

# Web-DL: An Experience in Making Digital Libraries from the Web

Pável P. Calado Altigran S. da Silva Berthier Ribeiro-Neto  
Alberto H. F. Laender Juliano P. Lage Davi C. Reis  
Pablo A. Roberto Monique V. Vieira

Department of Computer Science  
Federal University of Minas Gerais  
31270-901 Belo Horizonte MG Brazil  
{pavel, alti, berthier, laender, palmieri, davi, pabloa, monique}@dcc.ufmg.br

Marcos A. Gonçalves Edward A. Fox  
Department of Computer Science  
Virginia Tech  
Blacksburg, VA 24061, USA  
{mgoncalv, fox}@vt.edu

## ABSTRACT

The Web contains a huge volume of information, almost all unstructured and, therefore, difficult to manage. In Digital Libraries, however, information is explicitly organized, described, and managed. In this paper, we propose an architecture that allows the construction of digital libraries from the Web, using standard protocols and archival technologies, and incorporating powerful digital library and data extraction tools, thus benefiting from the breadth of the Web contents, but supporting services and organization available in digital libraries. The proposed architecture was applied to the Networked Digital Library of Theses and Dissertations, providing an important first step toward rapid construction of large DLs from the Web, as well as a large-scale solution for interoperability between independent digital libraries.

## 1. INTRODUCTION

The Web contains a huge volume of information. Almost all of it is largely unstructured and, therefore, difficult to manage. Access is mostly through browsing and search engines, which make no assumptions about users' tasks or their specific information needs. In Digital Libraries (DL), however, users have more specific interests and information is explicitly organized, described and managed for communities with specific information needs and tasks. In this paper we propose Web-DL, an architecture that allows the construction of digital libraries from the Web. This architecture supports services and organization available in digital libraries, but benefiting from the breadth of the Web contents.

Digital libraries involve rich collections of digital objects and community-oriented specialized services such as searching, browsing, and recommendation. Many DLs are built as federations of autonomous, possibly heterogeneous DL systems, distributed across the Internet. One such federated digital library is the Networked Digital Library of Theses and Dissertations (NDLTD)<sup>1</sup>, an international federation of universities, libraries, and other supporting institutions focused on efforts related to electronic theses and dissertations (ETDs).

A difficulty in creating large federations is increasing motivation. So, some recent efforts aim to create looser groupings of digital libraries. The underlying concept is that the participants make some small efforts to enable some basic shared services, without specifying a complete set of agreements. The best example is illustrated by the Open Archives Initiative (OAI)<sup>2</sup>, which promotes the use of Dublin Core as a standard metadata format and defines a simple standard metadata harvesting protocol.

If the various organizations are not prepared to cooperate in any formal manner, a base level of interoperability is still possible by gathering openly accessible information, for instance, through the Web. Although there are a series of well known problems with the Web's data quality, many have collected data from the Web in order to develop collections of suitable size for various DL-like systems. The Harvest system, one of the first systems to apply focused gathering, had simple HTML-aware extraction tools [1].

In the following, we present our Web-DL architecture which: 1) combines harvesting and gathering to broaden the scope of interoperability in federated digital libraries; and 2) provides a framework to integrate a number of tools such as focused crawling, powerful data extraction and digital library toolkits, ultimately providing an infrastructure for build-

<sup>1</sup>See <http://www.ndltd.org/>.

<sup>2</sup>See <http://www.openarchives.org/>.

ing high-quality digital libraries from Web content. The NDLTD has particular characteristics that complicate interoperability and transparent resource discovery across its members, like member autonomy, decentralization, heterogeneity, and a massive amount of data. For this reason, the Web-DL approach is here illustrated by applying it to NDLTD to integrate OAI and non-OAI-complaint members.

## 2. OVERVIEW OF THE WEB-DL ARCHITECTURE

To build an ETD archive from the Web, ETD metadata must be collected from Web sites and integrated into a DL system. This operation has 3 main steps: (1) crawl the ETD Web sites to collect the pages containing ETD metadata, (2) parse the extracted pages to extract the relevant data, and (3) make the data available through a standard protocol. Figure 1 shows the Web-DL architecture for the integration and building of an ETD digital library from the Web.

Collecting Web pages with ETD information is done by using the ASByE tool. ASByE (Agent Specification By Example) is a user driven tool that generates agents for automatically collecting sets of dynamic or static Web pages. The ASByE tool features a visual metaphor for specifying navigation examples, automatic identification of collections of related links, automatic identification of threads of answer pages generated from queries, and dynamically filling of forms from parameters provided for the agents, by the user. In a typical interaction with the tool, through a graph-like interface, the user provides examples of how to reach the target pages, filling any form if needed, and how to group together related pages. The output of the tool is a parameterized agent that fetches the selected pages. The ASByE tool is fully described in [2].

