The SciLink Project — From Document-centric to Resource-oriented Publications

Bernhard Haslhofer
Cornell University
Department of Information Science
301 College Avenue, Ithaca, NY, USA
bernhard.haslhofer@cornell.edu

ABSTRACT

The World Wide Web has changed the way of publishing and distributing scholarly results. However, scholarly publications are still organized linearly and point to supplemental or related information only by textual references or at most by hyperlinks embedded into PDF documents. They are stored in closed repositories and we can hardly access, navigate, and use scholarly resources the way we do it with other Web resources. The goal of the SciLink project is to analyze scholarly practices in certain pilot communities, to learn how scholars currently use Web resources in their publications, and to design tools that help scholars in aggregating and publishing the building blocks of their works in a way that integrates with the resource-oriented Architecture of the World Wide Web. In that way, we want to allow humans and machine agents to access and interact with scholarly resources just like with any other Web resource.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.3.7 [**Digital Libraries**]: System issues

General Terms

Design, Management, User Behavior

1. INTRODUCTION

The World Wide Web has revolutionized the way of publishing and distributing scholarly results. While printed books and journals were the primary publication medium during the past centuries, today's scientific results are increasingly disseminated via the Web, mostly in the form of scholarly publications.

Currently, however, publications on the Web still resemble the traditional print production process: they are static documents organized linearly into chapters and sections and include mostly text and figures. They are usually stored in isolated, often closed environments such as publication databases or institutional repositories, and are hardly linked with other resources on the Web. Because of their document-centric nature they often do not provide information relevant for the research context. They often ignore data sets required for reproducing the described experiments and other supplemental materials such as conference presentation, blog entries, etc. So it seems that scholarly communication does not yet make full use of the potentials of the World Wide Web and we share the opinion of others (see e.g., [1, 3]) that it is time to reassess the traditional way of publishing scholarly results.

The overall goal of the SciLink project is to analyze scholarly practices in some pilot research communities and investigate how and to what extent scholars make use of Web resources in their publications. From these findings we want to derive design decisions and implement tools that help scholars in organizing the building blocks of their publications in a way that integrates with the architectural principles of the World Wide Web. We will refrain from the current document-centric point of view, where publications are isolated artifacts stored in closed repositories, and want to apply a resource-oriented view, in which the building blocks of publications become aggregations of Web resources. Linked Data [2], as a method for exposing data on the Web, and existing aggregation data models such as OAI-ORE [4] will play a major role in our efforts to achieve that goal.

2. APPROACH

Within the scope of the SciLink project, we will analyze existing corpora of scholarly literature from some pilot communities in different research areas such as physics, chemistry, economy, and social sciences. Given the large amount of publications we are dealing with, we will, at least in the early project stage, pursue a purely quantitative study on existing full text corpora in order to generate first insights for a possible deeper investigation. We would like to understand in which parts of their publications scholars reference Web resources, to what kind of Web resources they are referencing (e.g., datasets, personal web sites, other publications) and how linking to Web resource in scholarly publications has changed over time.

From these findings we hope to be able to derive design decisions and develop mechanisms that allow scholars to organize their results as aggregations of Web resources. Figure 1 illustrates the concept of resource-oriented publications by a real-world example. The left-hand side shows an excerpt of a scholarly publication, which has been published in the Nature Journal and is now available as PDF document at

iConference 2012, February 7-10, 2012, Toronto, Ontario, Canada.

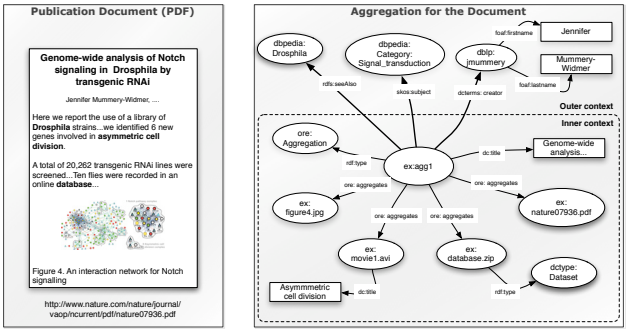

Figure 1: Sample publication document represented as Web-accessible resource aggregation. The prefixes (ex, dblp, ore, etc.) denote dereferencable URI spaces.

