Bridging the Gap between Linked Data and the Semantic Desktop

Tudor Groza, Laura Drăgan, Siegfried Handschuh, and Stefan Decker

DERI, National University of Ireland, Galway
IDA Business Park, Lower Dangan, Galway, Ireland
{tudor.groza,laura.dragan,siegfried.handschuh,stefan.decker}@deri.org
http://www.deri.ie/

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. Finding relevant work in a particular field and exploring the links between publications is currently 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 paper, we propose an innovative approach for bridging these three directions with the overall goal of alleviating the information overload problem burdening early stage researchers. Our application combines harmoniously document engineering-oriented automatic metadata extraction with information expansion and visualization based on linked data, while the resulting documents can be seamlessly integrated into the semantic desktop.

1 Introduction

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. As an example, in the biomedical domain, the well-known MedLine [1] now hosts over 18 million articles, having a growth rate of 0.5 million articles / year, which represents around 1300 articles / day [1]. In addition, we can also mention the lack of uniformity and integration of access to information. Each event has its own online publishing means, and there is no central hub for such information, even within communities in the same domain. Consequently, this makes the process of finding and linking relevant work in a particular field a cumbersome task.

On the desktop, we can find a somewhat similar problem, though on a smaller scale. A typical researcher acquires (and 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

[1] http://medline.cos.com/

A. Bernstein et al. (Eds.): ISWC 2009, LNCS 5823, pp. 827–842, 2009.

them), and may, in the best case scenario, be structured in intuitive folder hierarchies. Thus, finding a particular publication or links between the existing ones represents quite a challenge, even with the help of tools like Google Desktop[2].

Semantic Web technologies have been proved to help at alleviating, at least partially, the above mentioned issues. And at the foundation of the Semantic Web we find semantic metadata. Used in particular contexts, semantic metadata enables a more fertile search experience, complementing full text search with search based on different facets (e.g., one can search for a publication by a specific author and with some specific keywords in its title). In addition, subject to its richness, it can also leverage links between publications, e.g. citation networks.

Looking at the status of semantic metadata for scientific publications in the two directions, i.e. the Web and the Desktop, we observe the following. With the emergence of the Linked Open Data (LOD)[3] initiative, an increasing number of data sets were published as linked metadata. Regarding scientific publications, efforts like the Semantic Web Dog Food Server started by Möller et al. [2] represent pioneering examples. The repository they initiated acts as a linked data hub, for metadata extracted from different sources, such as the International or European Semantic Web conferences, and now hosts metadata describing over 1000 publications and over 2800 people. The manual creation of metadata is their main drawback, as well as of the other similar approaches. Within the second direction, i.e. on the desktop, different Semantic Desktop efforts improve the situation, by extracting shallow metadata, either file-related (e.g. creator, date of creation), or even publication-related, such as title or authors. In conclusion, we currently have two directions targeting similar goals and having the same foundation: (i) the <LOD — semantic metadata> bridge, linking publications on the web, and (ii) the <Semantic Desktop — semantic metadata> bridge, linking publications and personal information, on the desktop.

In this paper, 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 the primary source. Our method consists of a three step process and starts from a publication with no metadata, each step carried out incrementally to enrich the semantic metadata describing the publication. The three steps are: (i) *extraction* – we extract automatically metadata from the publication based on a document-engineering oriented approach; (ii) *expansion* – we use the extracted raw metadata to search the linked data cloud, the result being a set of clean and linked metadata; and (iii) *integration* – the metadata is further enriched by embedding it within the semantic desktop environment, where it is automatically linked with the already existing personal metadata. The main results of our process are: a simple and straightforward way of finding related publications based on the metadata extracted and linked automatically, and the opportunity of weaving the linked publication data on the desktop, by means of usual desktop applications (e.g. file and Web browser).

[2] http://desktop.google.com/
[3] http://linkeddata.org/

The remainder of the paper is structured as follows: Sect. 2 introduces the scenario used for exemplifying our approach, while in Sect. 3 we detail the technical elements of each of the three steps in the process. Sect. 4 describes the preliminary evaluation we have performed, and before concluding in Sect. 6, we have a look at related efforts in Sect. 5.

2 Scenario

To illustrate the problem mentioned in the previous section in a particular context, in the following, we will consider a typical scenario for an early stage researcher (or any kind of researcher) that steps into a new field. The amount of existing publications, and their current growth rate, makes the task of getting familiarized with the relevant literature in a specific domain highly challenging. From an abstract perspective, the familiarization process consists of several stages, as follows: (i) the researcher starts from a publication provided by the supervisor; (ii) she reads the publication, thus grasping its claims and associated argumentation; (iii) reaching a decision point, she either decides to look for another publication, or she follows backwards the chain of references, possibly including also publications by the same authors. This last step usually involves accessing a search engine (or a publication repository) and typing the names of the authors, or the title of the publication, to be able to retrieve similar results.

