;rangyFragment=0/187/1/1:0,0/208/1/1:14">
    <rdfs:label xml:lang="en">Dynamic languages such as JavaScript are more difficult to
compile than statically typed ones</rdfs:label>
    <semannp:isA rdf:resource="http://eis.iai.uni-bonn.de/semann/owl#Background" />
</rdf:Description>
<rdf:Description rdf:about="http://eis.iai.uni-bonn.de/semann/owl#Background">
    <rdfs:label xml:lang="en">background</rdfs:label>
```

[35] See version 1.0 of the tool at https://github.com/AKSW/semann/tree/v1.0
[36] Implementation relies on the DBpedia Lookup API: https://github.com/dbpedia/lookup
[37] http://downloads.dbpedia.org/3.9/en/mappingbased_properties_en.nt.bz2
[38] See namespace information in Appendix A

```
</rdf:Description>
<rdf:Description rdf:about="http://eis.iai.uni-bonn.de/semann/property#isA">
    <rdfs:label xml:lang="en">is a</rdfs:label>
</rdf:Description>
```

**Figure 22 - An example of the old SemAnn v.0.1 triple structure in RDF/XML format**

As a result of the requirements in chapter 4.1, the implemented annotation functionality provides end-users with maximum freedom in the type of annotations they want to make. That means support both for simple annotations that are mere classifications and more complex annotation tasks e.g. annotations that are connected via some relation.

### 4.4.3 Recommendation Functionality

The use of the ontologies mentioned in chapter 4.4.1 made it possible to reason over the context in which the annotations appear. This is a novel approach that transformed the annotations of a PDF file from flat constructs to hierarchical constructs.

When developing the architecture for the purpose of semantic annotation of scholarly papers, it is important to keep in mind how it would eventually support the semantic search functionality. After all, this is where the real benefit of annotated scholarly information lies. To help focus on this aspect better, the designed prototype implements a recommendation functionality for similar papers, with respect to the currently open document’s annotations.

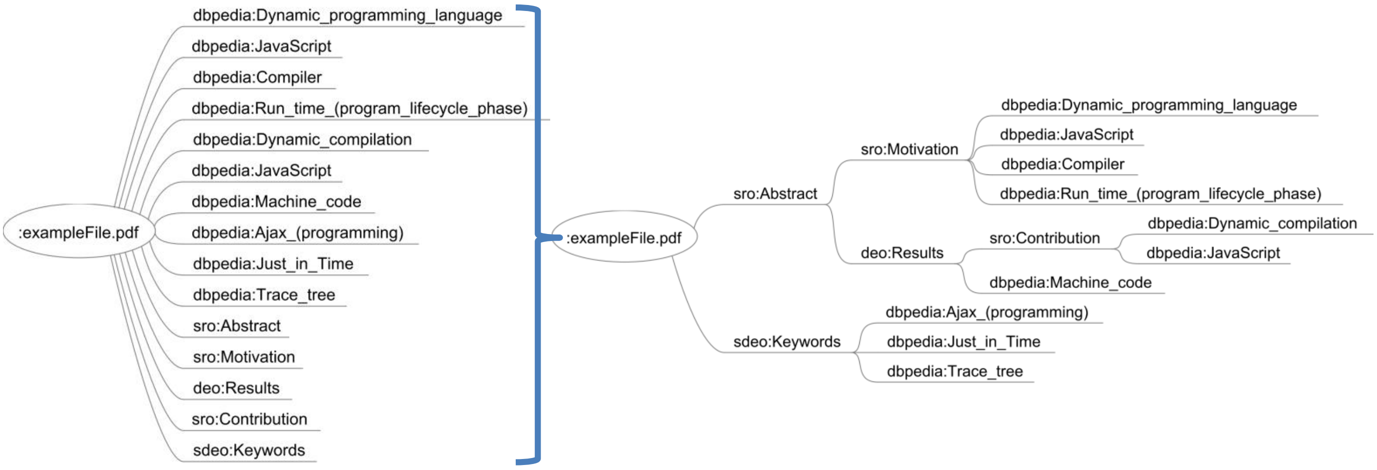


**Figure 23 - Hierarchical annotations**

As a result, more meaningful queries can be performed (see examples in chapter 4.4.1.2). Instead of being limited to querying whether a paper contains an annotation of some type, one can now check whether it appears in the context of an abstract or some other structural element relevant to scientific discourse. The query example in Figure 24 is a compact yet powerful query that a user familiar with SPARQL can easily understand and write. This was achieved by keeping the SemAnn Annotation Ontology as lightweight as possible and it does not require much effort in familiarising oneself with the ontology in order to start writing queries.

```
# return publications that have dbpedia resources in the abstract
prefix semann: <http://eis.iai.uni-bonn.de/semann/0.2/owl#>
prefix : <http://purl.org/dc/terms/>

SELECT ?file ?dbpediaResource
FROM <http://eis.iai.uni-bonn.de/semann/graph/meta> # hierarchy of annotations
FROM <http://eis.iai.uni-bonn.de/semann/graph> # annotation details
WHERE {
    ?file a semann:Publication .
    ?file :hasPart* ?abstract . # transitive property path
    ?abstract a <http://salt.semanticauthoring.org/ontologies/sro#Abstract> .
    ?abstract :hasPart* ?abstractTerm . # transitive property path
    ?abstractTerm a ?dbpediaResource .
    FILTER (STRSTARTS(STR(?dbpediaResource), "http://dbpedia.org"))
}
```

**Figure 24 - Example query for returning all DBpedia resources within abstracts of papers**

The implemented recommender functionality of the prototype (see Appendix B) is currently limited to comparing annotations of type DBpedia resource. Similar papers are found in the following way:

1. Finds all papers where the annotations of the currently open paper match annotations of the same type in other papers.
2. Finds all papers where the currently open paper and some other paper share annotations that point to the same DBpedia subject category.
3. Checks whether any of the found papers have their annotations in the same structural context as the currently open paper.

Recommendations fetched according to the above logic are displayed to the end-user in the format shown in Figure 9.

The ability to query within the specific context of some annotation type (note that one is not limited to structures in SDEO ontology only) is very useful in the context of recommendations. Since research has proven that the most valuable matches are often made based on abstracts, keywords and the title, special weights can be given to semantic matches in that context and display them towards the beginning of the list of similar papers. There are various ways recommendations can be implemented but since the focus of this thesis was not on the specific implementation of its functionality, the current implementation of the recommendation functionality is intentionally kept fairly basic. There is no ordering or filtering of matches, all semantic matches to similar papers are displayed to the user and the details include a list of explanations why the paper is a match. If a match is found in the same structural context, then that is emphasised with a corresponding label next to the specific explanation. The current implementation is therefore limited in its usefulness but served as a good starting point for collecting user feedback for future development and above all served as an excellent reference point for keeping the focus on the usefulness of the triples that were created.

# 5 Evaluation of the Solution

This chapter looks at a few initial hypotheses that were set up and analyses the feedback gathered from a group of ten test users. These users were given a set of tasks to perform (see Appendix C for details) and then interviewed about their experience and asked to provide a rating to reflect that. In addition, an expert in the field of research was interviewed who gave valuable feedback on the various use cases of the tool (current and potential new ones).

The following is the list of problems that this tool solves:

- Ability to semantically annotate text in PDF documents.
  This functionality has been evaluated with the users in the annotation task of the evaluation. A technical limitation of an external library that was used for the prototype has some opportunities of improvement in the user experience when annotating. This is not a deeply fundamental limitation of the architecture or a bug in the implementation of the prototype. It will be discussed in more detail later in the evaluation.
- Support for multiple ontologies.
  This functionality has been implemented in the code and makes the architecture very flexible for usage in any scientific domain and ontology preference. It has not been a focus in the evaluation with the test users, however the implementation of the feature is present in the UI and the general architecture of the tool puts no limitations to its use.
- Support for complex vocabularies with properties.
  The architecture of the tool supports the use of properties on annotations. I.e. it is not a tool that only supports classification ontologies but allows the user to model more complex knowledge as well. This has been implemented in the tool by allowing the user to incorporate an annotation into a relationship with another annotation or an instance of some ontology class. This use case is for expert users and has been evaluated with only selected expert users. Further evaluation and feedback on the user experience of this functionality would be a part of the future work.
- Support for multiple users.
  The general architecture assumes that this is a web-based tool and that data is collected into a central server where it can benefit all users. Hence this architecture supports crowdsourcing of annotations. The specific implementation of the prototype currently runs on a local Virtuoso backend but it would be easy to move this to an online server.

## 5.1 Limitations of the Solution

- The precision of selecting text in the PDF document is limited by the capabilities of the PDF.js library.
- The semantic reasoning of the recommender functionality is currently limited by the fact that DBpedia has a fairly broad general purpose vocabulary that does not necessarily contain very specific terms one might expect to find.

## 5.2 Evaluation Results

The following is an overview of the results from the evaluation of the tool:

- Semantic matches between seemingly unrelated papers were discovered.
- A surprisingly good coverage of related papers was achieved with only 5 annotations per paper when focusing on annotations in the key sections of the paper (e.g. abstract, keywords)

- There is a linear correlation between the number of pre-annotated papers in the database and the number of recommendations of similar papers.
- The average ratings for the evaluated prototype functionalities were mid-range, and some further future work is required to improve the user experience when using the tool.

The background of the users who participated in the evaluation can be found in the user profile descriptions in Appendix D.

### 5.2.1 Hypotheses

The following hypotheses were made for the recommendation task of the evaluation:

1. No unrelated papers will be offered to the user by the recommendation functionality.
2. The recall of related papers returned by the recommendation functionality will be strongly dependent on:
    - the quantity of annotations made by the user
    - the quantity of annotations per paper of already annotated papers in the database

Both hypotheses were proven wrong by the evaluation. At the end of this section there is a graph (Figure 26) that displays how the number of previously annotated papers in the database influenced the number of recommendations of related papers.

Please note that the recall values mentioned in this section refer to the author's own subjective categorisation of the pre-annotated dataset. This approach was favoured due to the users mostly not being familiar with all the papers and the fact that no ordering or filtering of recommendations has yet been implemented for the recommendation functionality. In fact, it was a good opportunity to gather feedback on what format users preferred to get their recommendations as valuable input for future development (see section 5.2.3 for details).

The following set up was used for testing the first hypothesis:

- A test dataset of 10 pre-annotated papers[39] and 1 unannotated paper was prepared. The papers in the pre-annotated dataset were divided into three categories as per author's own assessment of their relation to the unannotated paper: (a) related papers, (b) vaguely related papers, (c) unrelated papers. Of this dataset 40% were assessed to be related, 30% vaguely related and the remaining 30% to be unrelated to the unannotated paper (see Appendix F for details)
- Annotations in the pre-annotated dataset included instances of DBpedia resources and structural annotations from the SDEO ontology. See Appendix F for an overview of annotations used.
- The unannotated paper (Oren, Möller, Scerri, Handschuh, & Sintek, 2006) was chosen to be an easy to understand introduction paper that the test users were asked to annotate[40] in the timespan of 10 minutes. Attention was paid to whether any papers previously categorised as unrelated to the user annotated paper would make an appearance in the list.
- The control group consisted mainly of users who either read research papers daily, had research experience, or were familiar with semantics and RDF triples. See Appendix D for an overview of profiles.

