sClippy: Connecting Personal Information and Linked Open Data

Tudor Groza, Laura Drăgan, Siegfried Handschuh and Stefan Decker

Digital Enterprise Research Institute
National University of Ireland, Galway
IDA Business Park, Lower Dangan
Galway, Ireland
{tudor.groza, laura.dragan, siegfried.handschuh, stefan.decker}@deri.org

ABSTRACT

The exponential growth of the World Wide Web in the last decade, brought an explosion in the information space, which has important consequences also in the area of scientific research. Thus, finding relevant work in a particular field and exploring the links between publications is quite a cumbersome task. Similarly, on the desktop, managing the publications acquired over time can represent a real challenge. Extracting semantic metadata, exploring the linked data cloud and using the semantic desktop for managing personal information represent, in part, solutions for different aspects of the above mentioned issues. In this poster/demo, we show an innovative approach for bridging these three directions with the overall goal of alleviating the information overload problem burdening early stage researchers.

Categories and Subject Descriptors

I.7.5 [**Document and Text Processing**]: Document Capture; H.3.1 [**Information Systems**]: Content Analysis and Indexing; H.3.3 [**Information Systems**]: Information Search and Retrieval

Keywords

Semantic metadata, Semantic Desktop, Linked Data

1. MOTIVATION

The World Wide Web represents an essential factor in the dissemination of scientific work in many fields. At the same time, its exponential growth is reflected in the substantial increase of the amount of scientific research being published. In addition, with the lack of uniformity and integration of access to information, even within communities in the same domain, there is no central hub that stores common information. Consequently, this makes the process of finding and linking relevant work in a particular field a cumbersome task.

On the desktop, we find a somewhat similar problem, though on a smaller scale. A typical researcher stores an significant number of publications over time. Generally, the files representing these publications have a non-intuitive name (often the same cryptic name assigned by the system publishing them), and may, in the best case scenario, be structured in intuitive folder hierarchies. Thus, finding publications or links between them represents quite a challenge, even with the help of tools like Google Desktop.

Semantic Web technologies, with semantic metadata at their foundation, have been proved to help at alleviating, at least partially, the above mentioned issues. Used in particular contexts, semantic metadata enables a more fertile search experience, complementing full text search with search based on different facets. Considering semantic metadata in the context of the two directions, i.e. the Web and the Desktop, we observe that: (i) the <semantic metadata – Linked Data Web[1]> pair, can provide the means for linking publications on the web, while (ii) the <semantic metadata – Semantic Desktop> pair, can help at linking publications and personal information on the desktop.

In this poster/demo, we propose a solution for bridging the two directions, with the goal of enabling a more meaningful searching and linking experience on the desktop, having the linked data cloud as a primary source. The application provides a simple and straightforward way of finding related publications based on the automatically extracted and linked metadata, in parallel with the opportunity of weaving the linked publication data on the desktop, by means of usual desktop applications (e.g. File and Web browser).

2. EXTRACT – EXPAND – INTEGRATE

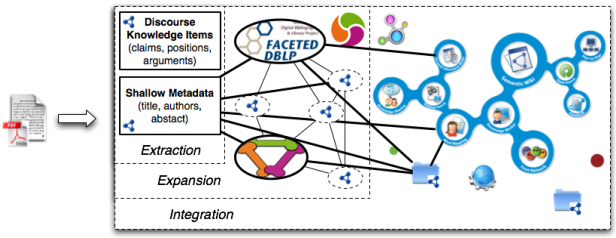

Figure 1: Incremental metadata enrichment process

The application's main goals are reducing the overhead imposed by collateral activities that need to be performed while researching a new field, in parallel with the actual reading of publications, and increasing the user's reward by ensuring a long-term effect of some of the achieved results. An overall figure of the three step process we propose is depicted in Fig. 1.