Each of the above activities has an associated corresponding time component: (i) the time assigned to reading the entire publication (or most of it) — needed to decide whether the publication is of interest or not, (ii) the time associated with finding appropriate links and references between publications,[4] and (iii) the time associated with searching additional publications, based on different metadata elements, and (manually) filtering the search results. This time increases substantially when an individual is interested in all the publications of a particular author.

Finally, analyzing the pairs <activity, time component> from the metadata perspective, and what it can do to improve the overall process, we can conclude that:

- searching and linking related publications as mentioned above, entails the (manual) extraction and use of shallow metadata. Thus, performing automatic extraction of shallow metadata and using it within the linked data cloud, will significantly reduce the time component associated with this activity.
- reading the publication to a large extent corresponds to mining for the discourse knowledge items, i.e. for the rhetorical and argumentation elements of the publication, representing its deep metadata. Consequently, extracting automatically such discourse knowledge items from a publication, will provide the user with the opportunity of having a quick glance over the publication's main claims and arguments and thus decrease the time spent on deciding whether the publication is relevant or not.

[4] Following all the references of a publication is obviously not a feasible option. Thus, the decision is usually done based on the citation contexts mentioning the references in the originating publication.

Transposing these elements into engineering goals led us to a three step process, detailed in the following section: *extraction* – automatic extraction of shallow and deep metadata; *expansion* – using the extracted metadata within the linked data cloud for cleaning and enriching purposes; *integration* – embedding the resulted linked metadata within the personal desktop to ensure a smooth search and browse experience, by using the ordinary desktop applications.

As a side remark to the proposed scenario, an interesting parallel can be made with the music domain. Similarly to publications, music items (e.g. music files, tracks, etc) are also acquired and stored by people on their personal desktops, in numbers usually increasing with time. And as well as publications, these can embed (depending on the format) shallow metadata describing them, such as, band, song title, album or genre. Thus, conceptually, the *extraction — expansion — integration* process we propose can be applied also in this domain. In practice, there already exist tools that deal with parts of this process. For example, on the extraction side, there are tools that help users to create or extract ID3 tags embedded into MP3 files or on the expansion side, there exist tools, such as Picard,[5] that clean the metadata based on specialized music metadata repositories (e.g. MusicBrainz). As we shall see in the next section, the result of our work is quite similar to these, but applied on scientific publications.

3 Implementation

One of our main goals was reducing as much as possible 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 at the same time, we targeted an increase of 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. The first step, *extraction*, 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

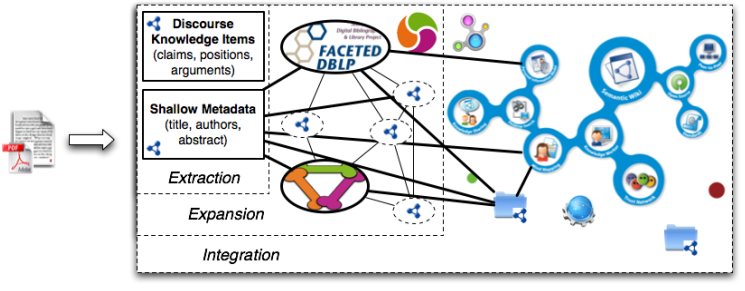

Fig. 1. Incremental metadata enrichment process

[5] http://musicbrainz.org/doc/PicardTagger

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. Since the extraction process is based on a hybrid 'document engineering – computational linguistic' approach, the resulting metadata may contain errors. These errors can be corrected in the *expansion* step, in addition to enriching the basic set of metadata with linked data, coming from different sources. As we shall see, we opted for a clear distinction of the semantics of the `owl:sameAs` and `rdfs:seeAlso` relations. 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.

In terms of implementation, the first two steps are developed as part of a standalone application[6]. The extraction currently targets publications encoded as PDF documents and preferably using the ACM and LNCS styles, while the expansion is achieved via the Semantic Web Dog Food Server and the Faceted DBLP[7] linked data repositories. The integration of the metadata is done using the services provided by the KDE NEPOMUK Server,[8] 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). The application 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.

To have a better understanding of the result of each step, we will use as a running example throughout the following sections, the metadata extracted from a publication entitled *Recipes for Semantic Web Dog Food – The ESWC and ISWC Metadata Projects*.[9] More precisely, we will assume that a user would start her quest from this publication, and show the incremental effect of using our application on the created metadata.

3.1 Extraction

The extraction of shallow metadata was developed as a set of algorithms that follow a low-level document engineering approach, by combining mining and analysis of the publication's text based on its formatting style and font information. The algorithms currently work only on PDF documents, with a preference for the ones formatted with the LNCS and ACM styles. Each algorithm in the set deals with one aspect of the shallow metadata. Thus, there are individual algorithms for extracting the title, authors, references and the linear structure.