[39] See https://github.com/AKSW/semann/tree/mergebranch/datasets/dataset%201

[40] User annotations available at https://github.com/AKSW/semann/tree/mergebranch/datasets/user%20annotations

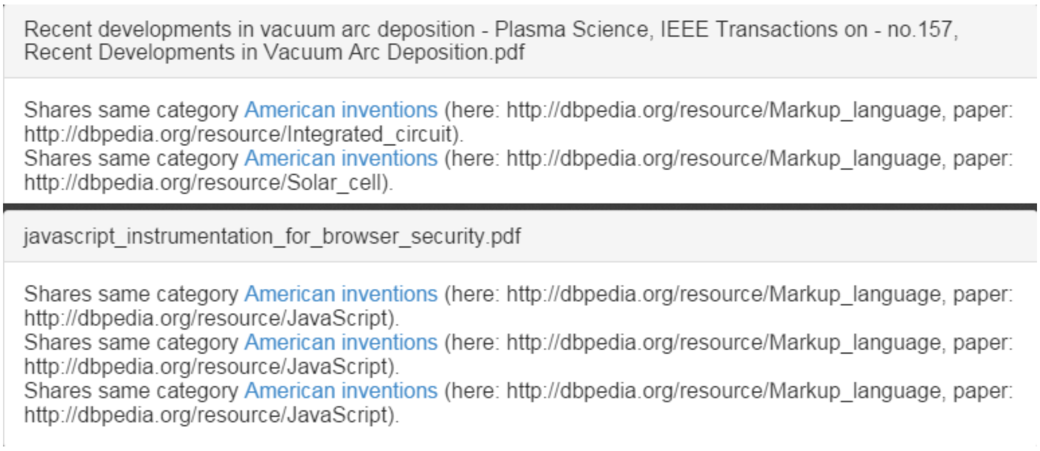
The following setup was used to test the second hypothesis:

- A second test dataset of 30 pre-annotated papers[41] was prepared with the same categorisation ratio as the papers in the initial dataset. Each paper was annotated with 5 unique DBpedia resources, mainly in the abstract of the paper.
- This pre-annotated paper dataset was used by the recommendation functionality and semantically matched against the annotations made by each test user. The number of recommended papers was observed and recall calculated.

The difference between the first and the second datasets is in the quantity of annotations per paper. E.g. the first dataset contains structural annotations and on average 3 times as many annotations of type DBpedia resources. This dataset was used for getting the feedback on the recommendation task of the evaluation. The second dataset was purely designed to observe how the number of previously annotated papers in the database influenced the number of recommendations of related papers.

**Hypothesis 1**

Firstly, a few surprising semantic connections were discovered with papers that the author had previously categorised as unrelated to the paper that the user was viewing. Serendipity played a role here when the same user got two recommendations for papers that were matched according to the very general DBpedia subject category “American Inventions”[42]. As displayed in Figure 25, this connection was made through the user added annotation dbpedia:Markup_Language that was matched against existing annotations in the pre-annotated dataset as Category:American_inventions. This disproved the first hypothesis. Due to the limited size of the dataset it was not possible to prove whether the likelihood of this happening was increased by the presence of above the average number of annotations in these papers, but one could speculate that there might be a correlation.



Recent developments in vacuum arc deposition - Plasma Science, IEEE Transactions on - no.157, Recent Developments in Vacuum Arc Deposition.pdf

Shares same category American inventions (here: http://dbpedia.org/resource/Markup_language, paper: http://dbpedia.org/resource/Integrated_circuit).
Shares same category American inventions (here: http://dbpedia.org/resource/Markup_language, paper: http://dbpedia.org/resource/Solar_cell).

javascript_instrumentation_for_browser_security.pdf

Shares same category American inventions (here: http://dbpedia.org/resource/Markup_language, paper: http://dbpedia.org/resource/JavaScript).
Shares same category American inventions (here: http://dbpedia.org/resource/Markup_language, paper: http://dbpedia.org/resource/JavaScript).
Shares same category American inventions (here: http://dbpedia.org/resource/Markup_language, paper: http://dbpedia.org/resource/JavaScript).

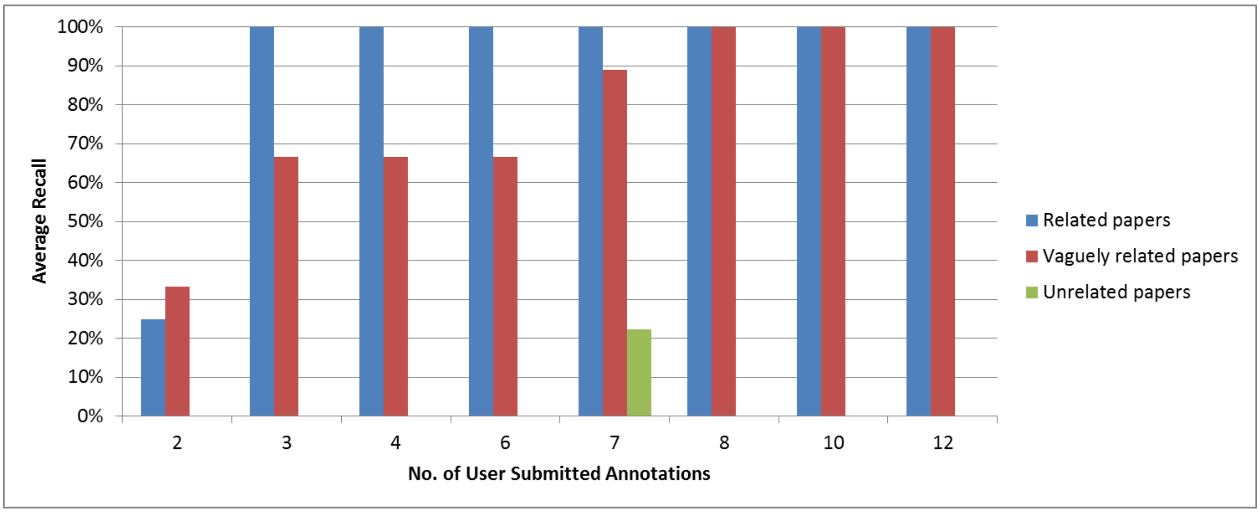
**Figure 25 - Semantic connections between seemingly unrelated papers.**

As mentioned in the previous paragraph, serendipity was the only cause for the “unrelated papers” category to show up in the results of recommendations. Otherwise a 92% recall of related papers was reached on average:

[41] See https://github.com/AKSW/semann/tree/mergebranch/datasets/dataset%202

[42] See http://dbpedia.org/resource/Category:American_inventions

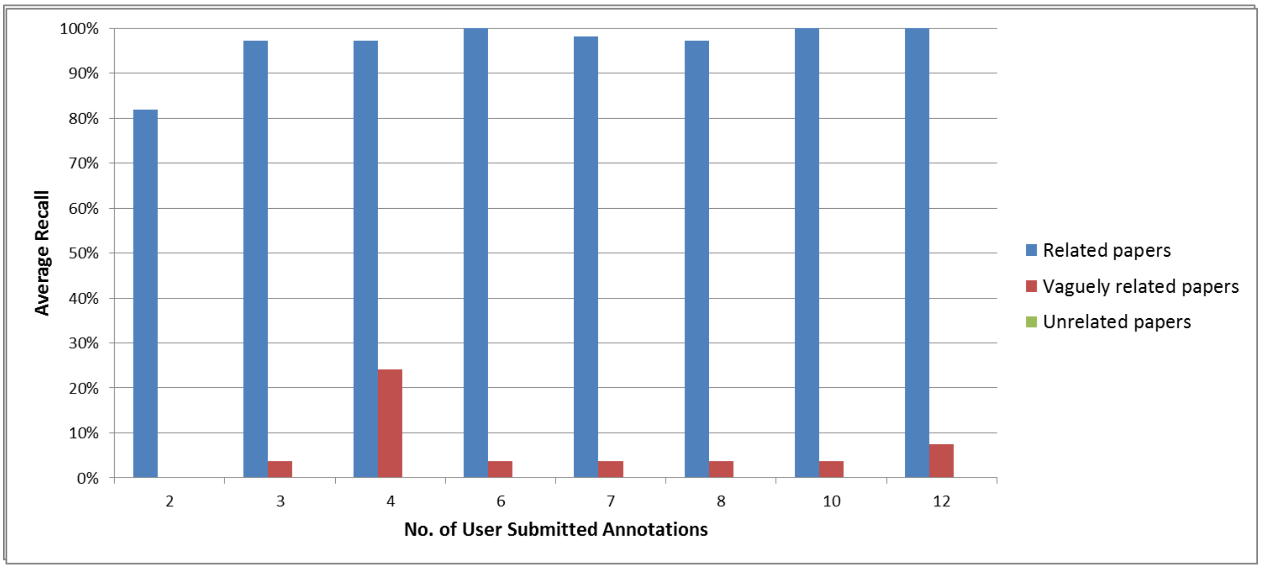
**Table 2 - Recall vs no. of User Submitted Annotations for dataset 1**



## Hypothesis 2

The second dataset as the one with the same amount of unique annotations of type DBpedia resource per paper was used for testing this hypothesis. The average recall of 97% for related papers leads the author to speculate that in order to take advantage of the semantic search one does not necessarily need that many annotations per paper. Only 5 well-chosen annotations from the linked open dataset DBpedia derived from Wikipedia[43] allowed sufficient semantic connections to be discovered at the DBpedia subject category level, as demonstrated with the second dataset. One has to take into account that annotations in that dataset were well chosen in the sense that they were fairly representative of the paper – i.e. annotations in the abstract section, keywords, introduction or the conclusion sections. As demonstrated by (Nascimento, Laender, da Silva, & Gonçalves, 2011) and (Jack, 2012), the abstract contains the most important keywords out of these and is therefore a good source for annotations that would be representative of the paper. Hence, it could be a good idea to encourage the users of this tool to prioritise the annotation of these key sections of a paper.

**Table 3 - Recall vs no. of User Submitted Annotations for dataset 2**



[43] DBpedia dataset v3.9: http://wiki.dbpedia.org/Downloads39

### 5.2.2 Correlation between the Number of Papers and Retrieved Recommendations

For this task, the second dataset was fed to the recommendation functionality in incremental batches of 10 in order to observe how it affected the number of recommended papers each user got after adding their own annotations to the previously unannotated paper. The number of recommendations returned was averaged over all the test users. It is important to note that the same ratio of related (40%), vaguely related (30%) and unrelated papers (30%) was enforced for each batch. A linear correlation was detected ($R^2 = 0.75$) as could have been expected.

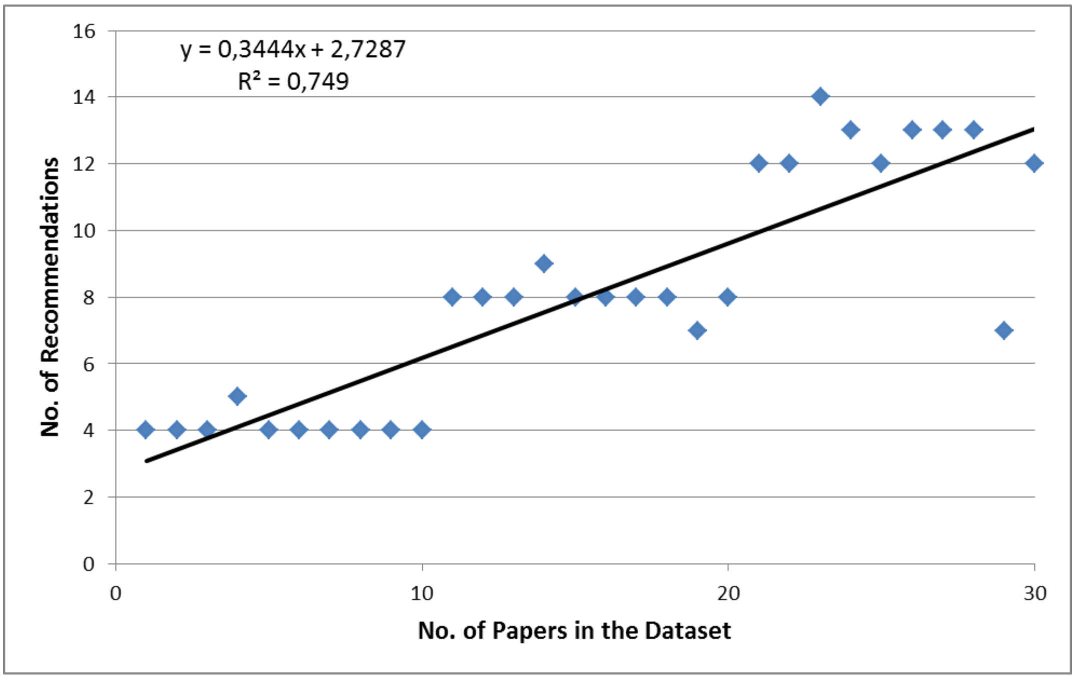


**Figure 26 - Linear correlation between the no. of recommendations and papers in the database**

This information is useful to know for scalability reasons and emphasises the importance of detecting the most relevant recommendations for the user when the number of similar papers grows in time.