[1]http://linkeddata.org/

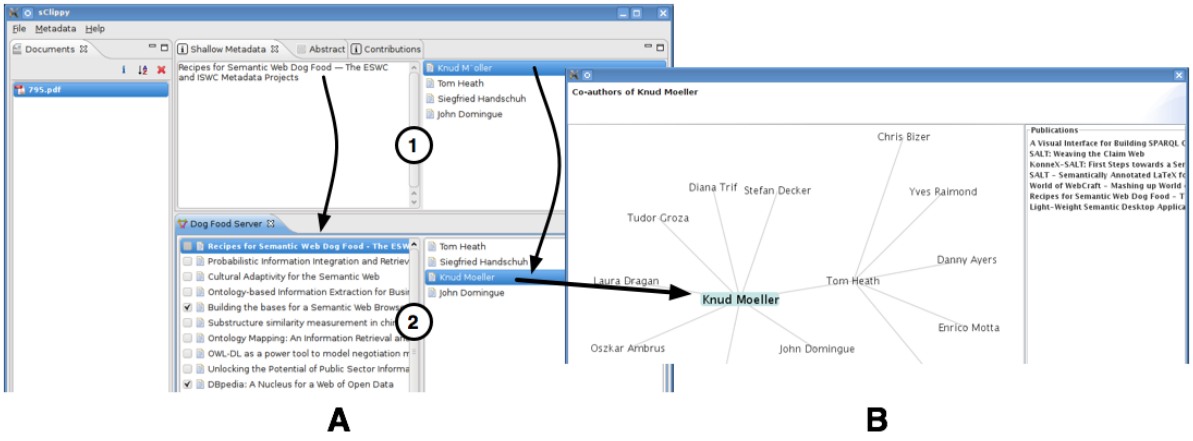

Figure 2: Screenshot of the application's interface: [A] – The main window; [B] – Co-authors graph visualization.

The first step, *extraction* (pointer `1` in Fig. 2 A), has as input a publication with no metadata and it outputs two types of metadata: (i) shallow metadata, i.e. title, authors, abstract, and (ii) deep metadata, i.e. discourse knowledge items like claims, positions or arguments. It represents the only step that appears to have no direct reward (or value) for the user (except for the discourse knowledge items). Nevertheless, it is compulsory in order to start the process, each subsequent step building on its results, and thus enabling an incremental approach to the enrichment of the semantic metadata describing the publication. The extraction process is based on a hybrid 'document engineering – computational linguistic' approach. This makes the resulting metadata prone to containing errors. In terms of document formats, the extraction currently works only on publications encoded as PDF documents and preferably using the ACM and LNCS styles.

The errors resulted from extraction can be corrected in the *expansion* step (pointer `2` in Fig. 2 A), in addition to enriching the basic set of metadata with linked data, coming from different sources. The enrichment is currently achieved the Semantic Web Dog Food Server [2] and the Faceted DBLP[2] linked data repositories. The outcome of the expansion features three elements: (i) a list of candidates, to be used for cleaning and linking the initially extracted metadata, (ii) a list of similar publications, that did not satisfy the shallow entity resolution performed for the previous ones, and (iii) for each author of the given publication found, the full linked model and the complete list of publications existing in the respective repository. For linking purposes, we chose a clear distinction of the semantics of the `owl:sameAs` and `rdfs:seeAlso` relations. We used the former to denote the same instance of an entity present in a different environment, while the latter is used to group together publications sharing common topics. An aside result of the linked metadata is the user's opportunity of navigating through the co-authors networks of a particular author (part B of Fig. 2). An interesting remark here, is that the visualization we have developed can act as a uniform graph visualization tool for any co-author networks emerging from a linked dataset.

Finally, the *integration* step embeds the linked metadata into the semantic desktop environment, thus connecting it deeper within the personal information space, and fostering long-term effects of the overall process. The integration is realized via the services provided by the KDE NEPOMUK Server[3] [1], while the searching and browsing experience is enabled via the usual KDE Desktop applications, such as Dolphin (the equivalent of Windows Explorer) and Konqueror (a KDE Web browser).

Overall, the application[4] we have developed is highly customizable, each step being represented by a module. Therefore, adding more functionality is equivalent to implementing additional modules, for example, an extraction module for MS Word documents, or an expansion module for DBpedia.

3. FUTURE WORK

Each step of the described process has associated a series of open challenges that we intend to address as part of our future work. Firstly, we intend to improve the *extraction* process by developing algorithms that accommodate any formatting style, as well as new extraction modules for other document formats, such as Open Document formats. Regarding the *expansion* step, we will develop modules that link additional repositories, and at the same time, add the option of creating ad-hoc mash-ups between them, thus allowing the user to see data coming from different sources in an integrated and uniform view. Last, but not least, we plan an even tighter *integration* within the Semantic Desktop, therefore enabling more meaningful queries and a richer browsing experience.

Acknowledgments

The work presented in this paper has been funded by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

4. REFERENCES

[1] Ansgar Bernardi, Stefan Decker, Ludger van Elst, Gunnar Grimnes, Tudor Groza, Siegfried Handschuh, Mehdi Jazayeri, Cedric Mesnage, Knud Möller, Gerald Reif, and Michael Sintek. *The Social Semantic Desktop: A New Paradigm Towards Deploying the Semantic Web on the Desktop*. IGI Global, 2008.

[2] Knud Möller, Tom Heath, Siegfried Handschuh, and John Domingue. Recipes for Semantic Web Dog Food – The ESWC and ISWC Metadata Projects. In *Proc. of the 6th Int. Semantic Web Conference*, 2007.

[2] http://dblp.l3s.de/

[3] http://nepomuk.kde.org/

[4] A video-demo of the complete features of the application can be found at http://sclippy.semanticauthoring.org/movie/sclippy.htm