A complete description of the algorithms can be found in [3]. Nevertheless, to provide the basic idea of how they work, we will describe shortly the authors extraction algorithm. There are four main processing steps: (i) We first merge the

[6] Demo at http://sclippy.semanticauthoring.org/movie/sclippy.htm
[7] http://dblp.l3s.de/
[8] http://nepomuk.kde.org/
[9] http://iswc2007.semanticweb.org/papers/795.eps

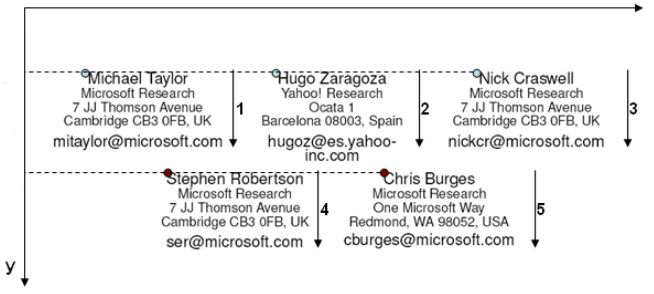

Fig. 2. Authors extraction algorithm example

consecutive text chunks on the first page that have the same font information and are on the same line (i.e. the Y coordinate is the same); (ii) then, we select the text chunks between the title and the abstract and consider them author candidates; (iii) the next step is the linearization of the author candidates based on the variations of the Y axis; (iv) finally, we split the author candidates based on the variations of the X axis.

Fig. 2 depicts an example of a publication that has the authors structured on several columns. The figure shows the way in which the authors' columns containing the names and affiliations are linearized, based on the variation of the Y coordinate. The arrows in the figure show the exact linearization order. The variations on the X axis can be represented in a similar manner.

The extraction of deep metadata, i.e. discourse knowledge items (claims, positions, arguments), was performed based on a completely different approach. Having as foundational background the Rhetorical Structure of Text Theory (RST) [4], we have developed a linguistic parser that mines the presence of rhetorical relations within the publication's content. In order to automatically identify text spans and the rhetorical relations that hold among them, we relied on the discourse function of cue phrases, i.e. words such as *however*, *although* and *but*. An exploratory study of such cue phrases provided us with an empirical grounding for the development of an extraction algorithm. The next phase consisted of an experiment for determining the initial probabilities for text spans to represent knowledge items, based on the participation in a rhetorical relation of a certain type and its block placement in the publication (i.e. abstract, introduction, conclusion or related work). The parser was implemented as a GATE[10] plugin. Detailing the actual extraction mechanism is out of the scope of this paper, as our focus is on the incremental process that realizes the bridging the Linked Web of Data and the Semantic Desktop. Nevertheless, it is worth mentioning that we do extract also deep metadata, as it brings added value to the user, and as we shall see later in this section, enables meaningful queries in the bigger context of the full set of extracted metadata.

As mentioned, the first two steps of our process are implemented as a standalone application. The left side of Fig. 3 depicts the main interface of this

[10] http://gate.ac.uk/

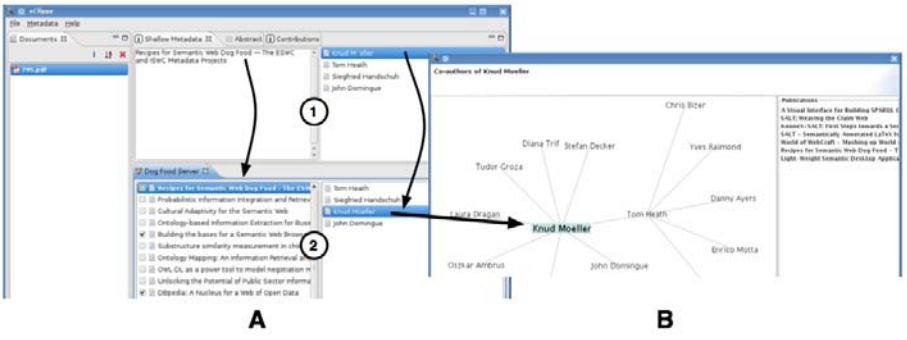

Fig. 3. Screenshot of the application's interface: [A] – The main window; [B] – Co-authors graph visualization

application, while with <1> we indicated the place where the result of the *extraction* is displayed. At the same time, the listing below summarizes elements of the metadata extracted after this first step, i.e. title, authors, the text of the abstract, and a list of claims (i.e. the most important contributions statements of the publication). For shaping the metadata, we used a mixture of ontologies and vocabularies, such as SALT (Semantically Annotated LaTeX) framework [5], DublinCore and FOAF. One particular element that should be noted here, is that this metadata may contain errors. As it can be seen in the listing below the name of the first author is incorrect: *Mo"ller* instead of *Möller*. The user has the chance to correct such mistakes manually, or advance to the next step, where the correction can be done automatically — if the publication under scrutiny is found in the linked data repository. In any case, already at this point, the user can decide to jump to the *integration* step, simply just export this metadata as an individual file, or embed it directly into the originating PDF. From the scenario's point-of-view, the researcher already gains value, as she can quickly grasp the main claims of the publication, by inspecting the extracted deep metadata.

```
<pub> a sdo:Publication .                          <knud> foaf:name ''Knud Mo"ller'' .
<pub> dc:title ''Recipes for Semantic Web ...'' . <abs> a sro:Abstract .
<pub> dc:creator knud .                            <abs> konnex:hasText ''Semantic Web ...'' .
<pub> dc:creator tom .                             <pub> dcterms:abstract abs .
<pub> dc:creator siegfried .                       <claim> a sro:Claim .
...                                                <claim> konnex:hasText ''This paper ...'' .
<knud> a foaf:Person .                             <pub> konnex:hasClaim claim .
```

