Scientific Knowledge Object Patterns

Fausto Giunchiglia[1], Hao Xu[1, 2], Aliaksandr Birukou[1], Ronald Chenu[1]

1. Department of Information Engineering and Computer Science, University of Trento
Via Sommarive, 5 I-38123 Povo, Trento, Italy
(+39) 0461 282092

2. College of Computer Science and Technology, Jilin University
Qianjin Street, 2699-130012, Changchun, China
(+86) 0431 85168832

{fausto, hao, birukou, ronald}@disi.unitn.it

ABSTRACT

Web technology is revolutionizing the way diverse scientific knowledge is produced and disseminated. In the past few years, a handful of discourse representation models have been proposed for the externalization of the rhetoric and argumentation captured within scientific publications. However, there hasn’t been a unified interoperable pattern that is commonly used in practice by publishers and individual users yet. In this paper, we introduce the Scientific Knowledge Object Patterns (SKO Patterns) towards a general scientific discourse representation model, especially for managing knowledge in emerging social web and semantic web.

Categories and Subject Descriptors

I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods – *representation languages, semantic networks*; I.5.1 [**Pattern Recognition**]: Models – *structural*.

General Terms

Design, Theory.

Keywords

Scientific Knowledge Object, Discourse Representation, SKO Patterns.

1. INTRODUCTION

Emerging web services technology is driving profound changes in the ways of scientific communication in academic societies. Scientific discourses, as the basic unit of dissemination and exploitation of research results, have steadily enhanced their discoverability and reusability in response to the advancement of web 2.0, semantic web, data-driven science, and open source science. A highly semantic enriched publication always makes its information and data much easier to search, navigate, disseminate and reuse, whereas most online articles of today are still electronic facsimiles of linear structured papers with shallow metadata described, lacking of semantic knowledge and interlinked relations among elementary modules of content.

In the last few years, a handful of models have been proposed for scientific discourse representation, which aim to externalize the rhetoric and argumentation within publications [3]. Harmsze’s Model [5] is one of the first comprehensive models for presenting rhetorical structure of scientific information in electronic articles. ABCDE Format [1] organizes papers by five types of rhetorical blocks, i.e. Annotation, Background, Contribution, Discussion, Entities, that is also similar to the IMRD (Introduction, Methods, Results, Discussion) structure [8]. SALT (Semantically Annotated LaTeX) [4] is constituted by three ontologies (Document Ontology, Rhetorical Ontology, Annotation Ontology) and dedicated to an authoring framework targeting enrichment of scientific discourses with metadata. Conceptually, all of these representation models for rhetorical structuring are analogous, whereas the theoretical foundations are different such as the Rhetorical Structure Theory (RST) [6] or Cognitive Coherence Relations [7].

In this paper, we propose the Scientific Knowledge Object Patterns (SKO Patterns) towards a general discourse representation model especially for the knowledge management in the emerging social web and semantic web. Such model not only draws on the essence of the above-mentioned rhetorical structured models, but also extends the capabilities of semantic annotation, semantic search, and strategic authoring grounded on logical reasoning (i.e. Deduction, Induction, and Abduction). Basically, a Scientific Knowledge Object (SKO) [2] is a four-layer scientific knowledge representation model capturing different aspects of scientific artifacts (content, semantics, serial order and presentation). The SKO Patterns mainly work in the semantic and serialization layers to help pattern users establish semantic documentations with flexible rhetorical structures, along with extensionable and interoperable metadata schemes.

Potential users of our proposed patterns include scientific publishers, digital libraries, knowledge base developers, or even individual researchers and authors who want to make scientific publications more modularized, expressive, semantic, and reusable.

2. SKO PATTERNS

By convention, pattern definition is described with the *Context* of use, the *Problem* that the pattern addresses, the *Forces* of scenario, the *Solution* to the problem, the *Rationale* of mechanism, the *Benefits* of the solution that resolves the forces, the *Liabilities* of such solution, and the *Known Uses* from the existing related projects and applications.

