How can Linked Data facilitate scientific publishing and knowledge exchange?

Sören Auer

Fraunhofer IAIS

universitätbonn
Rheinische Friedrich-Wilhelms-Universität Bonn

LOD2
Creating Knowledge
out of Interlinked Data

Scientific Knowledge Exchange

Raw data – Data Portals, Data Citation

Publications – Semantic Annotation

Textbooks, CourseWare – OCW

Informal communication
(social networks)

LOD2
Creating Knowledge out of Interlinked Data

State of Play – Semantic Annotation of Publications

There are quite comprehensive approaches for semantic representation of scholary content

OMDoc for mathematical knowledge

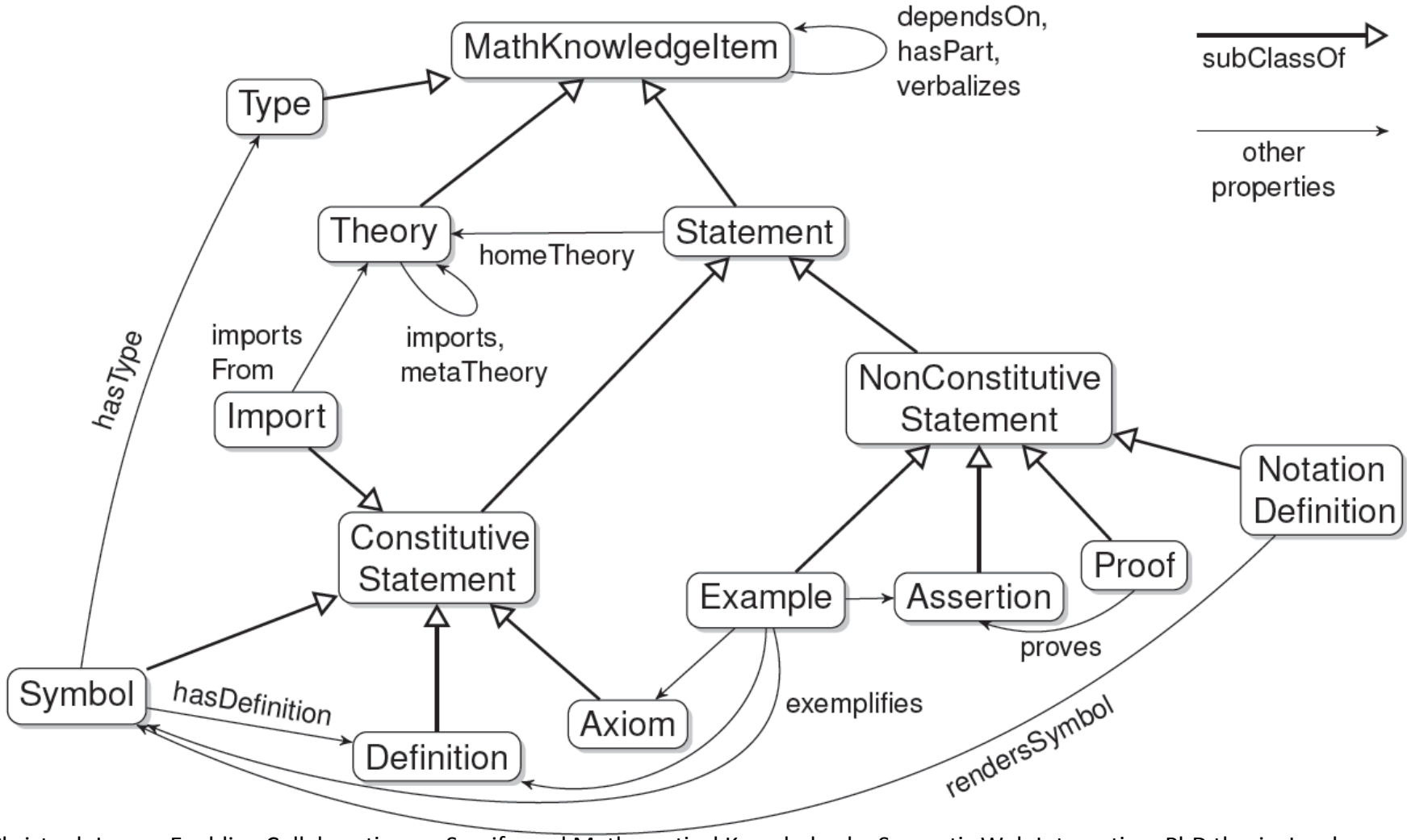
Semantic Annotation for LaTeX

…

But we lack a truly disruptive **architecture of participation** for linking and contextualizing