3.2 Expansion

The *expansion* step takes the metadata extracted previously and, under the user's guidance, corrects existing errors and enriches it, by using Linked Data repositories. We have currently implemented expansion modules for the Semantic Web

Dog Food Sever and Faceted DBLP. The actual expansion is done based on the extracted title and authors. On demand, these are individually used for querying the SPARQL endpoints of the Linked Data repositories. As a prerequisite step, both the title and the authors (one-by-one) are cleaned of any non-letter characters, and transformed into regular expressions. The title is also chunked into multiple variations based on the detected nouns, while each author name is chunked based on the individual parts of the full name, discarding the parts that are just one letter long. Consequently, each element will have an associated array of sub-strings used for querying.

In the case of the title, the query result will be a list of resources that may contain duplicates, and among which there might also be the publication given as input. In order to detect this particular publication, we perform a shallow entity identification. First, to mask possibly existing discrepancies in the title, we use string similarity measures. An empirical analysis led us to using a combination of the Monge-Elkan and Soundex algorithms, with fixed thresholds. The first one analyzes fine-grained sub-string details, while the second looks at coarse-grained phonetic aspects. The titles that pass the imposed thresholds (0.75 and 0.9) advance to the next step. Secondly, we consider the initially extracted authors and compare them with the ones associated with the publications that pass over the above mentioned thresholds. The comparison is done using the same similarity measures, but with different thresholds (0.85 and 0.95). The publications satisfying both conditions have their models retrieved and presented to the user as candidates. A similar approach is also followed on the authors' side.

The outcome of the expansion features three elements: (i) a list of candidates, to be used for cleaning and linking the initially extracted metadata (with their linked model and authors' models), (ii) a list of similar publications, mainly the ones that did not satisfy the two conditions of the shallow entity resolution (again with their linked model and authors' models), 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. From the scenario perspective, this outcome provides the researcher with the chance of analyzing both publications that might have similar approaches and inspect all the publications of a particular author.

At this stage, there are three options that can be followed. The first option is to use the best retrieved candidate to correct and link the initial metadata. Both the publication and the authors will inherit the discovered `owl:sameAs` links, that will later provide the opportunity to browse different instances of the same entity in different environments. The second option is to link other publications that she considers relevant to the one under scrutiny. While at the interface level this is done based on the user's selection (see pointer 2 in Fig. 3), at the model level we use the `rdfs:seeAlso` relation. We thus make a clear distinction in semantics between `owl:sameAs` and `rdfs:seeAlso`. The former represents a strong link between different representations of the same entity, while the latter acts as a weak informative link, that will later help the user in re-discovering similarities between several publications. The third and last option is specific for authors, and allows the user to navigate through the co-authors networks of a particular author (part

B of Fig. 3). 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.

Returning to our running example, the output of this step is an added set of metadata, presented briefly in the listing below. Thus, in addition to the already existing metadata, we can now find the above mentioned `owl:sameAs` and `rdfs:seeAlso` relations, and the incorrectly extracted name *Mo"ller*, now corrected to *Moeller*, based on the `foaf:name` found in the linked data.

```
<knud> foaf:name ''Knud Moeller'' .
<knud> owl:sameAs http://data.semanticweb.org/person/knud-moeller .
<knud> owl:sameAs http://dblp.l3s.de/d2r/resource/authors/Knud_M\%C3\%B6ller .
...
<pub> owl:sameAs http://data.semanticweb.org/conference/iswc-aswc/2007/.../papers/795 .
<pub> owl:sameAs http://dblp.l3s.de/d2r/resource/publications/conf/semweb/MollerHHD07 .
<pub> rdfs:seeAlso <pub2> .
...
<pub2> a sdo:Publication .
<pub2> dc:title ''DBPedia: A Nucleus for a Web of Open Data ... '' .
<pub2> dc:creator <richard> .
<pub2> dc:creator <georgi> .
<pub2> owl:sameAs http://dblp.l3s.de/d2r/resource/publications/conf/semweb/AuerBKLCI07 .
<pub2> owl:sameAs http://data.semanticweb.org/conference/iswc-aswc/2007/.../papers/715 .
```

3.3 Integration

The last step of our process is the *integration*, which embeds the extracted and linked metadata into the personal information space, managed by the Semantic Desktop, and thus realizing the actual bridge between the Linked Data and the Semantic Desktop. To achieve this, we have used the NEPOMUK–KDE implementation of the Semantic Desktop. This provides a central local repository for storing structured data and it is well integrated with the common desktop applications, such as the file and Web browsers. In terms of foundational models, it uses the NEPOMUK Ontologies[11] suite. In our case, the actual integration was done at the metadata level, where we had to align the instances previously extracted with the ones already existing in the local repository.