2.1 Context

We want to publish a research paper and make it easy to read, search, and reuse by others.

A scientific publication is always written and read in a linear structure as an indivisible knowledge unit. Its complex composition makes readers hard access the target information directly, especially for those non-expert readers. A rhetorical structure unveils precise semantics of the paper under the processes of intuitive thinking. Moreover, metadata as supportive material makes related data and knowledge linked. These would definitely facilitate the reading, dissemination, information retrieval, and semantic search.

2.2 Problem

A traditional paper doesn’t represent its rhetorical structure explicitly and lacks semantic information.

2.3 Forces

- A traditional paper is always a self-contained narrative with linear structure ordered by sections.
- A traditional paper has shallow metadata support for navigation and search.
- In a traditional paper, the conceptual structure is implicitly expressed to readers.
- It is difficult to automatically extract information and meta-information from a traditional paper.
- It is difficult to import/ export/ integrate annotations of a paper from other researchers.
- In traditional papers, text is not linked to the underlying data.
- Different audiences have different interesting parts in a paper, and it’s hard to access them directly in a traditional paper.

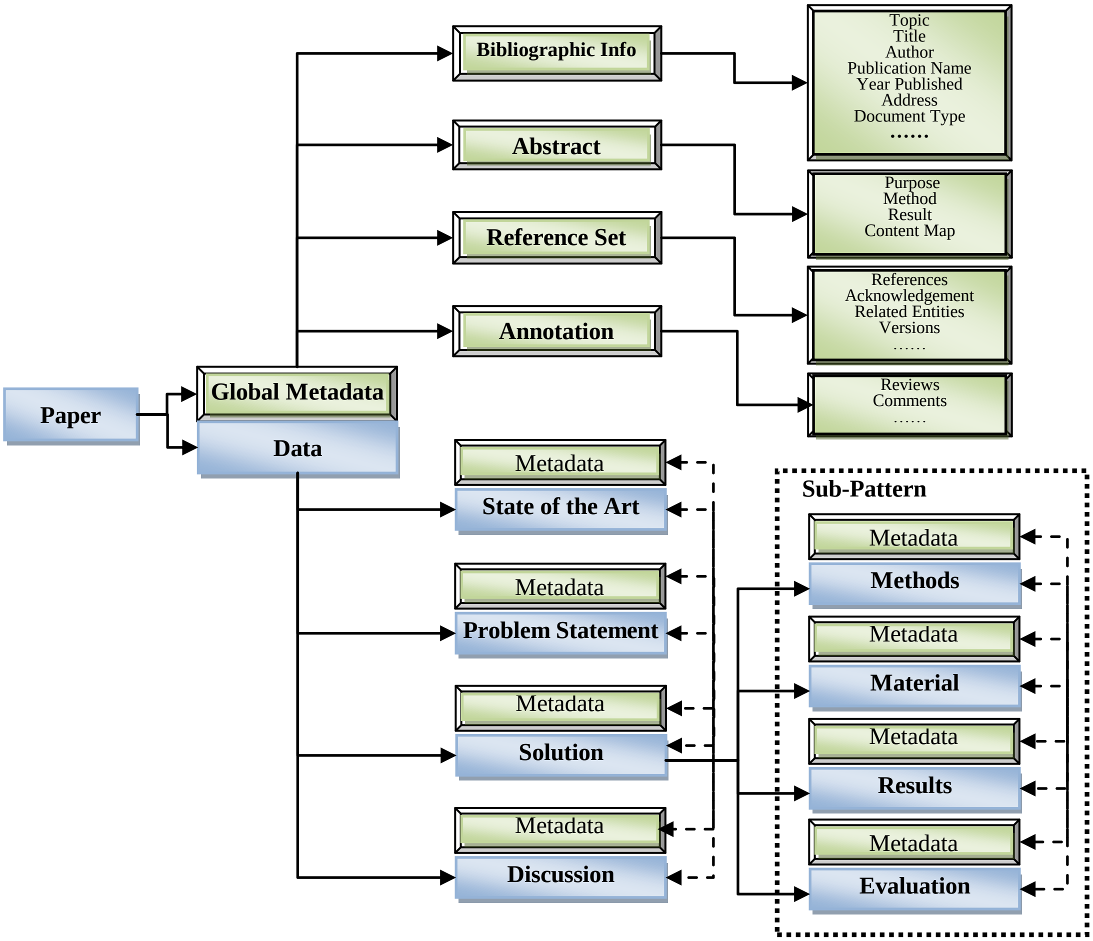

Figure 1. SKO Patterns.

- Low capabilities of social dissemination and collaboration, e.g. tagging, commenting, annotating, and sharing.

2.4 Solution

Compose a SKO paper with rhetorical structure and semantic metadata.

We modularize a scientific paper by logical functions of the information and reorganize it by rhetorical structure as our pattern solution for discourse representation. Above all, we divide a discourse into Metadata and Data parts. Herein, the Metadata consists of *bibliographic information*, *abstract*, *reference set*, *annotation*, etc., while the Data is the main body of a paper that is constructed via the general scientific method. The basic element of rhetorical structure is called Rhetorical Block in our methodology. Figure 1 gives an overview of the SKO Patterns for scientific papers.

Metadata

- Bibliographical Information: Topic, Title, Author/Editor (Name, Affiliation, Email), Keywords, Category, Source (Journal, Conference, Inproceedings, Inbook, Article, Thesis, Techreport, Misc, Other), Publisher, Year, Volume, Number, Pages, Series, Edition, Month, Document Type, etc.