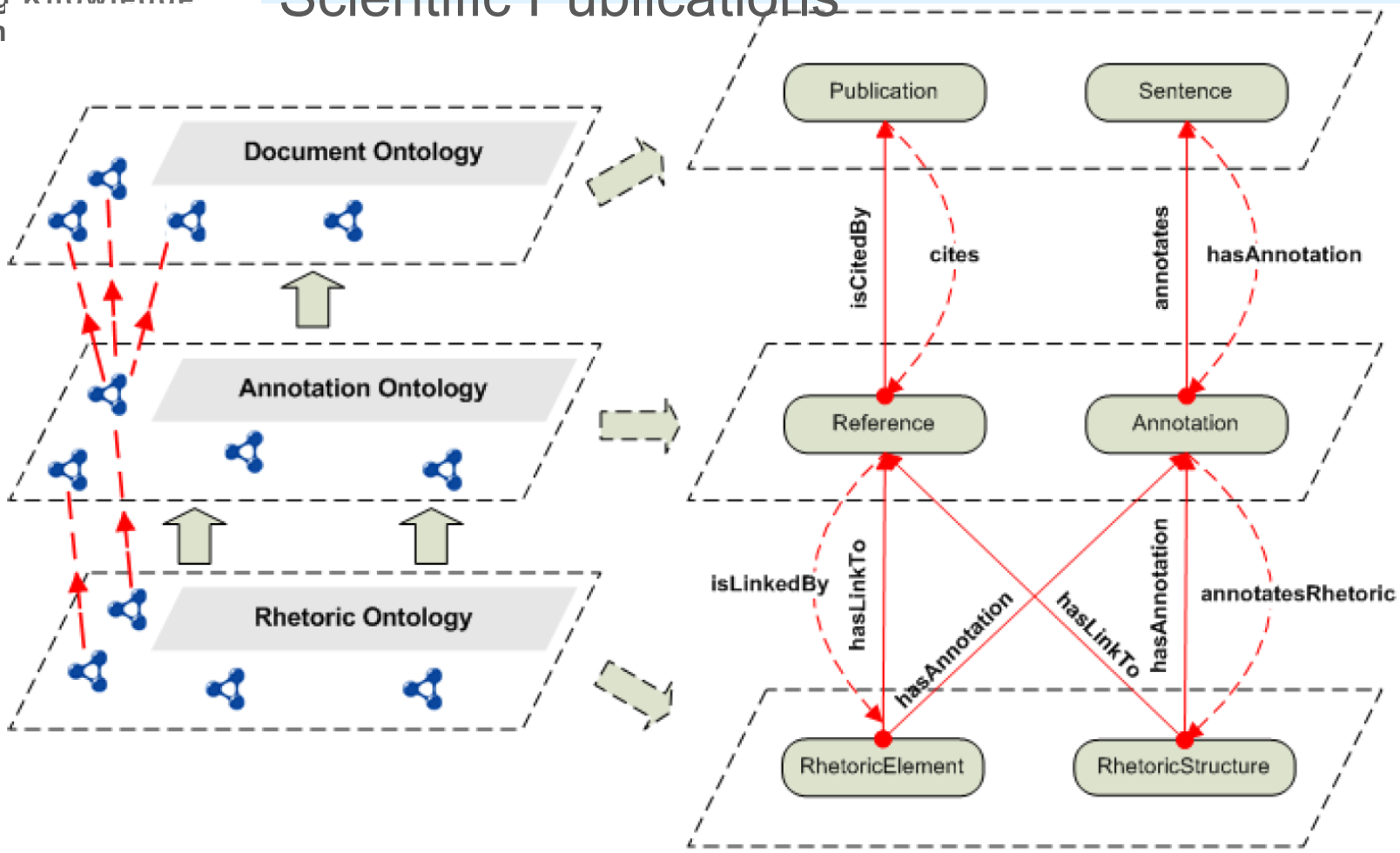
LOD2
Creating Knowledge
out of Interlinked Data

OMDoc



Christoph Lange. Enabling Collaboration on Semiformal Mathematical Knowledge by Semantic Web Integration. PhD thesis, Jacobs University Bremen, 2011, number 11 in Studies on the Semantic Web. AKA Verlag and IOS Press, Heidelberg and Amsterdam, 2011.

