ScienceWISE: A Web-based Interactive Semantic Platform for Paper Annotation and Ontology Editing

Anton Astafiev[3], Roman Prokofyev[4], Christophe Guéret[6], Alexey Boyarsky[123], and Oleg Ruchayskiy[5]

[1] Ecole Polytechnique Fédérale de Lausanne, Switzerland
{firstname.lastname}@epfl.ch
[2] Instituut-Lorentz for Theoretical Physics, Universiteit Leiden, Nederlands
[3] Bogolyubov Institute for Theoretical Physics, Kiev, Ukraine
[4] eXascale Infolab, University of Fribourg, Switzerland
{firstname.lastname}@unifr.ch
[5] CERN TH-Division, PH-TH, Geneva 23, Switzerland
oleg.ruchayskiy@cern.ch
[6] Vrije Universiteit Amsterdam, Nederlands
c.d.m.gueret@vu.nl

Abstract. The ScienceWISE system is a collaborative ontology editor and paper annotation tool designed to help researchers in their discovery. In this paper, we describe the system currently deployed at sciencewise.info and the exposition of its data as Linked Data. During the "RDFization" process, we faced issues to encode the knowledge base in SKOS and find resources to link to on the LOD. We discuss these issues and the remaining open challenges to implement some target features.

1 Introduction

Organizing scientific knowledge in systematic ways becomes increasingly important. However, the creation of intra- and inter-disciplinary knowledge bases is hindered by the heterogeneity and the scale of the information to consider. This calls for *scientific community-run systems* (replacing classical publishers of encyclopedias) allowing to combine presentation of new results, in-depth discussions, "user-friendly" introductions for young scientists, and meta-data to relate semantically similar concepts or pieces of content. Today, there are no standard tools to insert, store and query such meta-data online, which mostly remains "in the heads of the experts". To address these pertinent issues and make a first step towards the creation of tools for the automated support of the scientific process, a group of physicists from EPFL and CERN together with computer scientists from EPFL, the University of Fribourg and VU created the *ScienceWISE system — a system for semantically importing, storing and searching scientific data.*

ScienceWISE[7] allows a community of scientists, working in a specific domain, to generate dynamically as part of their daily work an *interactive semantic en-*

[7] Accessible at http://ScienceWISE.info

vironment, *i.e.*, a field-specific ontology with direct connections to the text of research papers. ScienceWISE is currently being essentially used by physicists and is connected with the `ArXiv.org` archive of papers. It is however designed not to be field specific and can be re-deployed to be used by other scientific communities. In the following, we will use ScienceWISE to refer to both the system and its deployed version at `http://ScienceWISE.info`.

The rest of the paper describes ScienceWISE general (Section 2) and the exposition of its data as Linked Open Data (Section 3). We conclude describing the problems faced while looking at exposing the data from ScienceWISE as Linked Data and sketch paths for future work (Section 4) .

2 Architecture of ScienceWISE

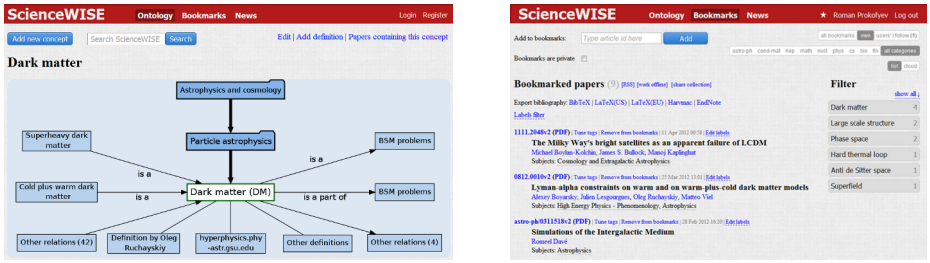
The ScienceWISE system is an eco-system currently compromising three type of entities. The **Ontology**: the knowledge-graphs which captures the concepts and their complex relationships; The **Users**: the social community of experts in a field which give local, noisy and incomplete knowledge on some parts of the ontology; The **Portal**: the web application which consolidates all local inputs with the current ontology and attempts to create a comprehensive, global and dynamic knowledge system. There are plans to add a fourth item to the list [1]: an intelligent assistant able to leverage the knowledge expressed in the Ontology and assist the Users in their research activities. The exposition of the data from ScienceWISE as Linked Data (see Section 3) is a first step in this direction.