- Abstract: a brief description of paper including Purpose, Method, Result and Content Map.
- Reference Set: A set of referenced entities, such as a list of “References”, Persons and Projects mentioned in “Related Work” and “Acknowledgement”, a set of URLs or other entities in the Footnotes and Endnotes, etc.
- Annotation: Comment, Review, Tag, etc.

Data

- State of the Art: Observations of phenomena, situation, foundational theories and related work, where the contextualized scientific problem addressed.
- Problem Statement: The description and an active challenge faced by researchers and aimed to be solved in the discourse.
- Methods: The specific techniques or methodology used in conducting a particular experiment.
- Material: Data collection, pretreatment, and analysis.
- Results: The outcome or the findings of the research.
- Evaluation: The evaluation methodology and its associated results.
- Discussion: Comparison of the results with related solutions or observations.

SKO Patterns provide a semantic approach for scientific discourse representation. Rhetorical blocks constitute the composition of metadata and data of discourse. Essentially, these rhetorical blocks are unordered, while they always have types of relations between each other instead of linear order. Examples of such relations include explanation relations, argumentation relations, etc. It is impossible to convince researchers follow a uniform serialization for writing various types of publications. However, there always are some sequential relations among the rhetorical blocks. For instance, we commonly address the problem first and find the solution then, as a problem-solving scientific method. During the solution, we need to collect data, carry out the experiment and find out the results. The further sequential relations (orders) of rhetorical blocks, which are based on three strategies of logical reasoning, will be discussed in the following Rationale subsection.

2.5 Rationale

The Rhetorical Blocks are derived from general scientific methods and three fundamental logical reasoning methods (Deduction, Induction and Abduction).

The SKO Patterns are constituted by unordered rhetorical blocks with links through semantic metadata and relations. In this subsection, we sequentially discuss the rationale and some possible solutions for ordering these atomic rhetorical blocks in an intuitive way for both writing and reading.