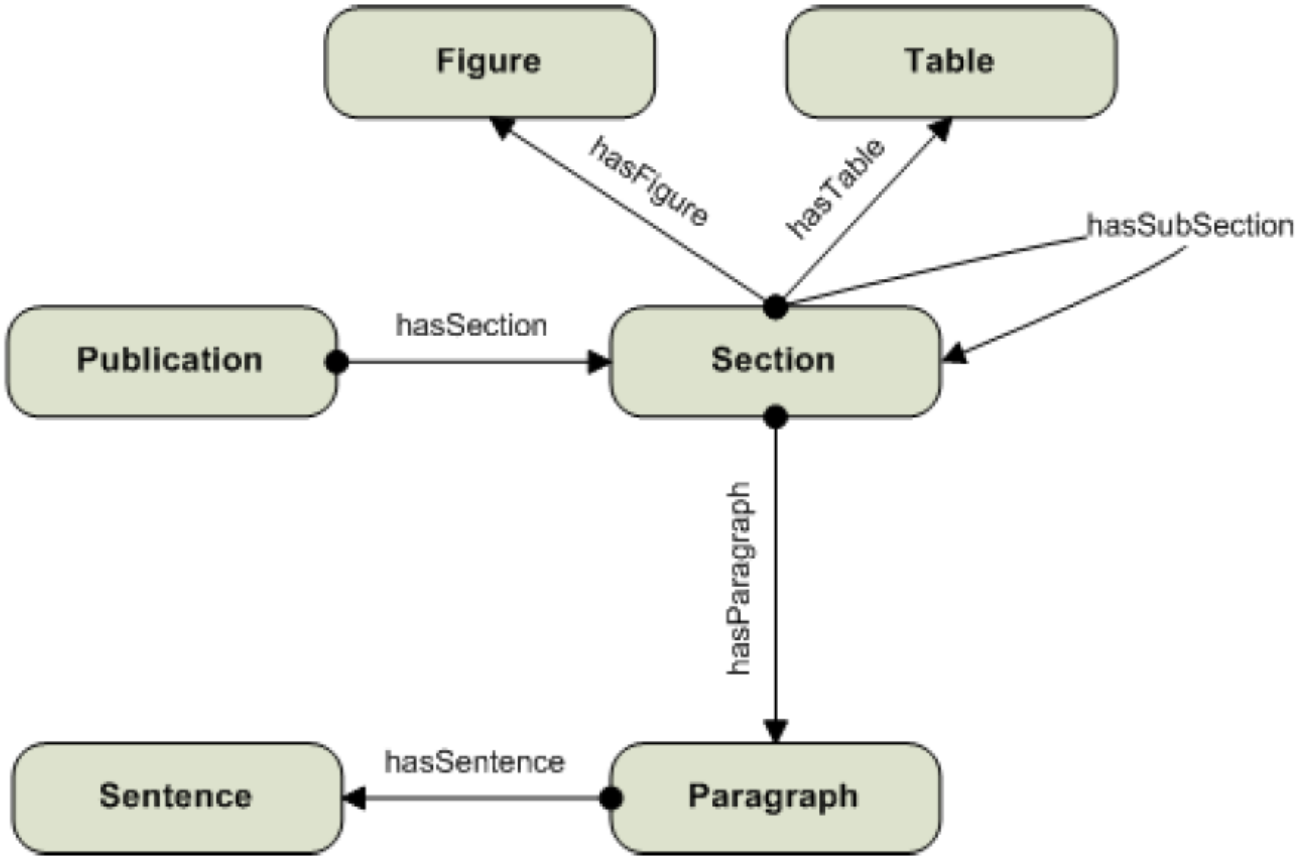
SALT - Semantically Annotated LATEX for Scientific Publications