Currently, we deal with two types of alignments: person alignment and publication alignment, that are described within the Semantic Desktop context by means of the NCO (NEPOMUK Contact Ontology), NFO (NEPOMUK File Ontology) and PIMO (Personal Information Model Ontology) ontologies.

The person alignment resumes to checking whether an extracted author is already present in the personal information model, and in a positive case, merging the two models in an appropriate manner. This is done based on a two step mechanism, similar to finding authors in a Linked Data repository. We first query the local repository for the name of the author and the associated substrings resulted

[11] http://www.semanticdesktop.org/ontologies/

from chunking the name into several parts. Using the same similarity measures, we filter out only the realistic candidates. These candidates are then analyzed based on the existing `owl:sameAs` links and their linked publications. If a candidate is found to have one identical `owl:sameAs` link and one identical publication with the initial author, we consider it a match and perform the merging of the two models. In a negative case, the author's model is stored as it is and we advance to the publication alignment. The result of this alignment is exemplified in the listing below. In addition to the already existing metadata, the author now has attached an email address and the birth date, both found within the user's personal information space.

The publication alignment is straightforward, as from a local and physical perspective, the publication is represented by a file. Thus, considering that the user started the process from such a file (which is always the case), we query the repository for the information element corresponding to that file, having the `fileUrl` (or path) as the main indicator. The conceptual model found to be associated with the file is then merged with the extracted publication model. The listing below shows this alignment as part of our running example, the last statement creating the actual grounding of the publication onto a physical file.

The *integration* enables, in particular, two important elements: (i) firstly, more meaningful ways of finding and linking publications on the desktop, and (ii) secondly, an opportunity of weaving the linked data present on the desktop, using ordinary desktop applications. Fig. 4 depicts the first aspect, using Dolphin, the

```
<knud> nco:birthDate ''1980-11-01'' .
<knud> nco:emailAddress knud.moeller@deri.org .
...
<pubFile> a nfo:FileDataObject .
<pubFile> nfo:fileSize 1353543 .
<pubFile> nfo:fileUrl file:///home/user/research/papers/p215.eps .
<pub> pimo:groundingOccurence <pubFile> .
```

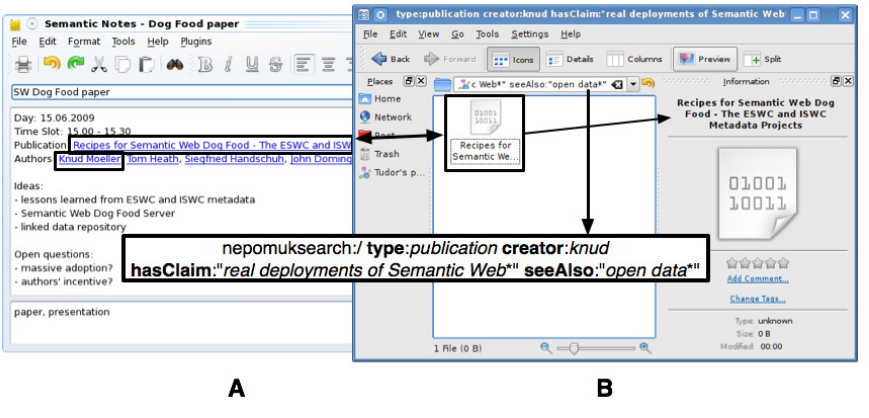

Fig. 4. Deep metadata integration in KDE applications: [A] SemNotes; [B] Dolphin

KDE file browser (part B), and SemNotes,[12] a KDE semantic note-taking application (part A). As shown in the figure, to retrieve all `publications` having `knud` among the authors, `claim`-ing that they deal with *"real deployments of Semantic Web..."* and being related to (`seeAlso`) publications that speak about *"open data"*, resolves to using the following query in Dolphin:

```
nepomuksearch:/ type:publication creator:knud hasClaim:''real
    deployments of Semantic Web*'' seeAlso:''open data*''
```

The result of the query will be a virtual folder that is automatically updated in time (enabling the long-term effect), thus showing also the publications that are added at a later stage and that satisfy the query, independently of the name of the physical file or its location. Similarly, while taking notes during the presentation of this particular paper, SemNotes will automatically link both the publication and the author mentioned in the note, therefore providing added information about the publication and its authors.

