ELPUB Digital Library v2.0

Application of semantic web technologies

Anand BHATT\textsuperscript{a} and Bob MARTENS\textsuperscript{b}
\textsuperscript{a}ABA-NET/Architexturez Imprints, New Delhi, India
\textsuperscript{b}Vienna University of Technology, Vienna, Austria

Abstract. This paper presents the ongoing efforts to further develop the ELPUB digital library, which highly supported the dissemination of published materials within the community in the past years. elpub.scix.net has been serving the ELPUB-community since a decade, predominantly aiming to archive the output of the annual conferences. Doubtless, there is still a need for a platform to maintain the “collective memory” of the association and so far over 700 entries from 17 conferences were recorded. The repository, which utilized the SciX-technology, delivered the user access to the individual papers and made the data available via an OAI-interface as well. However, in the course of time digital library technology evolved and an association dedicated to Electronic Publishing ought to be at the forefront of novel developments. For this reason a shift was performed towards the Architexturez platform (library.elpub.net) aiming to implement advanced semantic web features. Especially the display of evolving topics and their gradual development is appealing and moreover the aggregation of individual bibliographies. Many of the features were designed in consultation with research communities in, among others, architectural computing and real estate. While deploying features and capabilities are well established in the digital library domain, the system is designed to support further research by the ELPUB community and this paper will elaborate on the transition and deliver an overview on the current prospects along with the technical capabilities.

Keywords. Digital repository, Open Access, Semantic Web, Mining, Social interaction

4. Introduction

Shortly before the turn of the millennium, comprehensive web developments took place and solutions to create topical repositories popped up. The rationale targeted to work out digital libraries - on a shoe-string budget -, which would not disappear shortly after their launch. Furthermore, there was no business model in the background aiming to keep the digital libraries running. Contrariwise, a closed pocket model based on volunteering capacities in academic environ-ments defined the starting point (likewise in ELPUB, which has no formal legal entity), which would then again donate an open access
contribution to the community.

The initial work to setup a digital library can be characterized in these days as the extraction of metadata. Fortunately hardly any back digitization had to be handled for ELPUB, but even after a couple of elapsed years it took some efforts to gather the digital data. Indeed by its original launch a critical mass was already available and the ELPUB Digital Library has been extended after every annual crop [1].

Given these framework conditions the SciX-technology was a convenient working environment, however, trends were not covered, as further development did not take place. Nonetheless, the maintenance as such has been secured with insignificant disruptions in the course of time [2]. For this reason a search towards alternatives was started. In terms of quantities, we’re moreover not talking about millions of records. However, even a couple of hundred recorded entries require a solution, which would go far beyond the basic idea of sustainable archiving.

2. Shift from SciX to the Architexturez Platform

The origin of the system goes back to 1999 when the creators of the system were required to develop experimental technology demonstrators, while writing technical standards for digital information management. The system has been developed over the years as the creators assisted in writing new standards, such as Unstructured Information Management Architecture (UIMA) and Extensible Resource Indicators (XRI) for the internet and provided the first statements of use.

The requirements for Architexturez were defined by way of a grant for Vienna University of Technology to initiate collaborative projects to initiate collaborative projects and proposed to the CumInCAD research community for review [3]. At the outset, it was determined that the novel platform should have added features such as named entity disambiguation, required to build lists of papers by entities such as authors and keywords, and semantic web features to better facilitate discovery and relationship mining activities in future.

2.1. New System Architecture

Figure 1 illustrates the reference architecture for the Architexturez platform. It is built with a Drupal web applications framework for the end-user facing tasks (items 5-7 in figure 1), from a heterogeneous set of
modules, primarily based on *Apache Solr* and *Mahout libraries* for the applications part (layers 2-4), and a storage layer for data required by system. Proprietary data may be retrieved at runtime from external sources, such as *social network containers, citations databases* and *institutional repositories*. Unstructured Information Management Architecture (UIMA) and Mahout libraries provide, respectively, for search and advanced text analysis and document clustering capabilities.

Figure 1 illustrates a *reference architecture* for the Architexturez platform, reading from bottom-to-top, the data storage layer, including data connectors to external services and repositories, a database abstraction layer, system core modules such as (3a) rate calculation logic and user finder and (3b) other modules; system services engines, such as faceted search engine, recommendation engine and people relevator engine, an access API; template engine for HTML 5 user interface and RDF/JSON, OAI-PMH and URN-NBN serializations for external customers, and user interface and a REST services endpoint for data extraction.

![Figure 1. The Architexturez repository system architecture.](image)

If we take one instance, author names and keywords, as over time, there will occur instances where an author has used different names or abbreviations, and synonyms for the same keyword are provided. At other times there may occur multiple authors of the same name. SciX relied on human judgment in disambiguating named entities. Comparable systems, such as *CiteSeer* [4], provided machine-learning frameworks of named entities [5] and others use rank-based algorithms...
Figure 2. Author disambiguation screen – the system proposes clusters of authors with similar names, and human editors can (1) merge authors (2) retain author names as alternate forms or (3) do nothing.

Architexturez preferred to combine automated methods and individual human judgment in disambiguating named entities. It unites automated and human-edited methods and clustering algorithms that build list of authors that may be related and then provides human editors with tools to make the final decision, shown in figure 2.

3. Novel Functionalities

The Architexturez-interface differs very much from the previous SciX-
solution. When entering the upgraded ELPUB Digital Library environment, the home page displays the previous set of conferences by way of the published proceedings. However, a user can also enter a search string at the top right. The bottom region of the page allows to embed aggregated information packages, such as news, calls for papers, or link collections, etc.

Once a user has selected certain proceedings, a new screen will open, where in the central part the recorded entries are displayed. On the left side, there is a collation of keywords, and on the right a list of contributing authors in these proceedings appears. Paper abstracts and keywords are searchable on this screen as well (see figure 4).
Figure 4. Display of proceedings with collated keywords and authors.

At this point the next step allows to follow a certain concept on the one hand to follow a keyword or, on the other hand, to trace back the publication output of a selected author. Especially novel members in the ELPUB community may take real advantage, in order to retrieve relationships between the evolving ELPUB topics and the involved authorship. At present predominantly PDF-files are serving as “multimedia” documents. However, other media types could be attached as well.

In term of intuitive searching/browsing, two features are being offered: (1) off-device navigation displaying citations and networks, (2) system-generated vector-space model to assist users in discovering new relationships. For the further development of the Digital Library as it stands, feedback from the side of the ELPUB-community is more than welcome.
The system embraces tools for ontology extraction for offline analysis of the documents and citations. It also provides a vector space model, for filtering and ranking relevant documents.

An important check such proposed and system-generated knowledge models must pass is the test of intuition, when authors and subject-experts review the related content. They should affirm that the related content contains documents that they would consider the most similar to the record being displayed.
4. Conclusion and Outlook

This paper described the switch of the ELPUB Digital Library towards a novel environment. The design parameters have been pointed out and a depiction of the system’s interface was presented. It is to be regarded as an open invitation to the ELPUB community to use the novel platform, i.e. to experiment and research beyond the annual conferences.

The contribution intends to think about possible options for further developmental tracks. Above all the ELPUB community should be on the edge of novelties as far as the own publication output is concerned. For example data aggregation can be regarded as an important issue, where “internal” and “external” data sets are combined in a structured way.

A system to harvest citations and references to ELPUB papers has been created, and it is currently processing references from external sources. It has been stated already many times, that the amount of information is exponentially growing and even so, the need for expedient navigation (sorting) is becoming more and more important.

References