### 5.2.3 Feedback

This chapter gives an overview of the feedback gathered from the test users (see Appendix D for interview questions). Feedback that was considered to be relevant by the author has also been registered in the issue tracker for the tool. Most of this feedback concerns the usability of the user interface of the prototype and how to improve it. Users were also asked to provide a numerical score for the two tasks they were assigned with, namely on the annotation user experience and recommender functionality user experience. The averages of these scores are respectively 3.2 and 3.5 out of 5 (as max.).

**Table 4 - Overall ratings for the user experience in evaluation tasks**

| User | Annotations | Recommendations | Research frequency |
|---|---|---|---|
| user 1 | 4 | 2 | Weekly |
| user 2 | 3 | 2 | Weekly |
| user 3 | 3 | 5 | Weekly |
| user 4 | 2 | 4 | Daily |
| user 5 | 3 | 4 | Daily |

| | | | |
|---|---|---|---|
| user 6 | 3,5 | 3,5 | Monthly |
| user 7 | 3 | 2 | Daily |
| user 8 | 3 | 3 | Weekly |
| user 9 | 3 | 4 | Yearly |
| user 10 | 4 | 5 | Yearly |
| **Average** | **3,15** | **3,45** | |

To see feedback per each individual user, please see Appendix H .

### 5.2.3.1 Annotation Task

This chapter outlines the user feedback from the evaluation of their annotation task (see Appendix C). This task got an average score of 3.2 out of 5 (as max.) when users were asked to assess their annotation experience while taking into account all the difficulties they experienced. This dropped to 2.6 out of 5 when rated by users who read research papers daily. The main reasons behind this score have to do with the precision of selecting text in the PDF document or lack of suggested matches from DBpedia.

During this task 97 annotations[44] were entered into the database, of which 65 (67%) were simple annotations of vocabulary terms from DBpedia, 16 (16%) were applications of the SDEO ontology. The remaining 16% were mainly annotations where the user did not specify the annotation to be an instance of any ontology. An overview of this statistics in more detail can be seen in Appendix G.

The below were identified as the main areas for improvement regarding the annotation task:

- **Improve the precision of text selection within PDF**
  50% of evaluation participants remarked that the precision of selecting text within the PDF was not very good – sometimes a letter would be left out when double-clicking on a word to select it or selecting multiple words resulted in the selection of the whole paragraph unless one took extra care to avoid it. This issue has been registered[45] with the proposal to check whether a newer version of PDF.js might solve this in the future.

- **Not all text selections produce a match from DBpedia**
  30% of evaluation participants noted that they would have expected a match from DBpedia when there was none. This was especially true in the case of multiple word selections within PDF, e.g. “semantic web annotations”. This underlined the limitations of using a very broad general purpose vocabulary like DBpedia for classification suggestions and the need to extend entity recognition to other ontologies with narrower focus as well. The above suggestion has been registered[46] with the proposal to investigate potential ways of best achieving this goal.

- **Three input fields is confusing to users**
  Those evaluation participants who were unfamiliar with RDF triples and the semantic web were sometimes puzzled by the presence of three input fields (“subject”, “property”, “object”) when annotating. It was suggested by 30% of evaluation participants to hide

[44] See https://github.com/AKSW/semann/tree/mergebranch/datasets/user%20annotations
[45] Registered as enhancement request: https://github.com/AKSW/semann/issues/21
[46] Registered as enhancement request: https://github.com/AKSW/semann/issues/26

them from the main annotation panel until the user expressed the will to insert more complicated annotations. Since the main annotation use case is very likely to be simple annotations that do not involve relations then the above suggestion has been registered[47] with the proposal to hide the "property" and "object" fields from the main view.

Other suggestions that were made in regard to annotations in order of preference by the author of the thesis:

- *Avoid the use of links for making a selection*[48] (example use case: selecting a DBpedia match from the list of suggestions displayed after selecting text). This is indeed counterintuitive as the user expects it to act as a link and should be considered in further development.
- *Consider using a different colour highlighting for annotations that are instances of the SemAnn Discourse Elements Ontology*[49]. In this way one can identify better whether such annotations contain further child annotations that are instances of other ontologies. Currently such difference is hard to perceive due to the overlap of highlights in the same colour. It would improve the user experience if this were implemented.
- *Add support for deleting annotations from the database*[50]. This would be useful in the case where incorrect annotations have been made that need fixing and should be considered in further development.
- *Reduce the amount of clicks needed for saving a simple annotation*. Currently one has to make a minimum of three clicks (selecting text, selecting DBpedia resource suggestion, "Add annotation" button) and that could be reduced to two if the selection of a DBpedia resource would also insert it into the database. The author of the thesis feels that this might need further consideration as to whether it is worth deviating from the default and expected behaviour in favour of reduced clicks or whether it would end up confusing the users instead.
- *Provide better graphical feedback for simple entity classification. E.g. when marking an annotation to be an instance of class sdeo:Author, display it visually as a triple "<selected text> rdf:type sdeo:Author, semann:Annotation". In this way the experienced users will understand better what is constructed in the background*. In essence the author of the thesis agrees with the above argumentation but also thinks that this might require similar changes to be made when complex annotations are entered. This on the other hand would result in an increased complexity of the triple graph and the use of some drawing library. Hence it might need further consideration whether such changes are worth the potential gain or if a similar result could be achieved in some easier way.
- *Replace input fields in favour of the graphical representation of the triple and allow users to drag selected text to the appropriate triple nodes ("subject" or "object") instead*. The author of the thesis feels that this needs further weighing of advantages and disadvantages as it might complicate things for unexperienced users who are not familiar with RDF triples.

[47] Registered as enhancement request: https://github.com/AKSW/semann/issues/22
[48] Registered as enhancement request: https://github.com/AKSW/semann/issues/25
[49] Registered as enhancement request: https://github.com/AKSW/semann/issues/23
[50] Registered as enhancement request: https://github.com/AKSW/semann/issues/24

- *Remove the right side pane and display PDF in full width of the window*. All the annotations would be performed via a popup window that opens after selecting text. The author of the thesis feels that this is not necessarily a good idea given that there are other use cases for this tool beyond simple annotations. However, this issue could be partly alleviated by allowing the user to decide how much of the screen width would be used for displaying the PDF document[51].

#### 5.2.3.2 Recommender Functionality

The following is user feedback from the evaluation of the similar paper recommendations functionality, triggered by the “Find Similar” button (see Appendix C). The below were identified as the main areas for improvement regarding the functionality:

- **The current format of the recommendation details is difficult to read**
  70% of evaluation participants found it hard to grasp what is meant with the wording “shares same category” and “mentions” in the details of recommendations. Some of the suggested improvements included changing the wording and displaying this information as a list instead (30% of evaluation participants) and removing full URLs from text. 20% of evaluation participants suggested displaying these details visually rather than as text in order to reduce the time spent processing this information. One of the expert users suggested using a browsable mind map format for this purpose in order to represent connections between the papers. Such a mind map could also serve as a starting point for further searches by allowing the user to filter along the connections of interest. 40% of evaluation participants noted that multiple uses of the same vocabulary term within a paper caused repetitions in the recommendation details and suggested it be merged together.

- **Display additional information**
  40% of evaluation participants suggested displaying an abstract of the recommended papers to help the user in understanding how good of a match a recommended paper is likely to be to what they are looking for. 10% of evaluation participants felt that including basic bibliographic data with the recommendations would improve the results. Since none of this information is currently available to the tool unless some user has made respective annotations about the papers, it is the opinion of the author of the thesis that displaying this additional information requires additional analysis as how to best achieve this. One could consider extracting such information automatically[52] when a paper is opened by the user and asking the user to verify the correctness of the extraction results after which they can be uploaded as annotations into the database and then be used for the above purpose. Information like the year of publication (requested by 20% of evaluation participants) might require the use of an external API, as this information is not often contained in the paper.

- **Ordering of recommendations based on the precision of the match**
  50% of evaluation participants commented on the importance of ordering the recommendations and potentially even showing some kind of a score to reflect the precision of the match. An experienced user suggested taking into account the semantic

[51] Registered as enhancement request: https://github.com/AKSW/semann/issues/19

[52] Examples of such tools are listed on CiteSeer’s page: http://csxstatic.ist.psu.edu/about/scholarly-information-extraction

> distance between matches and the length of papers when ordering recommendations. He also commented that experienced users might be interested in the ability to apply a custom filter on the results depending on whether they want search results to be more generic or specific (i.e. whether to check for common supercategories or subcategories). Another experienced user suggested including the ability to select which annotations of the current paper to take into account when looking for similar papers. Other suggestions included giving preference to papers where keywords match**.**

It is evident from the feedback that the current format the recommendations are presented in is considered to be too wordy and future work should include respective changes.

## 5.3 Expert Opinion

As part of the evaluation, an experienced researcher was interviewed for his feedback. In fact, the interviewee Dr. David P. had a unique perspective of the potential of such a tool, given his background as an engineer and researcher. He also suggested a new use case for the tool's application that he saw a lot of potential in.

Dr. David P. worked at the RWTH Aachen University at the Institute for Fluid Power Drives and Controls for 9 years. The last 4 years he was the Head Engineer at the institute. Before that that he worked as a group leader and a Deputy Head Engineer for 2 years and another 3 years as a Research Assistant. Throughout his career at the institute he participated in the authoring of multiple research papers. Even though he has a basic understanding of what semantic search means, he has not previously had any experience with it nor used a similar tool to what was demoed to him.

The interviewee was presented with the various use cases of the tool and a live presentation of the tool in its current maturity. The following is the overview of the feedback that was collected during this interview.

On the whole the interviewee saw a lot of potential in the future of the tool. Taking into account the peculiarities of the engineering field where everyone tries to patent their inventions, he made the following observations. Firstly, he advised against forcing users to create profiles in order to use this tool – if the idea were to present itself. He said that a lot of the research work that a researcher does is something he does not necessarily want others to be aware of, especially in engineering. With that he meant mainly research for commercial purposes as not all research is done with the purpose of publishing it for common domain use. Also, a lot of the academic researchers work at the industry and therefore need to be careful with what they share with others or even avoid asking too specific questions from other experts in order to not reveal too much about their own work. He feared that a user profile of a researcher could present a potential security risk if access to his profile were to be obtained somehow. He felt it would potentially drive away some of the researchers from using such a tool, especially if they are working on sensitive material with the hope of patenting their inventions later. Hence for people like him it would be important to know that any annotations made with this tool could not be traced back to them if they wished so.