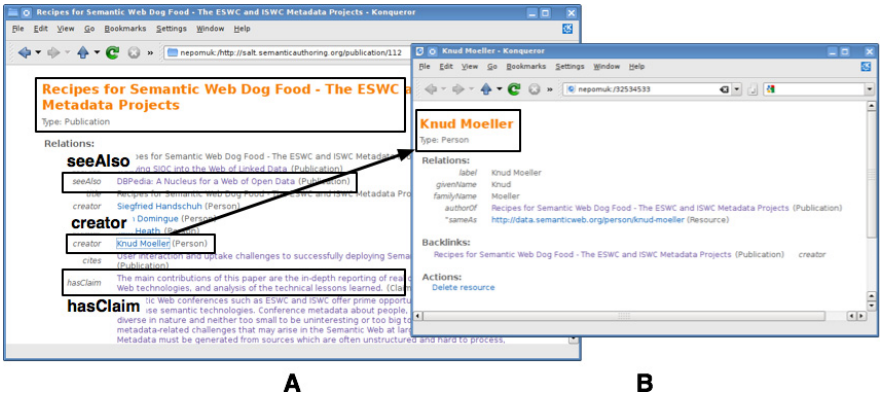

Fig. 5. Browsing deep integrated resources with Konqueror

The second aspect is presented in Fig. 5, which shows how resources such as publications or authors can be visualized by means of an ordinary Web browser (here, Konqueror). More important, this enables the visualization of the rich information space surrounding a publication, both from a local and linked data perspective. Without even opening the actual publications, the user can: (i) quickly grasp the main ideas of the publications, via the presented claims, (ii) see related publications, via the `rdfs:seeAlso` links, or (iii) inspect the publications' authors, either via their personal contact information, or via their different instances on the Web (`owl:sameAs` links). We believe that this approach combines harmoniously the Linked Data perspective with the Semantic Desktop perspective, thus enabling the weaving of Linked Data on the desktop.

[12] http://smile.deri.ie/projects/semn

4 Preliminary Evaluation

We evaluated the extraction of semantic metadata and performed a short-term usability study of the overall approach. The shallow metadata extraction achieved high accuracies for the title (95%) and abstract (96%) extraction, and a lower accuracy for authors extraction (90%). The evaluation method and complete results can be found in [3]. In this section, we focus on the usability study, as we believe that the developed application has to be easy to learn and use, and to provide the most appropriate information.

The study was conducted together with 16 evaluators, a mixture of PhD students and Post Doctorands from our institute, that were asked to perform a series of tasks covering all the application's functionalities. Example of such tasks included: extraction and manual correction of metadata from publications, expansion of information based on the same publications or exploration of the co-authors graph. At the end, the evaluators filled in a questionnaire, comprising of 18 questions, with Likert scale-based or free form answers, concentrating on two main aspects: (i) suitability and ease of use, and (ii) design, layout and conformity to expectancies. The complete results of the questionnaire can be found at http://smile.deri.ie/sclippy-usabilitystudy

Overall, the application scored very well in both categories we have targeted. The vast majority of the evaluators (on average more than 90%) found the tool well suited for the extraction and exploration of shallow and deep metadata. The same result was achieved also for the exploration of the information space surrounding the chosen publication, based on the extracted and linked metadata. In addition, the information presented by the application, both for publications and authors, was found helpful (100% for publications and 72.8% for authors), while 93.8% of the evaluators found an added value in our tool when compared to the original expansion environment.

In the other category, all evaluators considered the application easy to learn and use (100%) while having the design and layout both appealing (87.5%) and suited for the task (93.6%). Issues were discovered in two cases: (i) the self-descriptiveness of the application's interface (only 68.8% found it self-descriptive), mainly due to the lack of visual indicators and tooltips, and (ii) the suggested list of similar publications (again only 68.8% found it relevant). Although the application always found the exact publication selected for expansion in the repository, the proposed list of similar publications created some confusion.

Apart from these findings, directly taken from the questionnaires, we observed that even without any training and documentation, the evaluators experienced a very smooth learning curve. Additionally, most of them enjoyed our exercise, while some were interested in using the application on a daily basis. On the other hand, the study pointed out a number of issues and led us to a series of directions for improvement. First of all, the need to make use of a more complex mechanism for suggesting similar publications. As we expected, the shallow similarity-based heuristics we used for building the list of suggested publications left plenty of space for improvement. Unfortunately, its improvement is directly dependent on the quantity and quality of information provided by the linked data repository. As

an example, while we could use the abstract provided by the Semantic Web Dog Food Server to extract discourse knowledge items, and then perform similarity measures at this level, this would not be possible when using the Faceted DBLP, where such information does not exist. For this case, a possible solution, would be to drill deeper into the linked web of data. Secondly, augmenting the expanded information with additional elements (e.g. abstract, references, citation contexts), thus providing a deeper insight into the publications and a richer experience for the users.

5 Related Work

To our knowledge, until now, there was no attempt to combine in such a direct manner automatic metadata extraction from scientific publications, linked open data and the semantic desktop. Nevertheless, there are efforts that deal with parts of our overall approach, and, in this section, we will focus on them. Hence, we will cover: (i) automatic extraction of shallow metadata, including the context of the semantic desktop, and (ii) information visualization for scientific publications.