Collected pages must then be parsed to extract the ETD metadata. Such is accomplished by the DEByE tool. DEByE (Data Extraction By Example) is a tool that generates wrappers for extracting data from Web pages. It is fully based on a visual paradigm which allows the user to specify a set of examples of the objects to be extracted. These example objects are taken from a sample page of the same Web source from which other objects (data) will be extracted. The tool then derives from these examples an *Object Extraction Pattern (OEP)*, that includes information on the structure of the objects and on the textual context in which the actual data composing the objects appears in the Web pages. The OEP is then passed to a general purpose wrapper that uses it to extract data from new pages from the same Web source, provided that they have structure and content similar to the sample page. For a full discussion of the DEByE tool and the DEByE approach, we refer the interested reader to [5].

In order to be used by the MARIAN system, or any other DL system, data must be stored in a more structured way, (e.g., MARC or XML), normally using community-oriented semantic standards (e.g., Dublin Core or FGDC for geospatial data). In our work, we use ETD-MS, a metadata standard for electronic theses and dissertations<sup>3</sup>. Nonetheless, data in ETD Web sites is frequently in a non-standard, non-structured formats. In fact, four main problems were

found, when converting data to ETD-MS: 1) mandatory information is not present in the ETD page; 2) information is present, but only implicitly; 3) data is not in a required format; and 4) the text fields are not in the appropriate coding system. The first problem can be solved by inserting some default value. The second problem, however, is more difficult to solve. For this work, we manually filled in the required information but, in the future, an automated solution should be found. For the third and fourth problems, several appropriate conversion routines, easily available in many programming languages, were used.

After the data in ETD-MS format is stored, an OAI server set up on top of the local database will make it available to anyone using the OAI metadata protocol. In our particular case, available for the MARIAN system. MARIAN is a digital library system designed and built to store, search over, retrieve, and browse large numbers of diverse objects in a network of relationships [4]. In the context of the Web-DL architecture, MARIAN provides searching and browsing services for NDLTD. Data from OAI providers and from non-OAI-complaint members, coming from the Web-DL architecture, is integrated into a Union Catalog. MARIAN is equipped with OAI harvesters able to collect data for the Union Catalog periodically. MARIAN is completely reconfigurable for different DL collections, using digital library generators and a special DL declarative language, called 5SL [3], for this purpose.

## 3. AN EXAMPLE WEB ETD DIGITAL LIBRARY

For this work, we collected pages containing ETDs from the sites of 21 different institutions, selected from the list of NDLTD members, available at <http://www.theses.org>. These sites contained a total of 9595 ETDs. It was not possible to collect from 7 sites, since these were off-line or available only through a search interface. Of the 6 mandatory ETD-MS fields, an average of 29.5% were missing in the collected pages, and were therefore filed with a default value. Table 1 shows number of ETDs where each mandatory field was missing. Some optional fields also were filled with information, if it was known.

Field name	ETDs missing
dc.title	43 (0.4%)
dc.creator	23 (0.2%)
dc.subject	2349 (24%)
dc.date	283 (3%)
dc.type	703 (7%)
dc.identifier	4800 (50%)

**Table 1: Mandatory fields missing from the collected ETDs.**

For each collected site only 1 example per site was needed to create the crawling agents, and an average of 1.5 examples per site to create the data parsers. This represented about 2 hours of work per site, by a specialized user. In the future, however, we expect to further automatize this process, in order to reduce the time needed, as more sites are harvested.

## 4. CONCLUSIONS

<sup>3</sup>See <http://www.ndltd.org/standards/metadata/>.