The interviewee very much liked the semantic aspect and potential of the tool. He felt this to be a major advantage over other tools that had similar use cases[53]. He emphasised how important it was to be able to find relevant research on a specific subject, because using knowledge that has been published is in the common domain and one cannot infringe any patents by doing so. He

[53] At the time the use case of SemAnn as a reference management system was discussed

was also happy to know that such a tool could easily support ontologies in multiple languages. He saw it as an important aspect and brought an example from the engineering field where a lot of the engineers around the world study German because it allows them to have access to information that is not always available in other languages like English. Also, German terms in the engineering field tend to be more precise and specific than in languages like English, which makes it easier to search specific material. Hence he felt that it is very likely that researchers from the engineering field might prefer to use vocabularies in the German language when annotating and even better if such a vocabulary had translations to other languages.

**New Use Case – Semantic Annotation of Patents**

There was one previously undiscovered use case which Dr. David P. presented for this tool as he felt strongly that this could have a very high potential and suggested to investigate this further. Namely, he felt that a tool like this could be an ideal fit for the semantic annotation of patents which are also in PDF format. He went on to describe the various difficulties in finding the right patents and how much money and time a tool like this could potentially save. He emphasised how important it is to find out as soon as possible whether one's invention is something that has already been patented, before one embarks on a very lengthy and costly process of trying to obtain a patent themselves. Any help that a tool like this could provide by finding relevant patents could potentially be very valuable. E.g., it is often the case that a relevant patent is not found in time and it will be the patent attorney who brings it to one's attention when one is already in the process of claiming their own patent. If lucky, then this is a patent that is not necessarily solving the same issue as the invention and it is then up to the person to describe the patent in a bit more detail as to emphasise that difference. Whether it is a bad case scenario and there is already a patent on the invention or the case of just needing to specify one's own patent in more detail, there are still unnecessary expenses made that could have been avoided, had one been aware of that patent sooner. This links to another peculiarity of patents which is that they are on purpose written in a very unspecific way, which makes them very difficult to understand. Not only does one need to be an expert of the field but one also needs to have some knowledge about patents. As a holder of a patent, it is not desirable to give away the key secret of the invention and the inventor is only bound to describe the details of the invention with the granularity where it becomes clear that it is not infringing other relevant patents. This on the other hand makes it very difficult to search for similar patents to one's own invention, it usually involves several days of searching and lots of reading, even for an experienced person. Currently, the only way of finding similar patents is to find one that describes what the person is looking for and then trace the other patents that were referred in the document and repeat the same process. If there were some less time consuming way of finding relevant patents, it would be very useful. Since patents are written by patent attorneys who adhere to a certain structure, such connections to other patents could even be automatically extracted to speed up the annotation process.

In conclusion, Dr. David P. thinks that if such a tool's use case were to include patents, it could potentially make a big impact. The main aim of such a use case would be to enable one to find similar patents to what they are looking for. Each patent refers to other patents that have a similar approach but don't solve the specific issue that the current patent solves. It is important to find those patents if one is interested in patenting their own invention or trying to decide if it would be worth patenting. Also, if one wants to know the state of the art in the engineering field, it is better to look at the patents rather than research. He stated that in engineering, one doesn't try to publish things but patent it instead, as once they publish it, they cannot patent it any more. He felt that such a tool could be very useful not only for researchers and inventors but potentially even the patent attorneys themselves.

# 6 Conclusions

This work presented the architecture of a semantic annotation tool for scientific publications in the PDF format. Compared to existing solutions available, the implemented prototype offers the following advantages:

- The tool supports the user in the semantic annotation of scholarly publications in the PDF format, a format greatly neglected by existing tools.
- The tool can be used with all ontologies, making it a general purpose semantic annotation tool of scholarly papers that is not limited in its functionality by focusing on some specific application within a specific domain.
- The tool's functionality goes beyond semantic classification capabilities. Various levels of expressivity are supported, including the ability to express relationships between annotations themselves.
- The tool is capable of viewing annotations in the context of scientific discourse (but not limited to it). This provides powerful reasoning capabilities which can answer questions such as "find papers where the problem statement of the paper addresses […]."

## 6.1 Summary of Conclusions

In chapter 3 the objectives of the thesis were stated. The solutions to these challenges have been implemented and screenshots of the respective functionalities can be found in chapter 4.2. More specifically, the following functionality was implemented:

(a) The tool's ability to semantically annotate text in PDF documents has been significantly improved compared to its original capability as shown in chapter 4.4.2.
(b) The tool offers support for multiple ontologies by allowing the user to load the relevant vocabularies into the tool and then control their visibility via a control panel.
(c) The tool supports vocabularies with properties and is not a mere classification tool for selected text.

Additional requirements from chapter 4.1 were also successfully implemented:

(d) The tool implements a recommender functionality of similar papers. Recommendations are displayed with detailed explanations as to why the user is seeing the recommendation and the matching structural context between the papers is emphasised as additional valuable information.
(e) Multiple users are supported by the general architecture of the tool – a web-based tool that stores annotations in a central server[54]. However, this work takes this a step further in chapter 4.4.1.2 as the architecture is capable of knowing the structural context of each annotation without the user needing to specify it and by simply deducing this from already existing annotations. This is an excellent example of how crowdsourcing of information adds additional value.
(f) The tool adheres to the principles of Linked Data in its implementation and increased compliance with it by allowing users to link up their annotations to other ontologies or the linked open dataset DBpedia derived from Wikipedia. Existing ontologies were extended where appropriate.

[54] Setting up a publicly accessable server is still work in progress: https://github.com/AKSW/semann/issues/9

The following valuable discoveries were made in the evaluation phase:

- Serendipity played a role in semantic matches between seemingly unrelated papers.
- A very good coverage (97%) of related papers was achieved with only five annotations per paper.
- There is a linear correlation between the number of pre-annotated papers in the database and the number of recommendations of similar papers.

## 6.2 Future Work

The implemented prototype introduced a lot of new features to the original SemAnn project. The evaluation feedback suggested mostly user interface improvements and tickets have been registered for addressing the more relevant ones in the project's repository.

Besides the improvement suggestions from the evaluation, the author considers the most important future work to be in extending this tool to support communication with the Annotopia[55] Open Annotation Server (Ciccarese, Annotopia: Open Annotation Server, 2014), an open universal hub for storing and publishing of annotations in the Open Annotation ontology. This means that in the true spirit of Linked Data, semantic annotations created with the SemAnn tool can then be used by other tools like the Utopia[56] PDF viewer, once uploaded to the Annotopia server. Likewise, SemAnn tool can take advantage of the data on the Annotopia server, an important contribution in terms of improving the reasoning capabilities of the SemAnn tool. As a result, such integration would considerably increase the visibility of the semantically annotated data produced by the tool, making it available in standard OA format. This would be a considerable step closer to what semantic publishing is about and open up the data to everybody, i.e. not only the scientific community.

The author also considers it important to automate some of the aspects related to annotating through the use of external APIs. E.g. DBpedia Spotlight[57] could be used for automatically annotating mentions of DBpedia resources in the PDF text and Annotopia comes with ready plug-ins for entity recognition that would extend this further to automatic recognition of ontology concepts. The correct selection of ontologies is in itself a non-trivial task and any help the tool can offer in regards to this would be useful, especially for people who are new to ontologies. Such automation techniques combined could considerably improve the annotation experience as observations from the evaluation tasks displayed that manual annotating is a very repetitive task when done for classification purposes. Automation of such tasks would reduce the user's role to verifying whether correct vocabulary terms were applied and save them time.

Future work should also look into further development of the recommendation functionality since very good conditions for meaningful reasoning have been created in the architecture. This includes the ability to reason about annotations in a certain context (e.g. abstract, motivation, etc.) and the ability to serve recommendations with precise explanations as to why the user is recommended a specific paper to read. Evaluation feedback also emphasised the preference of users to consume this information in a visual format, e.g. as a browsable map of connections similar to RelFinder[58] in concept. This is an interesting approach and a high-level overview map of this kind could provide a good starting point for finding relevant papers by selecting connections of interest.

[55] https://github.com/Annotopia

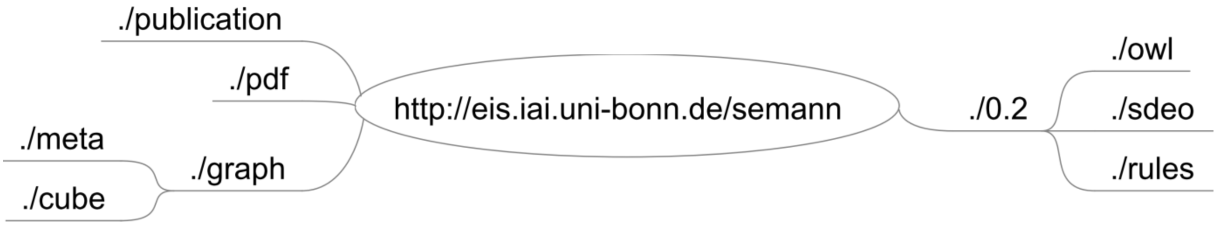
[56] http://utopiadocs.com

[57] http://spotlight.dbpedia.org/

[58] http://www.visualdataweb.org/relfinder.php

# Appendix A Namespaces

The prototype uses the following namespaces:



| Namespace | Use case |
|---|---|
| http://eis.iai.uni-bonn.de/semann | Base URI of the prototype |
| http://eis.iai.uni-bonn.de/semann/0.2 | Refers to 0.2 version of ontologies |
| http://eis.iai.uni-bonn.de/semann/0.2/owl | Annotation ontology |
| http://eis.iai.uni-bonn.de/semann/0.2/sdeo | SemAnn Discourse elements ontology |
| http://eis.iai.uni-bonn.de/semann/0.2/rules | Inference rules |
| http://eis.iai.uni-bonn.de/semann/publication | Annotation instances |
| http://eis.iai.uni-bonn.de/semann/pdf | Base URI of a PDF document. |
| http://eis.iai.uni-bonn.de/semann/graph | Refers to annotation details RDF dataset |
| http://eis.iai.uni-bonn.de/semann/graph/meta | Refers to annotation meta data RDF dataset |
| http://eis.iai.uni-bonn.de/semann/graph/cube | Refers to table annotation RDF dataset[59] |

[59] Not relevant to current thesis, refers to development in https://github.com/saifulnipo/eis-semantic-annotation

# Appendix B Similar Papers

The following is an example query that returns the same papers as the recommendation functionality of the prototype, grouped by the type of semantic match made.

```
# Returns similar papers for publication <http://eis.iai.uni-bonn.de/semann/pdf/example.pdf>
# -      exactMatch = no. of times there was an exact match for a DBpedia resource.
# -      categoryMatch = no. of times there was a match based on DBpedia subject category.
SELECT ?file SUM(?exactMatch) AS ?exactMatch SUM(?categoryMatch) AS ?categoryMatch
{
    {
        SELECT ?file COUNT(*) AS ?exactMatch 0 AS ?categoryMatch
        WHERE
        {
            {
                SELECT DISTINCT ?curr_aType
                FROM <http://eis.iai.uni-bonn.de/semann/graph>
                WHERE
                  {
                    <http://eis.iai.uni-bonn.de/semann/pdf/example.pdf> <http://eis.iai.uni-
bonn.de/semann/0.2/owl#hasAnnotation> ?curr_a .
                    ?curr_a a ?curr_aType .