ScienceWISE Ontology The ontology used to tag the papers is enriched and curated the users of the system. To create the initial version of the ontology, we have performed a semi-automated import from many science-oriented ontologies and online encyclopedias. After this initial step, ScienceWISE users (who are domain experts) are allowed to edit elements of the ontology (*e.g.*, adding new definitions or new relations) in order to improve its quality. Presently, the ScienceWISE ontology counts more than 60 000 unique entries, each with its own definitions, alternative forms, and semantic relations to other entries. The semantic relations are both of general (e.g., *is a part of*) and field-specific (*is a model of*, *is observed in*) nature.

ScienceWISE Users The system is public since April 2011 and accessible by scientists via `ArXiv.org` as well as via the CERN Document Server[8] and Inspire[9]. The system currently counts above 200 active users, thousands of conceptually indexed and annotated papers, and is now receiving several new registrations *daily*.

[8] `http://cds.cern.ch`
[9] `http://inspirebeta.net`: comprehensive bibliographic database in high-energy

ScienceWISE Portal The ScienceWISE portal is the main interface for interacting with the system. The ontology explorer (see Fig. 1a) shows information about concepts. Competing scientific viewpoints about the same concept are represented as alternative resources and definitions. In the tagging interface (see Fig. 1b) a user is presented with a automatically populated list of relevant concepts to pick from to annotate the paper. There is also the possibility to create and explore collection of papers. The portal is implemented in Python and uses PostgreSQL as a data back-end. It uses TeXpp[10] to browse the content of LaTeXfiles and spot concepts from the ontology.



(a) Representation of the ontology

(b) Paper tagging interface

Fig. 1: Two screen captures of the ScienceWISE portal

3 ScienceWISE and Semantic Web technologies

Already now ScienceWISE enable the construction of a highly-structured, collaboratively curated and expressive ontology. However, *the ultimate goal of our system* is to create a performant, user-friendly and customizable integrated environment to help scientists save time and effort in their daily work, while building complex ontological networks that capture their scientific findings [1]. This knowledge acquisition process will be conducted in a pervasive way, "in the background", harvesting data from different source and combining it with the ScienceWISE ontology.

The need for smooth data integration capabilities with other data sets on the web and the necessity for reasoning processes able to help scientists drove us toward considering the usage Semantic Web technologies. The Semantic Web enhanced version of ScienceWISE is depicted on Fig. 2.

"RDFization" To leverage on the existing Semantic Web technologies (deductive reasoning, relation finding, ontology matching) we have performed an "RDFization" of the ScienceWISE ontology. D2R periodically exports the content from the relational database as RDF and pushes it to an OWLIM triple store. The vocabulary used is mainly a combination of SKOS, RDF and RDFS.

[10] `http://code.google.com/p/texpp/`

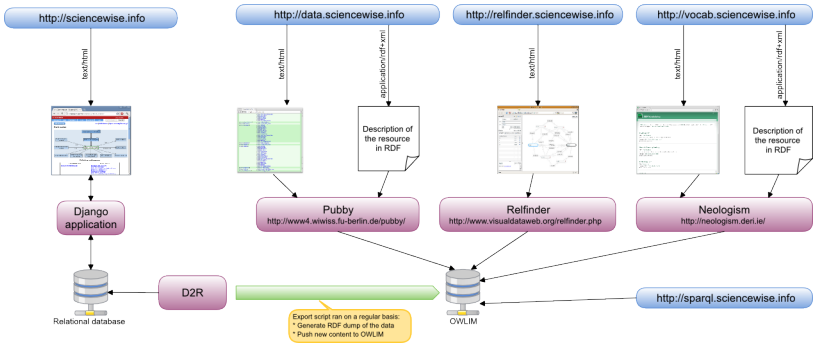
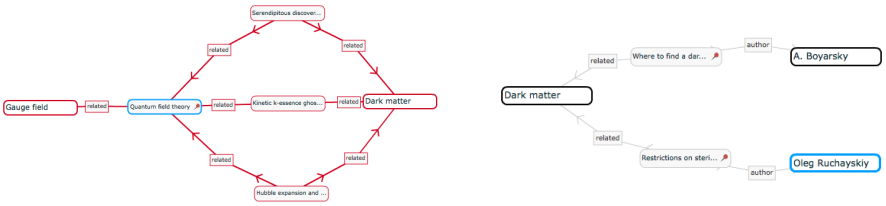