We derive three fundamental patterns for scientific discourse’s serialization from the three basic types of logical reasoning method, i.e. Deduction, Induction and Abduction. A logical reasoning contains three elements for inferences, that is, Precondition, Rule, and Conclusion.

$$\text{Precondition} \xrightarrow{\text{Rule}} \text{Conclusion}$$

- Deduction is a process of applying the Rule to Precondition and determining the Conclusion. For example, "When it rains, the road gets wet.” is the Rule. “It rains.” is the Precondition. Then we can deduct the Conclusion “The road is wet." Mathematicians are commonly associated with this style of reasoning.
- Induction is using Precondition and Conclusion to find the Rule that can explain the transition. For example, "The road has been wet every time it has rained. Therefore, when it rains, the road gets wet." Scientists are commonly associated with this style of reasoning.
- Abduction is using the Rule and the Conclusion to support that the Precondition could explain the Conclusion. For example, "When it rains, the rood gets wet. The road is wet, therefore, it may have rained." Diagnosticians and detectives are commonly associated with this style of reasoning.

In practice, when we do research and write a paper, problems always have to be solved by steps (states). We take a deduction as an instance:

We start from State 0 (S_0) as the Precondition and Theory 0 (T_0) as the Rule. Using T_1 and S_0 we may deduct S_1 as the intermediate Conclusion, while the rest may be deduced by analogy. So we can reach the State Final (S_F) as the Conclusion.

$$T_0, S_0 \xrightarrow{T_1,\, S_0} S_1 \xrightarrow{T_2,\, S_1} S_2 \dots\dots \xrightarrow{T_i,\, S_{i-1}} S_i \dots\dots \xrightarrow{T_F,\, S_{F-1}} S_F$$

During these reasoning periods, we also need to make the Observation, Hypothesis, and Experimentation for obtaining and validating the related States and Theories. In the following subsections, we propose three rhetorical structure patterns according to the three logical reasoning methods.

2.5.1 Deduction

Deductive Method (Figure. 2) works from general rule or principle to specific solution. (1) Theory and Observation - it begins with a theory and observation of our interests. (2) Hypothesis - then we narrow down it to a specific hypothesis that may solve the problem we face. (3) Experimentation - we narrow down further to test the hypotheses by specific experimentation. (4) Conclusion - a conclusion follows logically from available theory and observations.

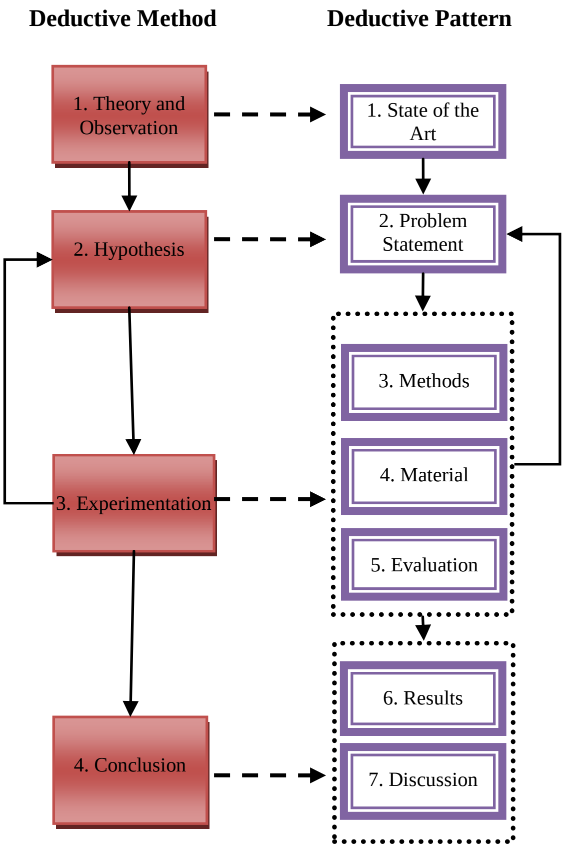

Figure 2. Deduction.

Deductive Pattern

1. State of the Art: Observe S_0, T_0, set i = 1;

Investigate existing Theories and Observations. Related phenomena, development and analysis construct the Initial State (S_0). Selected theories and techniques will support inference and argumentation as T_0.