                  }
                LIMIT 1000
            }
            {
                SELECT DISTINCT ?file ?aType
                FROM <http://eis.iai.uni-bonn.de/semann/graph>
                WHERE
                  {
                    ?file <http://eis.iai.uni-bonn.de/semann/0.2/owl#hasAnnotation> ?a .
                    ?a a ?aType .
                    FILTER (?file != <http://eis.iai.uni-bonn.de/semann/pdf/example.pdf>)
                    FILTER (STRSTARTS(STR(?aType), "http://dbpedia.org"))
                  }
                LIMIT 10000
            }
            FILTER (?curr_aType = ?aType)
        }
    }
    UNION
    {
        SELECT ?file 0 AS ?exactMatch COUNT(*) AS ?categoryMatch
        WHERE
        {
            {
                SELECT DISTINCT ?curr_aType
                FROM <http://eis.iai.uni-bonn.de/semann/graph>
                WHERE
                  {
                    <http://eis.iai.uni-bonn.de/semann/pdf/example.pdf> <http://eis.iai.uni-
bonn.de/semann/0.2/owl#hasAnnotation> ?curr_a .
                    ?curr_a a ?curr_aType .
                    FILTER (STRSTARTS(STR(?curr_aType), "http://dbpedia.org"))
                  }
                LIMIT 1000
            }
            {
                SELECT DISTINCT ?file ?aType
                FROM <http://eis.iai.uni-bonn.de/semann/graph>
                WHERE
                  {
                    ?file <http://eis.iai.uni-bonn.de/semann/0.2/owl#hasAnnotation> ?a .
                    ?a a ?aType .
                    FILTER (?file != <http://eis.iai.uni-bonn.de/semann/pdf/example.pdf>)
                    FILTER (STRSTARTS(STR(?aType), "http://dbpedia.org"))
                  }
                LIMIT 10000
            }
            FILTER (?curr_aType != ?aType)
            GRAPH <http://dbpedia.org> {
                ?curr_aType <http://purl.org/dc/terms/subject> ?curr_category .
                ?aType <http://purl.org/dc/terms/subject> ?curr_category .
            }
        }
    }
}
GROUP BY ?file ORDER BY ?file
```

# Appendix C Evaluation Tasks

The following is an overview of the tasks that the evaluation participants were asked to perform.

**Preparations**

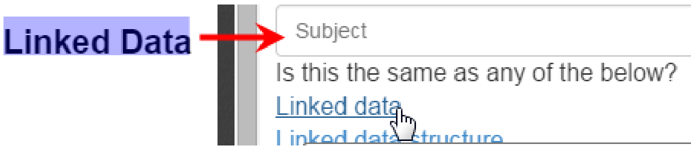
The evaluation participants were given a general introduction to the tool – general goal of the tool, how to make annotations and perform search on similar papers. Each user had their own isolated environment for evaluation so that annotations by other users could not influence the results.

**Tasks performed by the users**

- **Task A – make annotations**

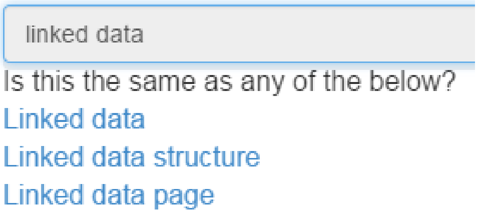
  Read a previously unannotated (Oren, Möller, Scerri, Handschuh, & Sintek, 2006) paper and annotate key points in 10 minutes. You are encouraged to make the following kind of annotations:

  1. Associate keywords with the correct vocabulary term: select a keyword from the PDF file, choose the correct vocabulary term by either selecting from the suggestions (on hover you can see an explanation of the term) or starting your own keyword search. Then press "Add Annotation" button.



Is this the same as any of the below?

**Figure 27 - Selecting a vocabulary term from the suggestions offered**



Is this the same as any of the below?

**Figure 28 - Searching vocabulary term via text input**

  2. Mark appropriate text in the research paper to be research paper structural element where applicable (an overview of the elements in the SDEO ontology is given in introduction).

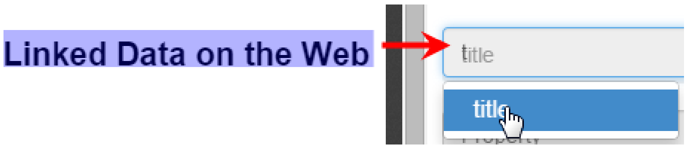
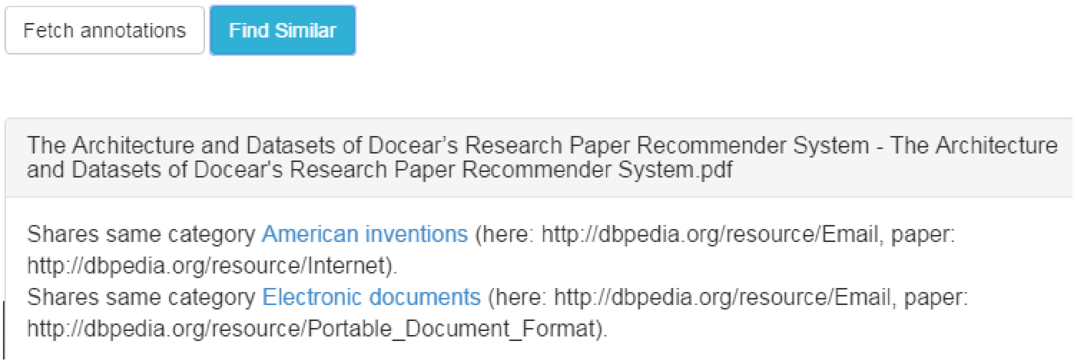


**Figure 29 - Assigning an annotation to be the title of the research paper**

- **Task B – view recommendations**

  Click on the “Find Similar” button to retrieve a list of similar paper recommendations that were compiled based on the annotations you entered. Skim through the list and pay attention to the explanations as to why the paper is recommended to you.



The Architecture and Datasets of Docear’s Research Paper Recommender System - The Architecture and Datasets of Docear's Research Paper Recommender System.pdf

Shares same category American inventions (here: http://dbpedia.org/resource/Email, paper: http://dbpedia.org/resource/Internet).
Shares same category Electronic documents (here: http://dbpedia.org/resource/Email, paper: http://dbpedia.org/resource/Portable_Document_Format).

**Figure 30 - An example of a retrieved recommendation for similar papers**

# Appendix D Evaluation Participants

The following outlines the profile of the participants used in the evaluation.

All the evaluation participants have the following in common:

- They are all heavy computer users, i.e. at least 4h per day.
- They have at least a very basic understanding of what "semantic search" means.
- They have experience with annotations and annotate documents at least every 3-6 months.
- They use annotation mostly in the context of work, school work or personal use.
- They are all between 24-40 years of age.

The main differences in user profiles are due to differences in the field they work in and the degree of involvement in research activity. 3 out of 10 users are considered to be participants with experience due to their daily contact with research.

**User 1**

- Background: This user is a student of Computer Science at Bonn University and has familiarity with the semantic technology field due to her choice of courses. She is also considering writing a thesis in this field.
- Research experience: Currently, she reads research papers weekly due to her thesis.

**User 2**

- Background: This user is a second year student of Computer Science at Bonn University and she has no familiarity with the semantic technology field. Her interests are focused around image processing.
- Research experience: She currently reads research papers weekly as part of her school work.

**User 3**

- Background: This user is a student of Computer Science at Bonn University and has familiarity with the semantic technology field due to her choice of courses. She is also considering writing a thesis in this field and has an interest in programming. She does not usually click on recommendations when browsing and views it mostly as a waste of time, preferring to look for anything she needs herself.
- Research experience: Currently, she reads research papers weekly due to her thesis.

**User 4 – experienced user**

- Background: This user is a student of Computer Science at Bonn University and has very superficial knowledge of the semantic technology field due to her choice of courses (which she did not complete). She is very interested in the field of human-computer interaction and is putting a lot of effort into building a career out of it. She is currently writing a thesis on the subject. She does sometimes follow recommendations when browsing. Specifically when looking for research papers, she does pay attention to any recommendations of similar papers that might be displayed.
- Research experience: Currently, she reads research papers daily due to her thesis.

**User 5 – experienced user**

- Background: This user is a postdoctoral researcher in the field of semantic technologies at Bonn University and is the most experienced user in the group.
- Research experience: He deals with research on a daily basis and has participated in the authoring of several scholarly publications in the semantic field.

**User 6**

- Background: This user is a student of Computer Science at Bonn University and has familiarity with the semantic technology field due to her choice of courses. She has recently started writing a thesis in this field. She does not normally click on recommendations when she is browsing and prefers to search herself rather than follow recommendations.
- Research experience: Currently, she reads research papers monthly due to her thesis.

**User 7 – experienced user**

- Background: This user is a PhD student who has also been working as a research assistant at Bonn University for nearly 2 years. Her PhD is related to the field of semantic technology and also involves scholarly papers. She is considered to be one of the expert users in the group.
- Research experience: She reads research papers daily due to her job and studies. Her research interests include semantic web and scientific communication.

**User 8**

- Background: This user studies Comparative Literature at the University of Düsseldorf. She also has two Bachelor’s degrees - in Germanistics and Journalism. She spends a lot of her time reading secondary literature and is well acquainted with difficulties of finding relevant research in the field of humanities where there is often an overabundance of information.
- Research experience: She reads research papers weekly due to her studies. Her research interests include comparative literature.

**User 9**

- Background: This user works as a R&D employee at a large consumer goods company. He is a mechatronics engineer by profession and through his work often needs to deal with technological innovations on production lines.
- Research experience: He reads research papers yearly, as it is not a regular part of his working life.

**User 10**

- Background: This user works as a R&D employee at a large consumer goods company. He is a mechanical engineer by profession and through his work often needs to lead and implement technological innovations on production lines.
- Research experience: He reads research papers yearly, as it is not a regular part of his working life.

# Appendix E Evaluation Interview

The following questions were asked from the control group during the evaluation.

### General questions

1. Do you have any general comments as feedback about the user interface of the tool?
2. How often do you use a computer?
3. How often do you read research papers with your computer?
4. Please rate how well do you think you know what semantic search means on a 5 point scale with 5 as maximum.
5. How often do you annotate documents?
6. If you annotate, what is the context of annotating?
7. If you do not use annotations, please give a reason why.
8. What is your current profession?
9. What is your age group?

### Task specific questions

| Task | Question | Answer format |
|---|---|---|
| Task A | How would you rate your annotation experience when taking into account all the difficulties you experienced | 5 point scale with 5 as max. |
| Task A | What issues did you experience and do you have any improvement suggestions? | Free form |
| Task B | How would you rate the recommendation functionality of the tool in terms of usefulness and the format it was returned in? | 5 point scale with 5 as max. |
| Task B | What issues did you experience when trying to understand the recommendations and do you have any improvement suggestions? | Free form |
| Task B | Was there any kind of information that you felt was missing and that would help you decide whether the suggested paper is what you would want to open next? | Free form |
| Task B | Were there any surprising or wrong suggestions among the recommended papers? | Free form |

# Appendix F Test Dataset no. 1

The (Oren, Möller, Scerri, Handschuh, & Sintek, 2006) paper that was selected for the user evaluation was selected due to this being a fairly easy introduction paper to the field.