Fig. 2: Global architecture of ScienceWISE showing the "RDFization" of the ScienceWISE data. The data exposed describe papers, concepts and authors

However, some of the field-specific relations (such as "is a mechanism of") can not be directly mapped to SKOS so we created a vocabulary which extends SKOS to match our needs. This vocabulary is published using Neologism and is available at `http://vocab.sciencewise.info/ontology`. Finally, Pubby is used to serve de-referencable URIs for the resources at `http://data.sciencewise.info/`.

Initial outcomes Having the data from ScienceWISE transcoded in a graph-db enables finding non trivial paths between the different nodes of this graph. In ScienceWISE, the nodes are papers, concepts and authors. We deployed RelFinder[11] on top the SPARQL end point to rapidly obtain a tool able to find and display these paths (*c.f.* Fig. 3). We have extended it by adding a possibility to ignore some nodes, defined via configuration file, and have plans to extend it further more.



(a) Scientific concepts ("*Dark matter*" and "*Gauge field*") related through their co-occurrence in research papers

(b) Authors who did not co-author but had written about the same subject ("*Dark Matter*")

Fig. 3: Finding relation between entities of different nature with Relfinder

[11] `http://www.visualdataweb.org/relfinder.php`

4 Current challenges and open questions

The publication of the data from ScienceWISE as Linked Data and the usage of Relfinder are a first step and a number goals can reached if we are able to utilize and extend beyond-the-state-of-the-art existing Semantic Web technologies. However we are facing some major issues that slow us down or block some aspects of the development:

Semantic structures: The level of abstraction of scientific concepts is highly non-trivial. For most of the concepts (apart from the "named entities": proteins, particles, celestial objects) it is typical to be an *instance* of one class and a *subclass* of the other class at the same time. Extensions of SKOS and reasoners aware of this are desired;

Resource discovery in LOD: One of the main problems, that we have encountered in bringing the ScienceWISE data to LOD was the absence of any "browsing" capabilities for resources in the LOD cloud (beyond clickable version of the picture[12] and some basic unstructured tags). Without an easy way to find re-usable URIs we had to mint URIs for entities which are actually not described in the system.

Resource matching: The matching between resources from various data sources and those from ScienceWISE is important. Our attempts at using semi-automated tools like SiLK resulted in too many false positive, even with a conservative strategy. We need more flexible matching tool and we need to integrate them within the Portal to ensure human validation of the links.

ScienceWISE is a working system that is being used by a growing amount of Physicists to annotate paper, discuss related concepts and express diverging opinions. In order to further develop the capabilities of the system and share its data, we have started using Semantic Web technologies and applied Linked Data publication principles. Our first results are promising, showing already some added value, but are limited by a number of problems and challenges we are facing. We described them in the paper and a path for future work and a call for guidance from the Semantic Web research community.

Acknowledgments This project is carried out as part of the "AAA/SWITCH – e-infrastructure for e-science" program under the leadership of the Swiss National Research and Education Network. It has been supported (in part) by funds from the ETH Board, the Swiss National Science Foundation under grant number PP00P2_128459, and the European Community's Seventh Framework Programme (FP7/20072013) under Grant Agreement n°256975, LOD Around The Clock (LATC) Support Action.

References

1. K. Aberer, A. Boyarsky, P. Cudré-Mauroux, G. Demartini, and O. Ruchayskiy. ScienceWISE : a web-based interactive semantic platform for scientific collaboration. In *Proceedings of ISWC2011 - "Outrageous ideas" track*, 2011.

[12] `http://richard.cyganiak.de/2007/10/lod/imagemap.html`