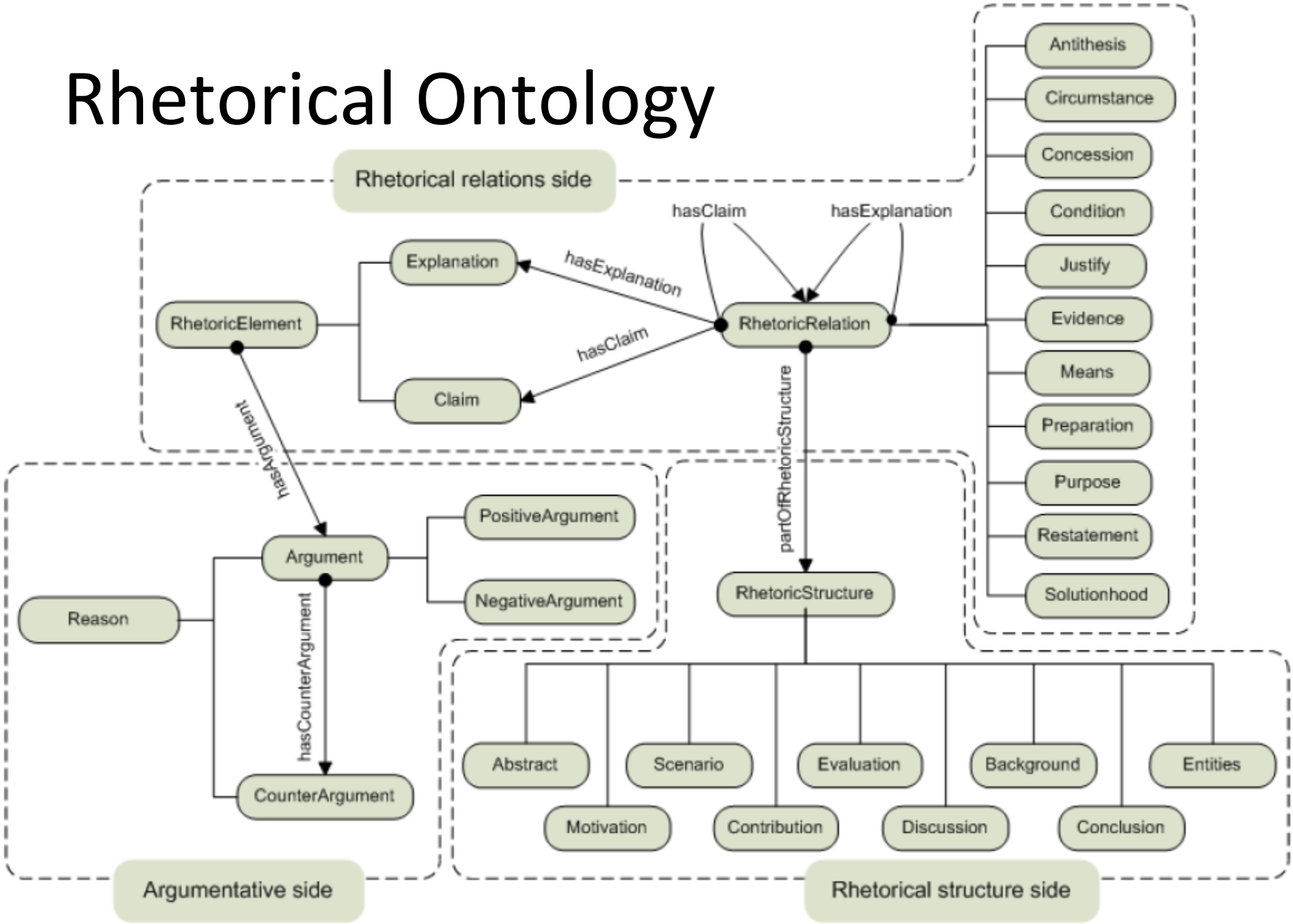
SALT-Semantically Annotated LaTeX for Scientific Publications

T Groza, S Handschuh, K Möller, S Decker - The Semantic Web: Research and Applications, 2007