The other 10 annotated papers in the test dataset[60] were selected for the following reasons given:

| ID | Title | Field | Reason for Choice |
|---|---|---|---|
| Paper 1 | Recent Developments in Vacuum Arc Deposition | Mechanical engineering field | This paper is unrelated to the unannotated paper. |
| Paper 2 | Carbon based tool coatings as an approach for environmentally friendly metal forming processes | Mechanical engineering field | This paper is unrelated to the unannotated paper. |
| Paper 3 | The Architecture and Datasets of Docear's Research Paper Recommender System | Computer Science field (software description) | This paper is vaguely related to the unannotated paper and has a good structure for the application of SDEO ontology. |
| Paper 4 | Folksonomy-based information retrieval in context-aware environment | Computer Science field (folksonomy) | This paper is vaguely related to the unannotated paper. |
| Paper 5 | Crowd-sourced Open Courseware Authoring with SlideWiki.org | Computer Science field (crowdsourcing) | This paper is vaguely related to the unannotated paper. |
| Paper 6 | Hybrid Approach for the Semantic Processing of Scientific Papers | Computer Science field (semantics) | This paper is closely related to the unannotated paper. |
| Paper 7 | Publishing on the semantic web | Computer Science field (semantics) | This paper is closely related to the unannotated paper. |
| Paper 8 | OntoWiki | Computer Science field (semantics) | This paper is closely related to the unannotated paper. |
| Paper 9 | Linked Data on the Web | Computer Science field (semantics) | This paper is closely related to the unannotated paper. |
| Paper 10 | JavaScript instrumentation for browser security | Computer Science field (programming language) | This paper is unrelated to the unannotated paper. |

Each paper in this dataset contains annotations. These annotations are of type that the current prototype supports in the inferencing of similar paper recommendations, i.e. they fall into the following categories:

1. Annotations that are instances of DBpedia resources.
2. Annotations that are instances of the SDEO ontology.

The following is an overview of all the annotations contained in the test dataset, organised according to its category.

[60] Available at https://github.com/AKSW/semann/tree/mergebranch/datasets/dataset%201

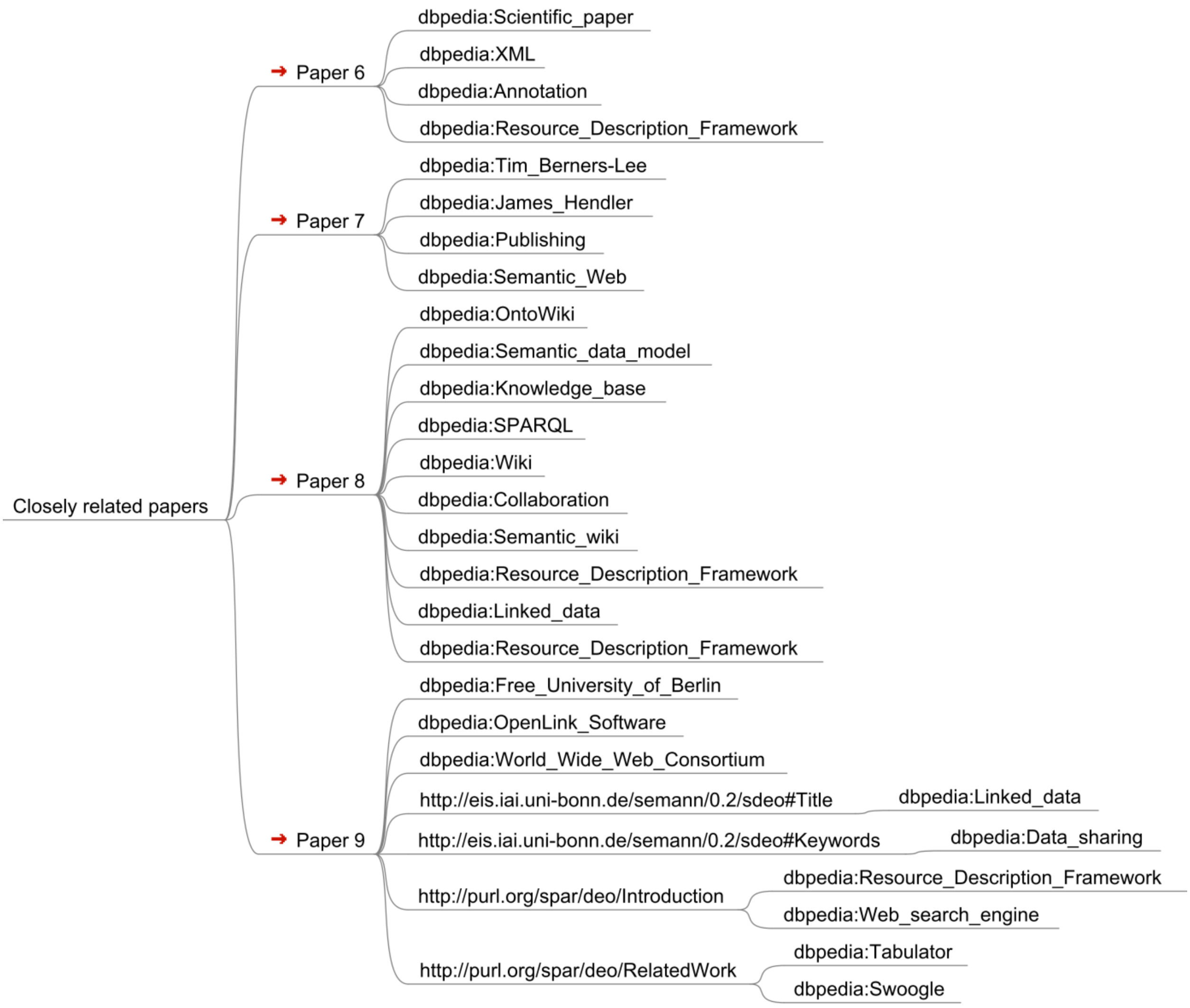

Closely related papers
➜ Paper 6
dbpedia:Scientific_paper
dbpedia:XML
dbpedia:Annotation
dbpedia:Resource_Description_Framework
➜ Paper 7
dbpedia:Tim_Berners-Lee
dbpedia:James_Hendler
dbpedia:Publishing
dbpedia:Semantic_Web
➜ Paper 8
dbpedia:OntoWiki
dbpedia:Semantic_data_model
dbpedia:Knowledge_base
dbpedia:SPARQL
dbpedia:Wiki
dbpedia:Collaboration
dbpedia:Semantic_wiki
dbpedia:Resource_Description_Framework
dbpedia:Linked_data
dbpedia:Resource_Description_Framework
➜ Paper 9
dbpedia:Free_University_of_Berlin
dbpedia:OpenLink_Software
dbpedia:World_Wide_Web_Consortium
http://eis.iai.uni-bonn.de/semann/0.2/sdeo#Title
dbpedia:Linked_data
http://eis.iai.uni-bonn.de/semann/0.2/sdeo#Keywords
dbpedia:Data_sharing
http://purl.org/spar/deo/Introduction
dbpedia:Resource_Description_Framework
dbpedia:Web_search_engine
http://purl.org/spar/deo/RelatedWork
dbpedia:Tabulator
dbpedia:Swoogle

**Figure 31 - Annotations in closely related papers 6-9 of dataset no. 1**

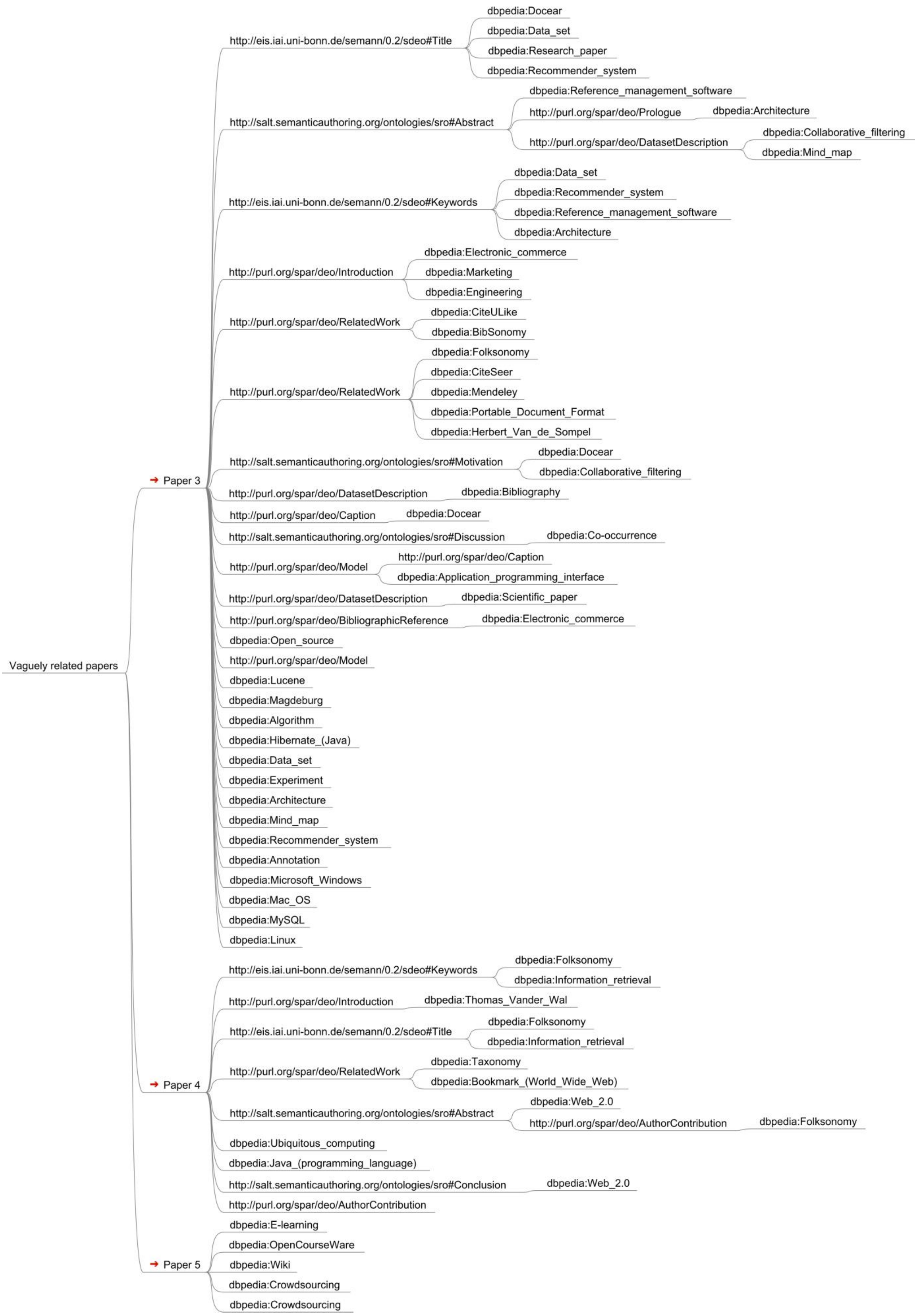

Vaguely related papers
Paper 3
http://eis.iai.uni-bonn.de/semann/0.2/sdeo#Title
dbpedia:Docear
dbpedia:Data_set
dbpedia:Research_paper
dbpedia:Recommender_system
http://salt.semanticauthoring.org/ontologies/sro#Abstract
dbpedia:Reference_management_software
http://purl.org/spar/deo/Prologue
dbpedia:Architecture
http://purl.org/spar/deo/DatasetDescription
dbpedia:Collaborative_filtering
dbpedia:Mind_map
http://eis.iai.uni-bonn.de/semann/0.2/sdeo#Keywords
dbpedia:Data_set
dbpedia:Recommender_system
dbpedia:Reference_management_software
dbpedia:Architecture
http://purl.org/spar/deo/Introduction
dbpedia:Electronic_commerce
dbpedia:Marketing
dbpedia:Engineering
http://purl.org/spar/deo/RelatedWork
dbpedia:CiteULike
dbpedia:BibSonomy
http://purl.org/spar/deo/RelatedWork
dbpedia:Folksonomy
dbpedia:CiteSeer
dbpedia:Mendeley
dbpedia:Portable_Document_Format