the given URL. On the right-hand side, we represent the building blocks of that publication as a Web-accessible aggregation: in the first place, it defines links to resources that are a direct product of the research work carried out. This includes the publication document (`ex:nature07936.pdf`), data sets used for experiments (`ex:database.zip`), but also multimedia material such as a video recording of the experiments (`ex:movie1.avi`). We denote these resources and the links between them as the inner context of a scientific publication. Secondly, it shows how aggregations (and also parts thereof) can be linked with semantically relevant resources that are not a direct product of the described research work, denoted as outer context of a scientific publication. This includes links that identify the authors and/or their affiliations as unique entities (`dblp:jmummery`), or information on the background of the described research topic, such as articles from the Online Encyclopedia Wikipedia (`dbpedia:Drosphila`). We could also explicitly categorize a publication using Wikipedia categories (`dbpedia:Category:Signal_transduction`) or with terms from any other Web-available knowledge organization system.

Introducing new tools without giving scholars an incentive to adapt established practices often does not lead to the desired result (see e.g., [5]). The challenging design question is how to integrate such mechanisms with the existing technical infrastructure currently used in scholarly authoring and dissemination processes. Our preliminary idea is to enhance article submission interfaces of existing repositories with functionality that performs the transformation from static and linear (mostly PDF) documents into networked resource aggregations at submission time. We can use existing software libraries for publishing the building blocks of scholarly publications as Web resources and provide domain-tailored semi-automatic solutions for linking scholarly publications with their outer context, i.e., other contextually relevant resources on the Web. We also would like to apply the Linked Data principles on the resulting resource aggregations so that scholarly resources become part of a global data space. Because of the ephemeral nature of Web resource it will also be necessary to investigate solutions for synchronizing Web resources.

3. PRELIMINARY RESULTS

We are still in an early project phase and are currently analyzing how scholars in our pilot communities are using HTTP links in their publications. We developed generic software libraries for extracting and analyzing Web resources in scholarly literature corpora and applied them to the arxiv.org corpus, which consists of approximately 700K full text documents. Initial analysis results based on the extracted raw dataset indicate a linear increase in the average amount of HTTP links per publication in arxiv.org, which covers mainly publications in the area of physics but also other areas such as mathematics and computer science. We also noticed a bimodal distribution in the relative position of HTTP links in scholarly publications, with modes differing by sub-discipline. This suggests that it is becoming best-practice to use HTTP links in the reference sections and in certain communities also at the beginning of publications.

4. SUMMARY AND FUTURE WORK

In the SciLink project we wan to learn how scholars in different areas use Web resources in their publications. From our findings we would like to derive design decisions for enhancing existing scholarly infrastructures so that the building blocks of scholarly publications better integrate with the resource-oriented architecture of the World Wide Web.

At the moment, we are still in the phase of analyzing the use of HTTP links in PDF documents in the arxiv.org publication corpus. In the next step, we would like to enhance our analysis framework with advanced functions that help us understanding what kind of resources people are linking to. Then we would like to extend the analysis to other publication corpora from other scholarly areas and identify differences in linking behavior.

5. ACKNOWLEDGMENTS

This work is supported by a Marie Curie International Outgoing Fellowship (IOF) within the 7th European Community Framework Program.

6. REFERENCES

[1] D. E. Atkins, K. K. Droegemeier, S. I. Feldman, H. G. Molina, M. L. Klein, D. G. Messerschmitt, P. Messina, J. P. Ostriker, and M. H. Wright. Revolutionizing science and engineering through cyberinfrastructure. Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, January 2003.

[2] C. Bizer, T. Heath, and T. Berners-Lee. Linked data — the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2009.

[3] K. H. Buetow. Cyberinfrastructure: Empowering a "third way" in biomedical research. *Science*, 308(5723):821–824, May 2005.

[4] C. Lagoze, H. Van de Sompel, P. Johnston, M. Nelson, R. Sanderson, and S. Warner. *Open Archives Initiative Object Reuse and Exchange (OAI-ORE)*. Open Archives Initative, October 2008. Available at: `http://www.openarchives.org/ore/1.0/primer.html`.

[5] T. Velden and C. Lagoze. Communicating chemistry. *Nat Chem*, 1(9):673–678, dec 2009.