Before detailing the two above-mentioned directions, we would like to discuss the position of the alignments described in the *expansion* and *integration* steps to the current state of the art. To a certain extent, these person and publication alignments are similar to performing coreference resolution. While in the person case the resolution is solved directly via string similarity measures, in the publication case we add the authors list as an extra condition. This makes our approach more simple and straightforward than the more accurate algorithms existing in the literature. Examples of such techniques include: naive Bayes probability models and Support Vector Machines [6], K-means clustering [7] or complex coreference based on conditionally trained uni-directed graph models using attributes [8].

Extensive research has been performed in the area of the Semantic Desktop, with a high emphasis on integration aspects within personal information management. Systems like IRIS [9] or Haystack [10] deal with bridging the different isolated data silos existing on the desktop, by means of semantic metadata. They extract shallow metadata from the desktop files and integrate it into a central desktop repository. Compared to our approach, the metadata extraction is file-oriented and shallow, whereas we extract specific publication metadata and integrate it within the already existing semantic desktop data. The closest effort to ours was the one of Brunkhorst et al. [11]. In their Beagle++ search engine, developed in the larger context of the NEPOMUK Semantic Desktop [12], the authors also perform metadata extraction from scientific publications, but limited to title and authors.

Regarding the general context of automatic extraction of metadata from publications, there have been several methods used, like regular expressions, rule-based parsers or machine learning. Regular expressions and rule-based systems have the advantage that they do not require any training and are straightforward to implement. Successful work has been reported in this direction, with emphasis on PostScript documents in [13], or considering HTML documents and use of natural language processing methods in [14]. Our approach is directly comparable

with these, even though the target document format is different. In terms of accuracy, we surpass them with around 5% on title and authors extraction, and with around 15% on linear structure extraction, while providing additional metadata (i.e. abstract or references).

Although more expensive, due to the need of training data, machine learning methods are more efficient. Hidden Markov models (HMMs) are the most widely used among these techniques. However, HMMs are based on the assumption that features of the model they represent are not independent from each other. Thus, HMMs have difficulty exploiting regularities of a semi-structured real system. Maximum entropy based Markov models [15] and conditional random fields [16] have been introduced to deal with the problem of independent features. In the same category, but following a different approach, is the work performed by Han et al. [17], who uses Support Vector Machines (SVMs) for metadata extraction.

With respect to information visualization of scientific publications, a number of methods and tools have been reported in the literature. The 2004 InfoVis challenge had motivated the introduction of a number of visualization tools highlighting different aspects of a selected set of publications in the Information Visualization domain. Faisal et. al. [18] reported on using the InfoVis 2004 contest dataset to visualize citation networks via multiple coordinated views. Unlike our work, these tools were based on the contents of a single file, which contained manually extracted and cleaned metadata. As noted by the challenge chairs, it was a difficult task to produce the metadata file [19] and hence the considerable efforts required made it challenging for wide-spread use. In [20], a small scale research management tool was built to help visualizing various relationships between lab members and their respective publications. A co-authorship network visualization was built from data entered by users in which nodes represented researchers together with their publications, and links showed their collaborations. A similar effort to visual domain knowledge was reported by [21], with their data source being bibliographic files obtained from distinguished researchers in the "network science" area. While this work was also concerned with cleansing data from noisy sources, the metadata in use was not extracted from publications themselves and no further information available from external sources such as Faceted DBLP was utilized. Another tool targeting the exploration of the co-authorship network is CiteSpace [22]. CiteSpace tries to identify trends or salient patterns in scientific publications. The source of information for CiteSpace is also from bibliographic records crawled from different publishers on the web, rather than extracted metadata.

6 Conclusion and Future Developments

In this paper we presented an approach for dealing, at least to some extent, with the information overload issue both on the Web and on the Desktop, and having as target early stage researchers. Our solution, inspired from the typical process of getting familiarized with a particular domain, combines elements from the Linked Web of Data and the Semantic Desktop, using semantic metadata as a common

denominator. The result consists of three steps (*extraction – expansion – integration*) that incrementally enrich the semantic metadata describing a publication, from no metadata to a comprehensive model, linked and embedded within the personal information space.

Each step has associated a series of open challenges that we intend to address as part of our future work. As currently the *extraction* works only on publications published as PDF documents, and formatted preferably with the LNCS and ACM styles, we plan to improve extraction algorithms to accommodate any formatting style, as well as develop new extraction modules for other document formats, such as Open Document formats. At the moment, the *expansion* uses only two Linked Data repositories, i.e. the Semantic Web Dog Food Server and the Faceted DBLP. Future developments will include also other repositories, in addition to means for 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, and ultimately a complete automatization of the process, thus reducing the overhead to the minimum possible.