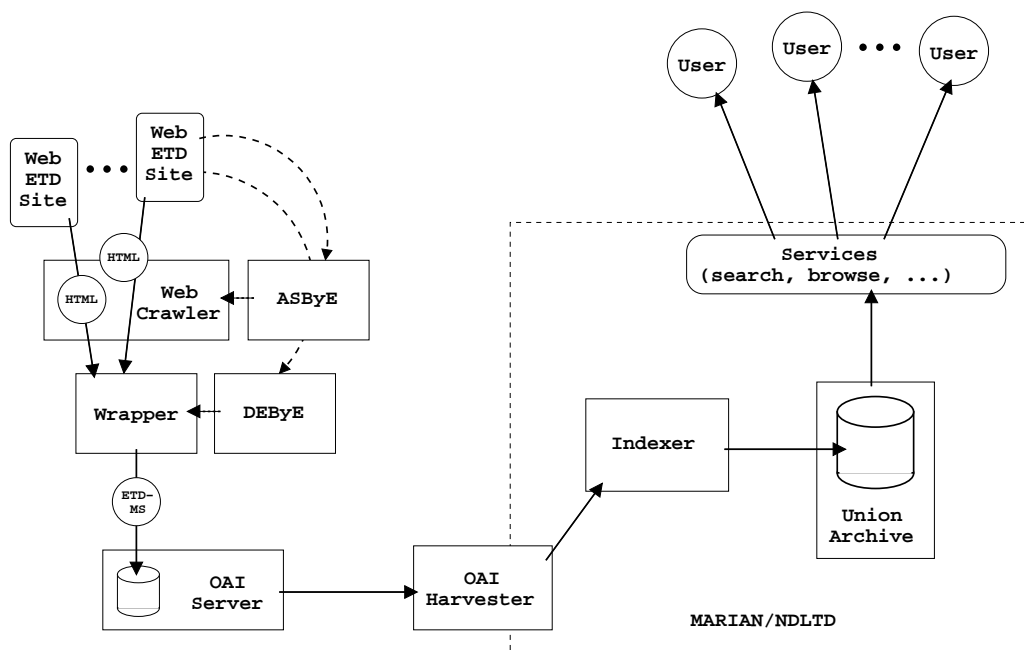


Figure 1: Proposed architecture for the integrated DEByE/MARIAN system.

Moving from the Web to a digital library is not a trivial task. Collecting ETD pages from the Web presented, in general, the same problems found by general Web crawlers, like offline sites, or sites that provide access only through search interfaces. For this reason, it might not always be possible to guarantee that all the information present in the Web may be covered by the DL.

Experimental results have shown the effectiveness of our approach for Web data extraction [5]. However, Web pages without any discernible syntactic context might be impossible to handle. Also, for some sites, pages containing ETDs are very different from each other and, therefore, it might not be possible to build a generic wrapper for the whole site. Data extraction is still possible, in such cases, however, wrappers must be built manually.

Finally, we face the problem of making the unstructured Web data fit a standard pattern. Information for mandatory fields may be missing, data may need to be converted to standard formats or controlled vocabulary. Although all these problems can be solved by more or less complex routines, they are still an obstacle if we intend to create a general conversion solution.

In sum, each of the tasks for extracting information from the Web, into a DL environment presents its own set of problems. A general solution for building digital libraries from the Web depends on general solutions for each of these tasks, and an efficient integration of such solutions. The solutions we presented for each task, although not completely general, provide important clues on how to further automate the process.

Currently, the Web-DL architecture for ETDs is in an experimental phase. However once the system is put in pro-

duction mode, it has the potential to double the current coverage of the NDLTD Union Archive in terms of the number of universities and available data. Web-DL can also be easily expanded to domains other than ETD, since the AS-ByE and DEByE tools are generic and the MARIAN system can be reconfigured through the 5SL language.

## 5. REFERENCES

- [1] C. M. Bowman, P. B. Danzig, D. R. Hardy, U. Manber, and M. F. Schwartz. The harvest information discovery and access system. *Computer Networks and ISDN Systems*, 28(1-2):119–125, Dec. 1995.
- [2] P. B. Golgher, A. H. F. Laender, A. S. da Silva, and B. Ribeiro-Neto. An Example-Based Environment for Wrapper Generation. In S. Liddle, H. Mayr, and B. Thalheim, editors, *Conceptual Modeling for E-Business and the Web, ER 2000 Workshops on Conceptual Modeling Approaches for E-Business and The World Wide Web and Conceptual Modeling*, pages 94–101. Springer, Berlin, Germany, 2000.
- [3] M. A. Gonçalves and E. A. Fox. 5SL: A Language for Declarative Generation of Digital Libraries. In *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL'02*, page To Appear, Portland, Oregon, USA, 2002.
- [4] M. A. Gonçalves, R. K. France, and E. A. Fox. MARIAN: Flexible interoperability for federated digital libraries. *Lecture Notes in Computer Science*, 2163:173–189, 2001.
- [5] A. H. F. Laender, B. Ribeiro-Neto, and A. S. da Silva. DEByE – Data Extraction by Example. *Data and Knowledge Engineering*, 40(2):121–154, 2002.