dbpedia:Herbert_Van_de_Sompel
http://salt.semanticauthoring.org/ontologies/sro#Motivation
dbpedia:Docear
dbpedia:Collaborative_filtering
http://purl.org/spar/deo/DatasetDescription
dbpedia:Bibliography
http://purl.org/spar/deo/Caption
dbpedia:Docear
http://salt.semanticauthoring.org/ontologies/sro#Discussion
dbpedia:Co-occurrence
http://purl.org/spar/deo/Model
http://purl.org/spar/deo/Caption
dbpedia:Application_programming_interface
http://purl.org/spar/deo/DatasetDescription
dbpedia:Scientific_paper
http://purl.org/spar/deo/BibliographicReference
dbpedia:Electronic_commerce
dbpedia:Open_source
http://purl.org/spar/deo/Model
dbpedia:Lucene
dbpedia:Magdeburg
dbpedia:Algorithm
dbpedia:Hibernate_(Java)
dbpedia:Data_set
dbpedia:Experiment
dbpedia:Architecture
dbpedia:Mind_map
dbpedia:Recommender_system
dbpedia:Annotation
dbpedia:Microsoft_Windows
dbpedia:Mac_OS
dbpedia:MySQL
dbpedia:Linux
Paper 4
http://eis.iai.uni-bonn.de/semann/0.2/sdeo#Keywords
dbpedia:Folksonomy
dbpedia:Information_retrieval
http://purl.org/spar/deo/Introduction
dbpedia:Thomas_Vander_Wal
http://eis.iai.uni-bonn.de/semann/0.2/sdeo#Title
dbpedia:Folksonomy
dbpedia:Information_retrieval
http://purl.org/spar/deo/RelatedWork
dbpedia:Taxonomy
dbpedia:Bookmark_(World_Wide_Web)
http://salt.semanticauthoring.org/ontologies/sro#Abstract
dbpedia:Web_2.0
http://purl.org/spar/deo/AuthorContribution
dbpedia:Folksonomy
dbpedia:Ubiquitous_computing
dbpedia:Java_(programming_language)
http://salt.semanticauthoring.org/ontologies/sro#Conclusion
dbpedia:Web_2.0
http://purl.org/spar/deo/AuthorContribution
Paper 5
dbpedia:E-learning
dbpedia:OpenCourseWare
dbpedia:Wiki
dbpedia:Crowdsourcing
dbpedia:Crowdsourcing

**Figure 32 - Annotations in vaguely related papers 3-5 in dataset no. 1**

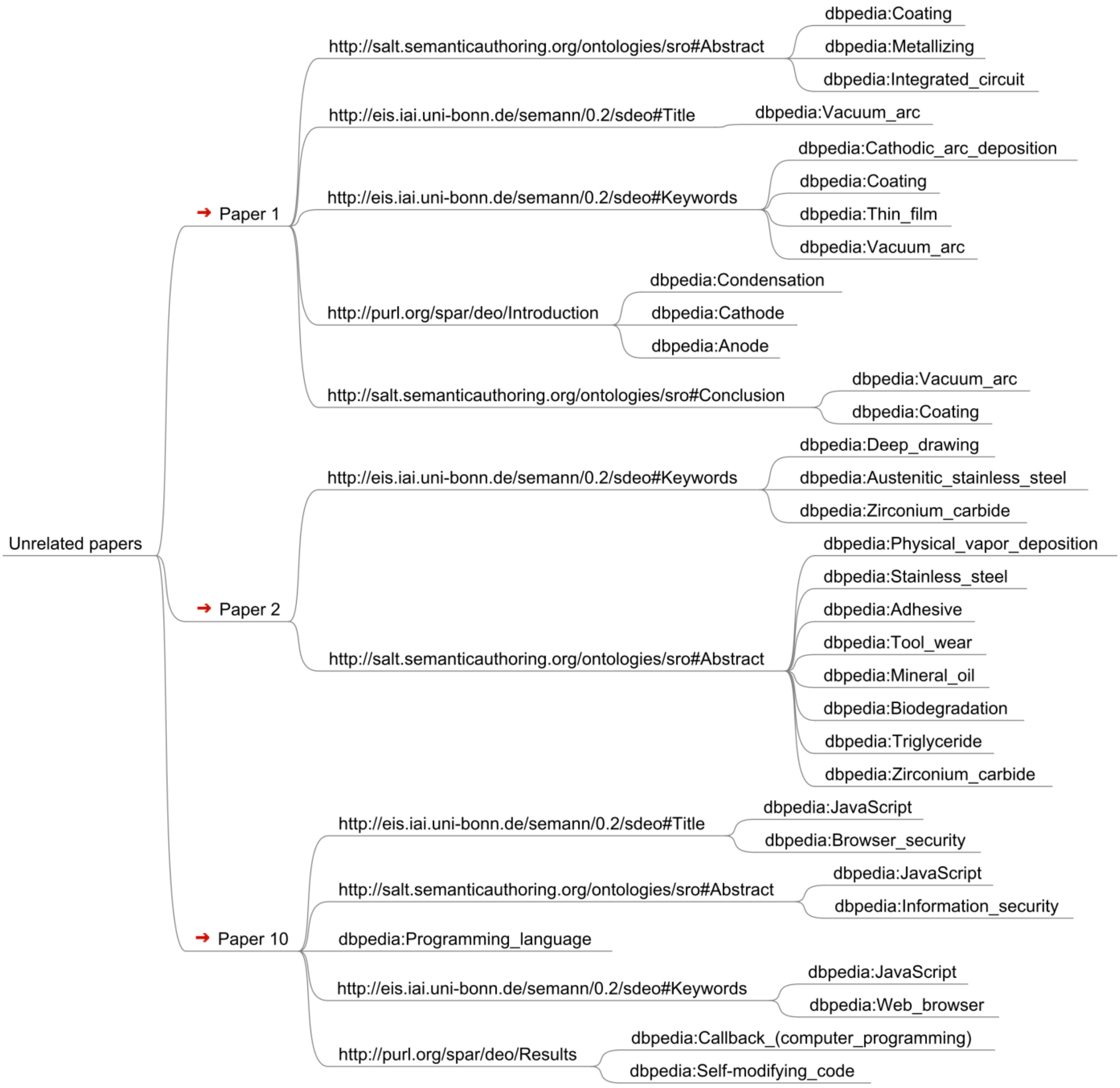

Unrelated papers
➜ Paper 1
http://salt.semanticauthoring.org/ontologies/sro#Abstract
dbpedia:Coating
dbpedia:Metallizing
dbpedia:Integrated_circuit
http://eis.iai.uni-bonn.de/semann/0.2/sdeo#Title
dbpedia:Vacuum_arc
http://eis.iai.uni-bonn.de/semann/0.2/sdeo#Keywords
dbpedia:Cathodic_arc_deposition
dbpedia:Coating
dbpedia:Thin_film
dbpedia:Vacuum_arc
http://purl.org/spar/deo/Introduction
dbpedia:Condensation
dbpedia:Cathode
dbpedia:Anode
http://salt.semanticauthoring.org/ontologies/sro#Conclusion
dbpedia:Vacuum_arc
dbpedia:Coating
➜ Paper 2
http://eis.iai.uni-bonn.de/semann/0.2/sdeo#Keywords
dbpedia:Deep_drawing
dbpedia:Austenitic_stainless_steel
dbpedia:Zirconium_carbide
http://salt.semanticauthoring.org/ontologies/sro#Abstract
dbpedia:Physical_vapor_deposition
dbpedia:Stainless_steel
dbpedia:Adhesive
dbpedia:Tool_wear
dbpedia:Mineral_oil
dbpedia:Biodegradation
dbpedia:Triglyceride
dbpedia:Zirconium_carbide
➜ Paper 10
http://eis.iai.uni-bonn.de/semann/0.2/sdeo#Title
dbpedia:JavaScript
dbpedia:Browser_security
http://salt.semanticauthoring.org/ontologies/sro#Abstract
dbpedia:JavaScript
dbpedia:Information_security
dbpedia:Programming_language
http://eis.iai.uni-bonn.de/semann/0.2/sdeo#Keywords
dbpedia:JavaScript
dbpedia:Web_browser
http://purl.org/spar/deo/Results
dbpedia:Callback_(computer_programming)
dbpedia:Self-modifying_code

**Figure 33 - Annotations in unrelated papers 1, 2, 10 in dataset no. 1**

# Appendix G User Annotation Statistics

**Table 5 - Breakdown of annotations[61] per evaluation users**

| **Annotation statistics** | **User 1** | **User 2** | **User 3** | **User 4** | **User 5** | **User 6** | **User 7** | **User 8** | **User 9** | **User 10** | **Total** | **%** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DBpedia terms | 3 | 7 | 12 | 4 | 7 | 10 | 7 | 6 | 2 | 7 | 65 | 67% |
| SDEO ontology | 0 | 0 | 0 | 1 | 1 | 1 | 3 | 0 | 7 | 3 | 16 | 16% |
| User ontology, i.e. annotations that are not instances of any other ontology | 1 | 6 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 6 | 15 | 15% |
| Triples, i.e. annotations that have relations. | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1% |
| **Total** | **4** | **13** | **14** | **5** | **9** | **11** | **10** | **6** | **9** | **16** | **97** | **100%** |

Since the recommendations of the tool are currently served based on the matches that were made between DBpedia resources that were used in annotations, the below table summarises the use of DBpedia resources per user.