2. Problem Statement: Hypothesis S_F, State the Problem $P = |S_F| - |S_{i-1}|$;

Predict a Target State S_F as a hypothesis for further test and confirmation. The problem statement presents the gap between S_F and S_{i-1}.

3. Methods: Propose Ti such that $|T_i| > |T_{i-1}|$;

The way of design/ refine/ apply a Theory Ti, which leads $S_{i-1} \longrightarrow S_i$. The method could be experimental method, numerical method, or theoretical method, etc.

4. Material: Compute $S_i = T_i (S_{i-1})$;

Material includes all the raw data, intermediary data, and pretreated data collected from the State of the Art that are used for Experimentation by proposed Method.

5. Evaluation: Evaluate S_i. if ($|S_F| - |S_i| > \varepsilon$) i = i + 1, goto (2) ;

Compare S_i with S_F. If S_i is not satisfied with expectation, repeat the loop 3-4-5 with the modifications of Theories until the ideal S_i is obtained. Here some new problem may arise during the whole loop 3-4-5. If this happens, go to 2 making a new sub problem statement and continue in recursion. When S_i (approximately) equals to S_F, then break and go down to next step 6.

6. Results: $S_F = S_i$;

Present Final State S_F.

7. Discussion: Discuss S_F and $|S_F| - |S_0|$;

Compare S_F and S_0 with related observations and findings from other scientists, always along with an old theory confirmed or applied within a new context.

2.5.2 Induction

Inductive Method works from specific observations to general theories and principles. (1) Observation - we begin with specific observations. (2) Hypothesis - then we formulate a generalized hypothesis to explore. (3) Experimentation - detect the patterns and regularities via various measures and experimentations. (4) Theory - it ends up developing some general theories.

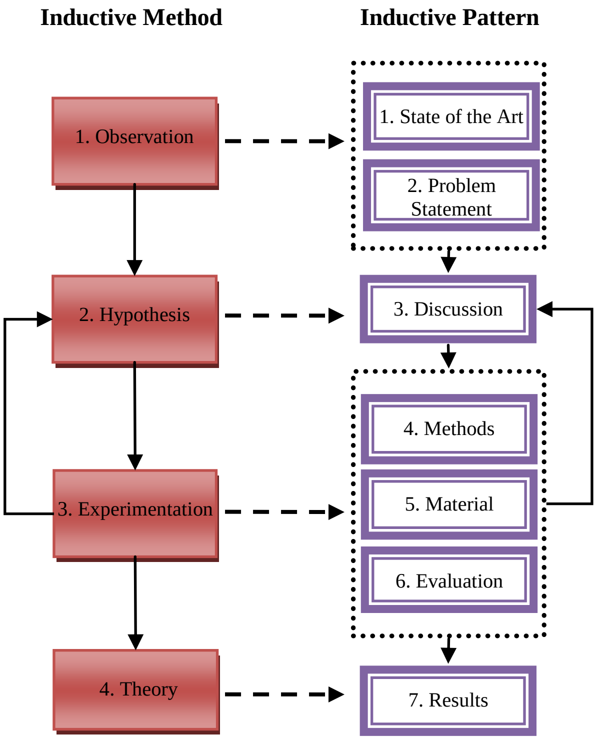

Figure 3. Induction.

Inductive Pattern

1. State of the Art: Observe T_0, S_0, S_F, $i = 1$;

Investigate existing Observations along with their theoretical explanations, and set them as T_0, S_0, S_F.

2. Problem Statement: Hypothesis T_F, $P = |T_F| - |T_0|$;

Pose some phenomena as a Final State S_F, which can't be explained by existing theories or described by existing models. The problem statement aims at finding a Theory T_F, where it possible implies $S_0 \longrightarrow S_F$.

3. Discussion: Discuss Property(S_F) and $|S_F| - |S_{i-1}|$;

Observe and analyze the specific phenomena and particular scenario in S_{i-1} and S_F. Generalize and patternize more general solution for a series of separated problems.