Document Ontology

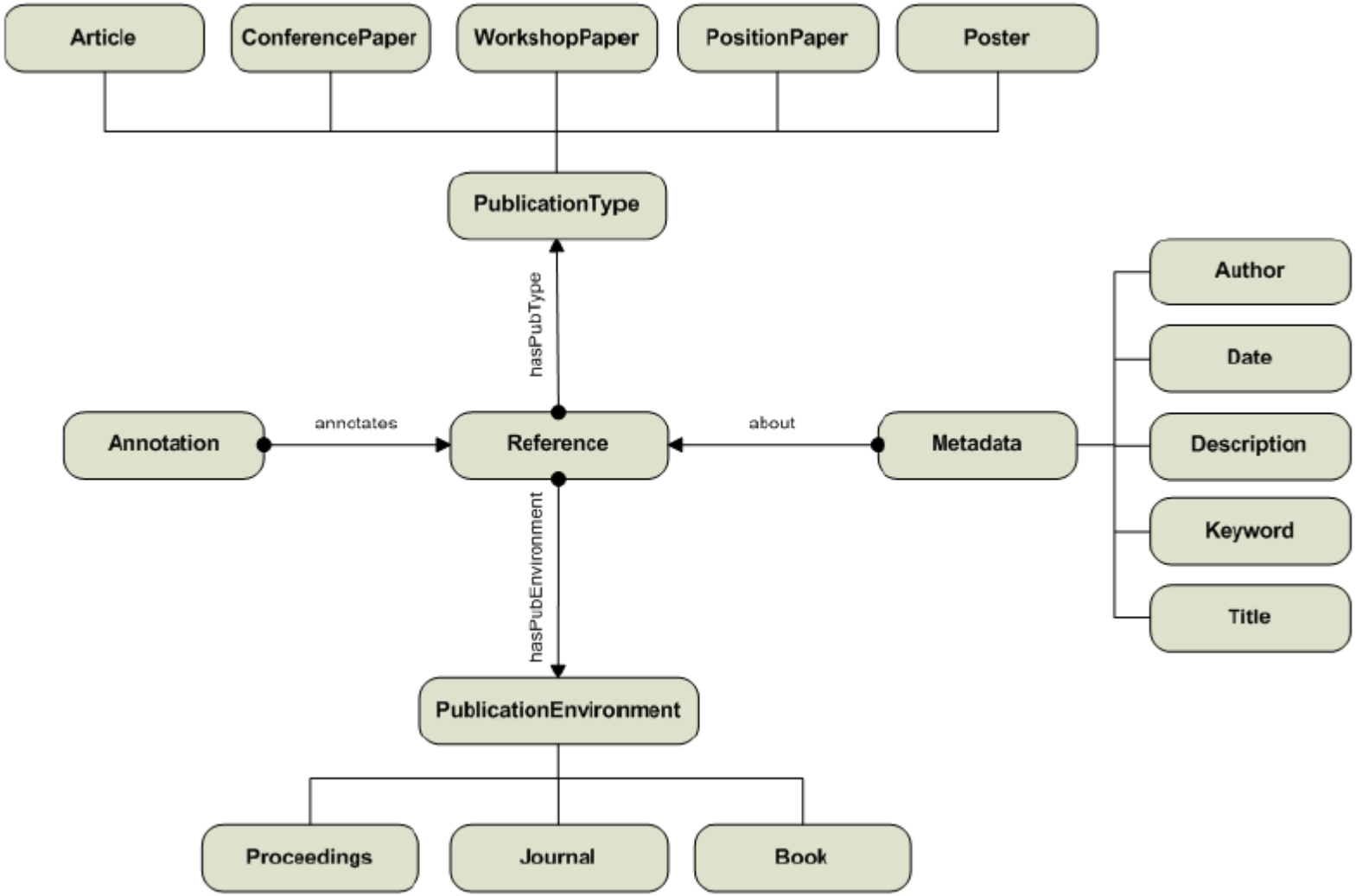


Rhetorical Ontology



LOD2
Creating Knowledge
out of Interlinked Data

Annotation Ontology



LOD2
Creating Knowledge
out of Interlinked Data

Semantically describing the content of scientific publications

LIMES — A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data

Axel-Cyrille Ngonga Ngomo, Sören Auer
AKSW/BIS, Institut für Informatik
Universität Leipzig
Postfach 100920, 04009 Leipzig, Germany
{ngonga|auer}@informatik.uni-leipzig.de

Abstract

The Linked Data paradigm has evolved into a powerful enabler for the transition from the document-oriented Web into the Semantic Web. While the amount of data published as Linked Data grows steadily and has surpassed 25 billion triples, less than 5% of these triples are links between knowledge bases. Link discovery frameworks provide the functionality necessary to discover missing links between knowledge bases. Yet, this task requires a significant amount of time, especially when it is carried out on large data sets. This paper presents and evaluates LIMES, a novel time-efficient approach for link discovery in metric spaces. Our approach utilizes the mathematical characteristics of metric spaces during the mapping process to filter out a large number of those instance pairs that do not suffice the mapping conditions. We present the mathematical foundation and the core algorithms employed in LIMES. We evaluate our algorithms with synthetic data to elucidate their behavior on small and large data sets with different configurations and compare the runtime of LIMES with another state-of-the-art link discovery tool.