**Table 6 - Summary of the use of DBpedia resources in user annotations**

| **Annotation is instance of** | **User 1** | **User 2** | **User 3** | **User 4** | **User 5** | **User 6** | **User 7** | **User 8** | **User 9** | **User 10** | **Total** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| dbpedia:Resource_Description_Framework | | + | + | + | + | + | | + | | + | 7 |
| dbpedia:Annotation | + | | + | + | | + | + | | + | | 6 |
| dbpedia:Semantic_Web | + | | | | + | + | + | | | + | 5 |
| dbpedia:Tag_(metadata) | | | + | | | + | + | + | | | 4 |
| dbpedia:Blog | | + | | | + | | + | | | | 3 |
| dbpedia:Semantics | | | + | | | + | | | | + | 3 |
| dbpedia:Uniform_resource_identifier | | + | | + | + | | | | | | 3 |
| dbpedia:Wiki | | | | | + | + | + | | | | 3 |
| dbpedia:Metadata | | | + | | | | + | | | | 2 |

[61] User inserted annotations from the evaluation can be viewed here: https://github.com/AKSW/semann/tree/mergebranch/datasets/user%20annotations

| **Annotation is instance of** | **User 1** | **User 2** | **User 3** | **User 4** | **User 5** | **User 6** | **User 7** | **User 8** | **User 9** | **User 10** | **Total** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| dbpedia:Ontology | | | | | | + | + | | | | 2 |
| dbpedia:Ontology_(information_science) | | | | | | | | + | | + | 2 |
| dbpedia:Paris | | | + | | + | | | | | | 2 |
| dbpedia:Semantic_wiki | | | | | | | | + | | + | 2 |
| dbpedia:Sigmund_Freud | | + | + | | | | | | | | 2 |
| dbpedia:Annotea | | | | | | | | | | + | 1 |
| dbpedia:Collaboration | | | + | | | | | | | | 1 |
| dbpedia:Conceptual_model | | | + | | | | | | | | 1 |
| dbpedia:Data_mapping | | | | | | | | | + | | 1 |
| dbpedia:Delicious_(website) | | | + | | | | | | | | 1 |
| dbpedia:Euclidean_space | | + | | | | | | | | | 1 |
| dbpedia:Flickr | | + | | | | | | | | | 1 |
| dbpedia:France | | | | | + | | | | | | 1 |
| dbpedia:Granularity | | | | | | | | + | | | 1 |
| dbpedia:Ireland | | | + | | | | | | | | 1 |
| dbpedia:Linguistics | | | | | | + | | | | | 1 |
| dbpedia:Markup_language | | | | | | | | | | + | 1 |
| dbpedia:Mathematical_model | | | | | | + | | | | | 1 |
| dbpedia:Model_of_computation | | | | | | | | + | | | 1 |
| dbpedia:Named-entity_recognition | | | | | | + | | | | | 1 |
| dbpedia:SAWSDL | + | | | | | | | | | | 1 |
| dbpedia:Technorati | | | + | | | | | | | | 1 |
| dbpedia:Tool | | | | + | | | | | | | 1 |
| dbpedia:WiMAX | | + | | | | | | | | | 1 |
| **Total** | **3** | **7** | **12** | **4** | **7** | **10** | **7** | **6** | **2** | **7** | **65** |

# Appendix H Evaluation Feedback

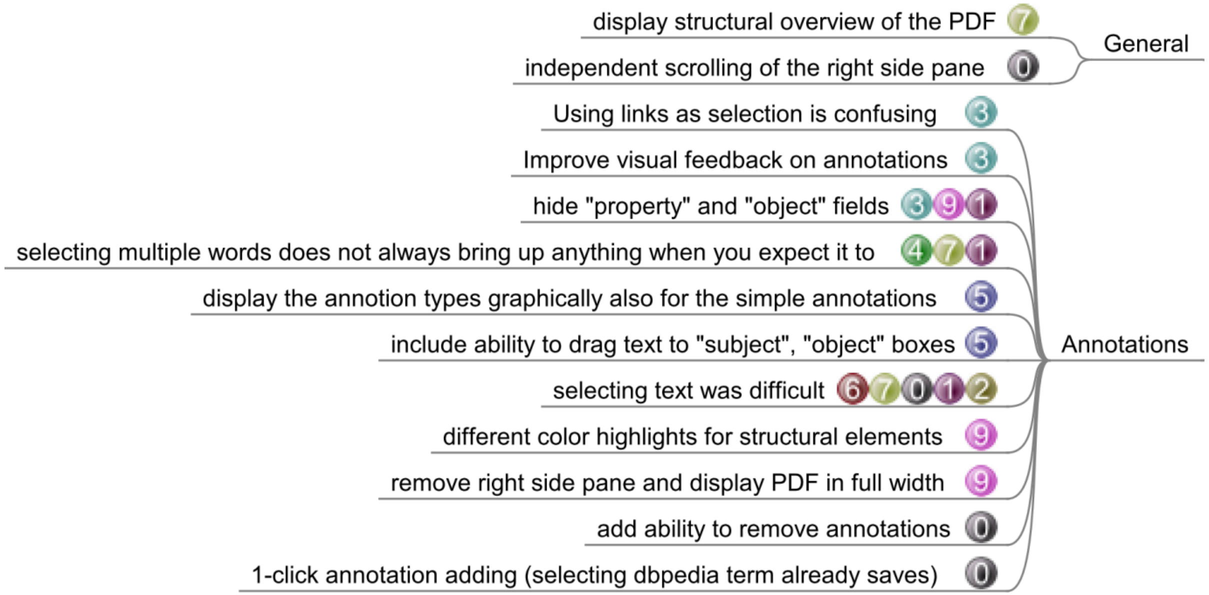

display structural overview of the PDF 7
independent scrolling of the right side pane 0
General
Using links as selection is confusing 3
Improve visual feedback on annotations 3
hide "property" and "object" fields 3 9 1
selecting multiple words does not always bring up anything when you expect it to 4 7 1
display the annotion types graphically also for the simple annotations 5
include ability to drag text to "subject", "object" boxes 5
Annotations
selecting text was difficult 6 7 0 1 2
different color highlights for structural elements 9
remove right side pane and display PDF in full width 9
add ability to remove annotations 0
1-click annotation adding (selecting dbpedia term already saves) 0

**Figure 34 - An overview of user feedback on the annotating task (numbers refer to user IDs)**

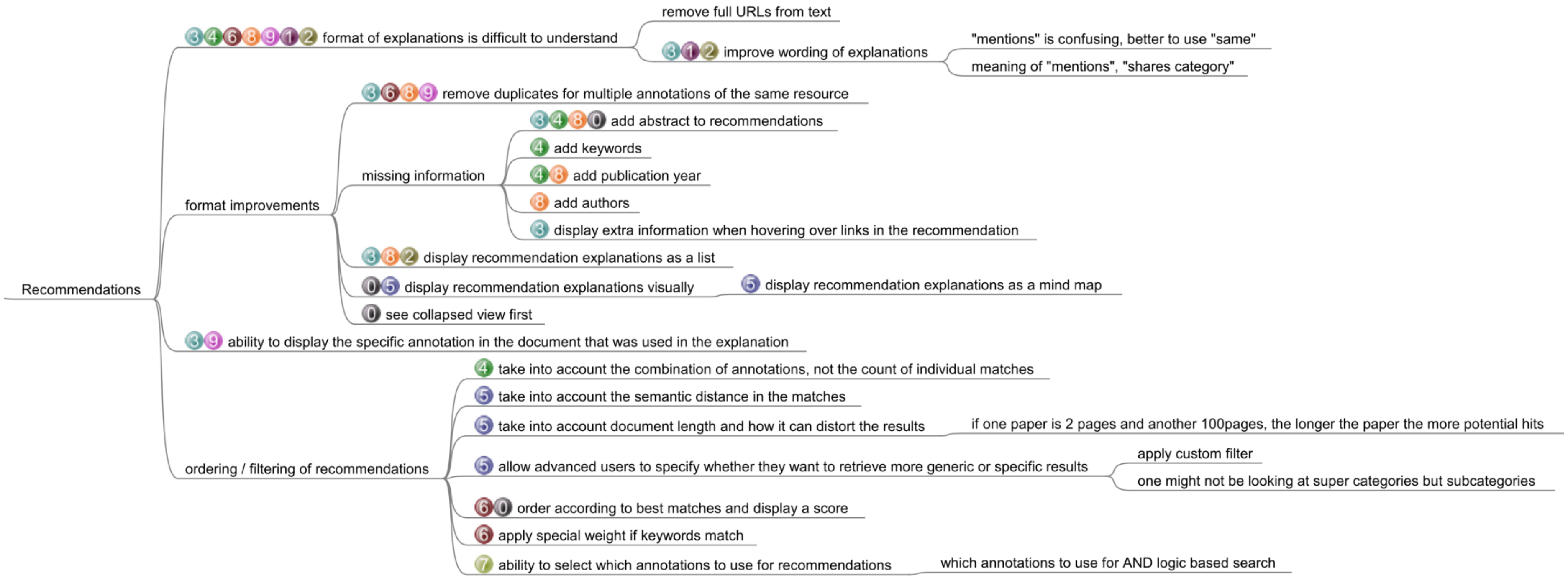

Recommendations
3 4 6 8 9 1 2 format of explanations is difficult to understand
remove full URLs from text
3 1 2 improve wording of explanations
"mentions" is confusing, better to use "same"
meaning of "mentions", "shares category"
format improvements
3 6 8 9 remove duplicates for multiple annotations of the same resource
missing information
3 4 8 0 add abstract to recommendations
4 add keywords
4 8 add publication year
8 add authors
3 display extra information when hovering over links in the recommendation
3 8 2 display recommendation explanations as a list
0 5 display recommendation explanations visually
5 display recommendation explanations as a mind map
0 see collapsed view first
3 9 ability to display the specific annotation in the document that was used in the explanation
ordering / filtering of recommendations
4 take into account the combination of annotations, not the count of individual matches
5 take into account the semantic distance in the matches
5 take into account document length and how it can distort the results
if one paper is 2 pages and another 100pages, the longer the paper the more potential hits
5 allow advanced users to specify whether they want to retrieve more generic or specific results
apply custom filter
one might not be looking at super categories but subcategories
6 0 order according to best matches and display a score
6 apply special weight if keywords match
7 ability to select which annotations to use for recommendations
which annotations to use for AND logic based search

**Figure 35 - User feedback on recommendation functionality (numbers refer to user IDs)**

# 7 Bibliography

Attwood, T., Kell, D., McDermott, P., Marsh, J., Pettifer, S., & Thorne, D. (2009). Calling International Rescue: Knowledge Lost in Literature and Data Landslide. *Biochemical* Journal.

Attwood, T., Kell, D., McDermott, P., Marsh, J., Pettifer, S., & Thorne, D. (2010). Utopia documents: Linking Scholarly Literature With Research Data. *Bioinformatics*.

Bikakis, N., Giannopoulos, G., Dalamagas, T., & Sellis, T. (2010). Integrating Keywords and Semantics on Document. ODBASE.

Ciccarese, P. (2014, 4 18). Annotopia: Open Annotation Server. I Annotate 2014. Retrieved 11 11, 2014, from https://www.youtube.com/watch?v=UGvUbFv0Zl8

Ciccarese, P., Ocana, M., & Clark, T. (2012). Open Semantic Annotation of Scientific Publications Using DOMEO. *Journal of Biomedical Semantics.*

Ciccarese, P., Ocana, M., Castro, L., Das, S., & Clark, T. (2011). An Open Annotation Ontology for Science on Web 3.0. *Journal of Biomedical Semantics*.

Eriksson, H. (2007). An Annotation Tool for Semantic Documents. ESWC.

European Commission. (2013, 12 11). Guidelines on Data Management in Horizon 2020, v.16. Retrieved 11 8, 2014, from http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

Geoghegan-Quinn, M. (2014). *Science 2.0: Europe Can Lead the Next Scientific Transformation.* European Commissioner for Research, Innovation and Science.

Iniesta, A. R., & Corcho, O. (2014). A Review of Ontologies for Describing Scholarly and Scientific Documents. Proceedings of the 4th Workshop on Semantic Publishing.

International Association of Scientific, Technical and Medical Publishers. (2007). Brussels Declaration. Retrieved 11 8, 2014, from http://www.stm-assoc.org/brussels-declaration/

Jack, K. (2012). *Mendeley: Recommendation Systems for Academic Literature.* Presentation at Technical University of Graz.

Liakata, M., Saha, S., Dobnik, S., Batchelor, C., & Rebholz-Schuhmann, D. (2012). Automatic Recognition of Conceptualization Zones in Scientific Articles and Two Life Science Applications. Bioinformatics.

Nascimento, C., Laender, A. H., da Silva, A. S., & Gonçalves, M. A. (2011). A Source Independent Framework for Research Paper Recommendation.

Oren, E., Möller, K., Scerri, S., Handschuh, S., & Sintek, M. (2006). What Are Semantic Annotations?

Osborne, F., & Motta, E. (2012). Mining Semantic Relations between Research Areas. *In Proceedings of ISWC 2012*. Boston, MA.

Osborne, F., & Motta, E. (2014). Understanding Research Dynamics. 11th ESWC.

Pettifer, S., McDermott, P., Marsh, J., Thorne, D., Villeger, A., & Attwood, T. (2011). Ceci n'est pas un hamburger: Modelling and Representing the Scholarly Article. *Learned Publishing*.

Shadbolt, N., Hall, W., & Berners-Lee, T. (2006). The Semantic Web Revisited. *Intelligent Systems*.

Shotton, D. (2009). Semantic Publishing: the Coming Revolution in Scientific Journal Publishing. *Learned Publishing*.