4. Methods: Propose T_i such that $|T_i| > |T_{i-1}|$;

The scientific methodology, logic, or philosophic approach for deriving Theory from transmission $S_{i-1} \longrightarrow S_i$.

5. Material: Compute $S_i = T_i (S_{i-1})$;

Evidences, intermediate data, observations, etc, which support analysis and evaluation via proposed Method.

6. Evaluation: Evaluate S_i. if (S_i != S_F) i = i +1, go to (3);

Compare S_i with S_t. Repeat the loop 3-4-5-6 with the modifications of Ti until the ideal Theory is obtained.

7. Results: $T_F = T_i$;

A new theory T_F is proposed.

2.5.3 Abduction

Abductive Method is the process of inference that produces a hypothesis as its end result. (1) Observation - observe a set of seemingly unrelated facts, armed with an intuition that they are somehow connected. (2) Theory - we then move to the related theories or principles that may explain some features of facts. (3) Experimentation - infer a possible precondition as an explanation of observable facts judging by existing theories. (4)Hypothesis - a hypothesis is detected.

Abductive Pattern

1. Problem Statement: Pose a problem that to derive explanations E of observations O according to theories T, namely

(1) $T \cup E \vDash O$ and

(2) $T \cup E$ is consistent.

2. State of the Art: Investigate related observations, phenomena, facts, and set them as the Final State S_F.

3. Discussion: Observe and analyze the set of seemingly unrelated facts, and discuss various possibilities that an Initial State S_i could be an explanation of S_F, where

$S_i \longrightarrow S_F$.

4. Methods: The way of deriving S_i, for example, enumerative method, exclusive method, etc.

5. Material: Evidences, facts, observations, etc, which support analysis and backtracking according to existing Rule.

6. Evaluation: Compare $T(S_i)$ with S_t. Repeat the loop 2-3-4-5-6 with the modifications of methods and replacement of rules until the ideal S_i is obtained.

7. Results: Phenomena detection or theory generation/ development/ appraisal.

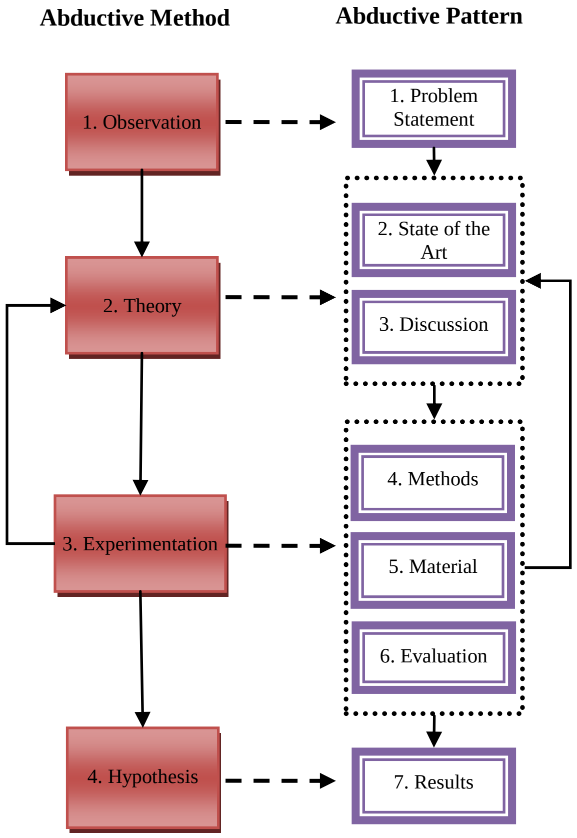

Figure 4. Abduction.

2.6 Benefits

- Rhetorical structured papers facilitate strategic reading.
- Rhetorical blocks enhance the discoverability of elementary knowledge within the context.
- Metadata and other annotated semantic information enable linking scholarly literature with research data.

