The Personal Publication Reader

Fabian Abel[1], Robert Baumgartner[2,3], Adrian Brooks[3], Christian Enzi[2],
Georg Gottlob[2,3], Nicola Henze[1], Marcus Herzog[2,3], Matthias Kriesell[4],
Wolfgang Nejdl[1], and Kai Tomaschewski[1]

[1] Research Center L3S & Information Systems Institute, University of Hannover
{abel, henze, nejdl, tomaschewski}@kbs.uni-hannover.de
[2] DBAI, Institute of Information Systems, Vienna University of Technology
{baumgart, enzi, gottlob, herzog}@dbai.tuwien.ac.at
[3] Lixto Software GmbH, Donau-City-Strasse 1/Gate 1, 1220 Vienna, Austria
{baumgartner, brooks, gottlob, herzog}@lixto.com
[4] Inst. f. Math. (A), University of Hannover
kriesell@math.uni-hannover.de

Abstract. This application demonstrates how to provide personalized,
syndicated views on distributed web data using Semantic Web technolo-
gies. The application comprises four steps: The **information gather-
ing step**, in which information from distributed, heterogenous sources
is extracted and enriched with machine-readable semantics, the **oper-
ation step** for timely and up-to-date extractions, the **reasoning step**
in which rules reason about the created semantic descriptions and addi-
tional knowledge-bases like ontologies and user profile information, and
the **user interface creation step** in which the RDF-descriptions result-
ing from the reasoning step are interpreted and translated into an appro-
priate, personalized user interface. We have developed this application for
solving the following real-world problem: We provide personalized, syn-
dicated views on the publications of a large European research project
with more than twenty geographically distributed partners and embed
this information with contextual information on the project, its working
groups, information about the authors, related publications, etc.[1]

Keywords: web data extraction, web data syndication, personalized
views.

1 Introduction

In today's information society, the World Wide Web plays a prominent role for
disseminating and retrieving information: lots of useful information can be found
in the web, from train departure tables to consultation hours, from scientific
data to online auctions, and so on. While this information is already available
for consumption by human users, we lack applications that can collect, evaluate,

---

combine, and re-evaluate this information. Currently, users retrieve online content in separate steps, one step for each information request, and evaluate the information chunks afterwards according to their needs: e.g. the user compares the train arrival time with the starting time of the meeting he is requested to participate in, etc. Another common scenario for researchers is that a user reads some scientific publication, gets curious about the authors, other work of the authors, on related work targeting on similar research questions, etc. Linking these information chunks together is a task that can currently not be performed by machines. In our application, we show how to solve this information integration problem for the latter mentioned "researcher scenario". We show, how to

1. extract information from distributed and inhomogeneous sites, and create semantic descriptions of the extracted information chunks,
2. maintain the web data extraction to ensure up-to-date information,
3. reason about the created semantic descriptions and ontological knowledge,
4. and create syndicated, personalized views on web information.

The Personal Publication Reader (PPR) extends the idea of Semantic Portals like e.g. SEAL [4] or others with the capability of extracting and syndicating web data from various, distributed sites or portals which do not belong to the ownership of the application itself.

2   Extraction and Annotation with Semantic Descriptions

In our application, the web pages from which we extract the information are maintained by partners of the research project REWERSE, thus the sources of the information are distributed and belong to different owners which provide their information in various ways and formats (HTML, Java-script, PHP-generated pages, etc.). Moreover, in each list, authors, titles and other entities are potentially characterized in a different way, and different order criteria are enforced (e.g. by year or by name). Such a web presentation is well suited for human consumption, but hardly usable for automatic processing. Nevertheless, the web is the most valuable information resource in this scenario. In order to access and understand these heterogeneous information sources one has to apply web extraction techniques. The idea of our application is to "wrap" these heterogeneous sources into a formal representation based on Semantic Web standards. In this way, each institution can still maintain their own publication list and at the same way we can offer an integrated and personalized view on this data by regularly extracting web data from all member sites.

This application is open in the sense that it can be extended in an easy way, i.e. by connecting additional web sources. For instance, abstracts from `www.researchindex.com` can be queried for each publication lacking this information and joined to each entry. Moreover, using text categorization tools one can rate and classify the contents of the abstracts. Another possibility is to extract organization and person data from the institution's web pages to inform the ontology to which class in the taxonomy an author belongs (such as full

professor). Web extraction and annotation in the PPR is performed by the *Lixto Suite*. Web data extraction is a hot topic – for an extensive overview of methods and tools refer to [3]. First, with the *Lixto Visual Wrapper* [1] for each type of web site a so-called wrapper is created; the application designer visually and semi-automatically defines the characteristics of publication elements on particular web sites based on characteristics of the particular HTML presentation and some possible domain knowledge. After a wrapper has been generated it can be applied to a given web site (e.g. publications of University of Munich) to generate an "XML companion" that contains the relevant information stored in XML using (in this application context meaningful) XML tags.