Acknowledgments

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

References

1. Tsujii, J.: Refine and pathtext, which combines text mining with pathways. In: Keynote at Semantic Enrichment of the Scientific Literature 2009, Cambridge, UK (2009)
2. Möller, K., Heath, T., Handschuh, S., Domingue, J.: Recipes for Semantic Web Dog Food — The ESWC and ISWC Metadata Projects. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 802–815. Springer, Heidelberg (2007)
3. Groza, T., Handschuh, S., Hulpus, I.: A document engineering approach to automatic extraction of shallow metadata from scientific publications. Technical Report 2009–06–01, Digital Enterprise Research Institute (2009)
4. Mann, W.C., Thompson, S.A.: Rhetorical Structure Theory: A theory of text organization. Technical Report RS-87-190, Information Science Institute (1987)
5. Groza, T., Handschuh, S., Möller, K., Decker, S.: SALT - Semantically Annotated LaTeX for Scientific Publications. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 518–532. Springer, Heidelberg (2007)
6. Han, H., Zha, H., Giles, C.: A Model-based K-means Algorithm for Name Disambiguation. In: Proc. of the Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data @ ISWC 2003, Sanibel Island, Florida, USA (2003)
7. Han, H., Giles, L., Zha, H., Li, C., Tsioutsiouliklis, K.: Two supervised learning approaches for name disambiguation in author citations. In: Proc. of the 4th ACM/IEEE-CS joint conference on Digital Libraries, Tuscon, AZ, USA (2004)

8. Wellner, B., McCallum, A., Peng, F., Hay, M.: An integrated, conditional model of information extraction and coreference with application to citation matching. In: Proc. of the 20th Conf. on Uncertainty in Artificial Intelligence, Banff, Canada (2004)
9. Cheyer, A., Park, J., Giuli, R.: IRIS: Integrate. Relate. Infer. Share. In: Proc. of the Semantic Desktop and Social Semantic Collaboration Workshop (2006)
10. Quan, D., Huynh, D., Karger, D.R.: Haystack: A Platform for Authoring End User Semantic Web Applications. In: Proc. of the 2nd International Semantic Web Conference, pp. 738–753 (2003)
11. Brunkhorst, I., Chirita, P.A., Costache, S., Gaugaz, J., Ioannou, E., Iofciu, T., Minack, E., Nejdl, W., Paiu, R.: The beagle++ toolbox: Towards an extendable desktop search architecture. Technical report, L3S Research Centre, Hannover, Germany (2006)
12. Bernardi, A., Decker, S., van Elst, L., Grimnes, G., Groza, T., Handschuh, S., Jazayeri, M., Mesnage, C., Möller, K., Reif, G., Sintek, M.: The Social Semantic Desktop: A New Paradigm Towards Deploying the Semantic Web on the Desktop. In: Semantic Web Engineering in the Knowledge Society. IGI Global (2008)
13. Shek, E.C., Yang, J.: Knowledge-Based Metadata Extraction from PostScript Files. In: Proc. of the 5th ACM Conf. on Digital Libraries, pp. 77–84 (2000)
14. Yilmazel, O., Finneran, C.M., Liddy, E.D.: Metaextract: an nlp system to automatically assign metadata. In: JCDL 2004: Proc. of the 4th ACM/IEEE-CS joint conference on Digital libraries, pp. 241–242 (2004)
15. McCallum, A., Freitag, D., Pereira, F.: Maximum entropy markov models for information extraction and segmentation. In: Proc. of the 17th Int. Conf. on Machine Learning, pp. 591–598 (2000)
16. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. of the 18th Int. Conf. on Machine Learning, pp. 282–289 (2001)
17. Han, H., Giles, C.L., Manavoglu, E., Zha, H., Zhang, Z., Fox, E.A.: Automatic document metadata extraction using support vector machines. In: Proc. of the 3rd ACM/IEEE-CS Joint Conf. on Digital libraries, pp. 37–48 (2003)
18. Faisal, S., Cairns, P.A., Blandford, A.: Building for Users not for Experts: Designing a Visualization of the Literature Domain. In: Information Visualisation 2007, pp. 707–712. IEEE Computer Society Press, Los Alamitos (2007)
19. Plaisant, C., Fekete, J.-D., Grinstein, G.: Promoting Insight-Based Evaluation of Visualizations: From Contest to Benchmark Repository. IEEE Transactions on Visualization and Computer Graphics 14(1), 120–134 (2008)
20. Neirynck, T., Borner, K.: Representing, analyzing, and visualizing scholarly data in support of research management. In: Information Visualisation 2007, pp. 124–129. IEEE Computer Society Press, Los Alamitos (2007)
21. Murray, C., Ke, W., Borner, K.: Mapping scientific disciplines and author expertise based on personal bibliography files. In: Information Visualisation 2006, pp. 258–263. IEEE Computer Society Press, Los Alamitos (2006)
22. Chen, C.: CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. J. of the American Society for Information Science and Technology 57(3), 359–377 (2006)