- SKO Patterns can be employed in various platforms or services, such as publishing workflow tools, semantic web tools, metadata exchange, social network, linked data, authoring and reviewing tools.

- SKO Patterns are compatible with other prominent scientific annotation ontologies.

2.7 Liabilities

- High cost of metadata generation.

- High cost of metadata maintenance.

2.8 Know Uses

"Article of the Future"

From the first issue of year 2010, the journal of Cell (http://www.cell.com) began to launch a new format for online presentation of all research articles. The "Article of the Future" initiative (http://beta.cell.com/) aims to evolve the concept of a scientific publication in step with the development of new technologies and functionalities. Cell aims to develop an online format which breaks the constraints of traditional linear structured paper and allows individual reader to create a personalized path through the discourse's content based on one's own interests or needs. "Article of the Future" proposed a new approach to organizing the traditional sections of the article, moving away from a strictly linear structure required by print towards a more integrated and linked structure. Tabbed and hyperlinked navigation through the Summary, Introduction, Results, Discussion, Experimental Procedures, Data, References, Supplemental Information, Related Information and Comments allows subject-area researchers to quickly access in-depth information on a specific experiment result, while providing more general readers a choice to gain the conceptual insights without being overwhelmed by additional details.

3. CONCLUSION

In this paper, we propose the Scientific Knowledge Object Patterns for solving problems of explicit representation of the semantics of scientific discourses. The patterns mainly serve in the semantic layer of SKO, and three possible serialization patterns derived from logical reasoning, i.e. Deduction, Induction and Abduction, have also been discussed.

Presently, we initiate a "Conference of the Future" project, which would be a first comprehensive scientific publishing platform equipped with SKO Patterns along with metadata schemes. Our ultimate goal is to provide a high-level pattern language for the externalization of the rhetoric and argumentation captured within Scientific Knowledge Objects, such as papers, which will facilitate discovery, dissemination and reuse of scientific knowledge in research communities.

4. ACKNOWLEDGMENTS

We are deeply grateful to Prof. Paris Avgeriou for his excellent shepherding. We thank Prof. James Noble, Andreas Ruping, Tim Wellhausen, James Siddle and Mandy Chessel for their detailed feedback during the workshop at Irsee. This research is partially supported by the European Union FP7 project LiquidPublication (grant no. 213360).

5. REFERENCES

[1] de Waard, A. and Tel, G. 2006. The abcde format enabling semantic conference proceedings. *In SemWiki*, 2006.

[2] Giunchiglia, F., and Chenu, R. 2009. *Scientifc knowledge objects v.1.* Technical Report DISI-09-006, University of Trento, January 2009.

[3] Groza, T., Handschuh, S., Clark, T., Buckingham Shum, S. and de Waard, A. 2009. A short survey of discourse representation models. *In: Proceedings 8th International Semantic Web Conference, Workshop on Semantic Web Applications in Scientific Discourse.* Lecture Notes in Computer Science, Springer Verlag: Berlin, 26 Oct 2009, Washington DC.

[4] Groza, T., Handschuh, S., M oller, K., Decker, S. 2007. SALT - Semantically Annotated LATEX for Scientific Publications. *In: Proceedings of the 4th European Semantic Web Conference (ESWC 2007)*, Innsbruck, Austria

[5] Harmsze, F.A.P. 2000. PhD thesis: *A modular structure for scientific articles in an electronic environment* (ISBN 90-9013486-7)

[6] Mann, W.C., Thompson, S.A.1987. *Rhetorical Structure Theory: A theory of text organization.* Technical Report RS-87-190, Information Science Institute

[7] Sanders, T.J.M., Spooren, W.P.M., Noordman, L.G.M.1993. Coherence Relations in a Cognitive Theory of Discourse Representation. *Cognitive Linguistics* 4(2) (1993) 93–133

[8] Swales, J.M. 1990. *Genre Analysis: English in Academic and Research Settings*, Cambridge University Press.