3   Extraction Maintenance

In the next step, in the *Lixto Transformation Server* application designer visually composes the information flow from web sources to an RDF presentation that is handed over to the PPR once a week. Then the application designer defines a schedule how often which web source is queried and how often the information flow is executed. Additionally, deep web navigation macros possibly containing logins, cookies and web forms as well as iteration over forms are created. As a next step in the data flow, the data is harmonized to fit into a common structure, and e.g. an attribute "origin" is added containing the institution's name, and author names are harmonized by being mapped to a list of names known by the system. Finally, the XML data structure is mapped to a pre-defined RDF schema structure. Once the wrappers are in place, the complete application runs without further human interference, and takes care of publication updates. In case future extractions fail the application designers will receive a notification.

4   Reasoning for Syndicated and Personalized Views on Distributed Web Data

In addition to the extracted dynamic information, we maintain data about the members of the research project from the member's corner of the REWERSE project web site. We have constructed an ontology for describing researchers and their involvement in scientific projects like REWERSE, which extends the known Semantic Web Research Community Ontology (`http://ontobroker.semanticweb.org/ontos/swrc.html`) with some project-specific aspects.

Personalization rules reason about all this dynamic and static data in order to create syndicated and personalized views. As an example, the following rule (using the TRIPLE[5] syntax) determines all authors of a publication:

```
FORALL A, P authors(A, P) <- P[dc:creator -> A]@'http:...':publications.
```

In this rule, `@'http:...':publications` is the name of the model which contains the RDF-descriptions of the extracted publication informations. Further rules combine information on these authors from the researcher ontology with the author information. E.g. the following rule determines the employer of a

project member, which might be a company, or a university, or, in general, some instance of a subclass of an organization (see line three below: here, we query for some subclass (direct or inferred) of the class "Organization"):

```
FORALL A,I works_at(A, I) <- EXISTS A_id,X (name(A_id,A)
  AND ont:A_id[ont:involvedIn -> ont:I]@'http:...#':researcher
  AND ont:X[rdfs:subClassOf -> ont:Organization]@rdfschema('..':researcher)
  AND ont:I[rdf:type -> ont:X]@'http:...#':researcher).
```

Disambiguation of results – here especially resource identification problems caused by varying author names – is achieved by an additional name identification step. For a user with specific interests, for example "interest in personalized information systems", information on respective research groups in the project, on persons working in this field, on their publications, etc., is syndicated.

5   User Interface Provision

We run the PPR within our Personal Reader framework for designing, implementing and maintaining personal Web Content Readers [2]. These personal Web Content Readers allow a user to browse information (the *Reader* part), and to access personal recommendations and contextual information on the currently regarded web resource (the *Personal* part). For the PPR, we instantiated a personalization Web service in our Personal Reader framework which holds the above mentioned rules. An appropriate visualization Web service for displaying the results of the reasoning step (which are provided as RDF documents and refer to an ontology of personalization functionality) has been implemented.

Availability of the Personal Publication Reader

The concept of the Personal Publication Reader and its functionality are summarized in a video, and so are the web data extraction and maintenance tasks. All demonstration videos and access to the application itself are available via `http://www.personal-reader.de/semwebchallenge/sw-challenge.html`.

References

1. R. Baumgartner, S. Flesca, and G. Gottlob. Visual Web Information Extraction with Lixto. In *Proc. of VLDB*, 2001.
2. N. Henze and M. Kriesell. Personalization Functionality for the Semantic Web: Architectural Outline and First Sample Implementation. In *1st Int. Workshop on Engineering the Adaptive Web (EAW 2004)*, Eindhoven, The Netherlands, 2004.
3. S. Kuhlins and R. Tredwell. Toolkits for generating wrappers. In *Net.ObjectDays*, 2002.
4. A. Maedche, S. Staab, N. Stojanovice, and R.Studer. Semantic portal - the seal approach. In D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster, editors, *Spinning the Semantic Web*, pages 317–359. MIT-Press, 2003.
5. M. Sintek and S. Decker. TRIPLE - an RDF Query, Inference, and Transformation Language. In *International Semantic Web Conference (ISWC)*, Sardinia, Italy, 2002.