1 Introduction

The core idea behind the Linked Data paradigm is to facilitate the transition from the document-oriented Web to the Semantic Web by extending the Web with a data commons consisting of interlinked data sources [Volz *et al.*, 2009]. While the number of triples in data sources increases steadily and has surpassed 25 billions, links still constitute less than 5% of the total number of triples available on the Linked Data Web[1]. In addition, while the number of tools for publishing Linked Data on the Web grows steadily, there is a significant lack of time-efficient solutions for discovering links between these data sets. Yet, links between knowledge bases play a key role in important tasks such as cross-ontology question answering [Lopez *et al.*, 2009], large-scale inferences [Urbani *et al.*, 2010] and data integration [Ben-David *et al.*, 2010].

[1] http://www4.wiwiss.fu-berlin.de/lodcloud/

To carry out a matching task, the distance measure as defined by the user is usually applied to the value of some properties of instances from the source S and target T so as to detect instances that should be linked. Instances whose distance is lower or equal to a given threshold are considered to be candidates for linkage. The a-priori complexity of a matching task is proportional to $|S||T|$, an unpractical proposition as soon as the source and target knowledge bases become large. For example, discovering duplicate cities in DBpedia [Auer *et al.*, 2008] alone would necessitate approximately 0.15×10^9 distance computations. Hence, the provision of time-efficient approaches for the reduction of the time complexity of link discovery is a key challenge of the Linked Data.

In this paper, we present LIMES (Link Discovery Framework for metric spaces) - a time-efficient approach for the discovery of links between Link Data sources. LIMES addresses the scalability problem of link discovery by utilizing the *triangle inequality* in metric spaces to compute pessimistic estimates of instance similarities. Based on these approximations, LIMES can filter out a large number of instance pairs that cannot suffice the matching condition set by the user. The real similarities of the remaining instance pairs are then computed and the matching instances are returned. We show that LIMES requires a significantly smaller number of *comparisons* than brute force approaches by using synthetic data. In addition, we show that our approach is superior to state-of-the-art link discovery frameworks by comparing their runtime in real-world use cases. Our contributions are as follows:

- We present a lossless and time-efficient approach for the large-scale matching of instances in metric spaces.
- We present two novel algorithms for the efficient approximation of distances within metric spaces based on the triangle inequality.
- We evaluate LIMES on synthetic data by using the number of comparisons necessary to complete the given matching task and with real data against the SILK framework [Volz *et al.*, 2009] with respect to the runtime.

The remainder of this paper is structured as follows: after reviewing related work in Section 2 we develop the mathematical framework underlying LIMES in Section 3. We present the LIMES approach in Section 4 and report on the results of an experimental evaluation in Section 5. We conclude with a discussion and an outlook on future work in Section 6.

```
limes-paper describes   appr123
appr123     a           approach
appr123     for         Link_Discovery
appr123     hasProp     looseless
...

limes-paper describes   impl123
impl123     a           implementation
impl123     implements  appr123
impl123     language    Java
...

limes-paper describes   eval123
eval123     a           evaluation
eval123     evaluates   impl123
eval123     uses        DBpedia
...
```

LOD2
Creating Knowledge out of Interlinked Data

How can we create an architecture of participation for semantic annotation

Provide **instant benefits** for semantic annotations, e.g.:

- Find related work
- Gain reputation on social networks
- Visualization
- Fun

Provide **medium/long-term benefits** for semantic annotations, e.g.:

- More citations
- More funding

LOD2
Creating Knowledge
out of Interlinked Data

Distributed Semantic Annotations

Semantic annotations represented in RDF and Linked Data can be stored in different Digital Libraries, repositories, OpenCourseWare …

- Using shared vocabuaries and semantic search (e.g. sindice) these can be glued together

A Semantic Wikipedia for Science

- Papers instead of Wiki articles

Thanks for your attention!

Sören Auer

http://www.iai.uni-bonn.de/~auer | http://aksw.org | http://lod2.org

auer@cs.uni-bonn.de