A Scalable Digital Library Infrastructure Expands Search and Beyond

Min Song[1], Shuyuan Mary Ho, Michael Bieber,
Eric Koppel, Vahid Hamidullah, and Pawel Bokota

New Jersey Institute of Technology, College of Computing Sciences,
Department of Information Systems,
University Heights, Newark, NJ 07102
{song, smho, bieber, erk7, vh22, pmb9} @ njit.edu

1 Introduction

A deep web book search becomes a challenge as the number of digital libraries increases, and so does the demand of sophisticated users for searching requirements over several digital libraries and search engines before getting to the desired book selections. A scalable, integrated search infrastructure is needed to help users to effectively search structured content information based on identified name entities across heterogeneous digital libraries. IntegraL, an NSF funded project, is a light-weight system that aims to integrate multiple digital libraries, including book search engines, and meta-search engines by using link analysis (Song & Bieber 2008). One of the objectives of IntegraL is to allow users to flexibly enable or disable the use of the subscribed libraries as plugins through a graphical administrative page.

In the following we would like to review why an infrastructure is necessary for integrating multiple digital libraries, describe how IntegraL is an innovative way of expanding the use of multiple digital libraries and book search engines at the user's preference, list benefits of using IntegraL, and conclude our study.

2 Background

Rao (2004) described the progression of information search from the 60's to the 90's. Users' information search has been drastically enhanced from simple query-in, result-out in the 60's, to information digest, indexing, extraction, categorization, visualization, and further to federated research. While users' search capability has been empowered, the design and development of digital libraries have become more sophisticated. Information retrieval will be more based on open and flexible infrastructures (Rao 2004; Kazai & Doucet 2008). However, this scalable archival infrastructure takes the collaboration among heterogeneous digital libraries. Being able to accurately retrieve documents from distributed uncooperative digital libraries becomes critical with foreseen and unforeseen problems. They include issues with

[1] Corresponding author

archival preservation of digital content, indexing in each collection, represent-able query phrase, effective use of metadata for search, robust retrieval algorithms, seemliness interactions between user and the data, integration between services and tools, and inevitably privacy and security considerations when accessing data for sensitive purpose.

3 System Infrastructure

IntegraL offers a scalable infrastructure that links among heterogeneous digital libraries through the development of a web-based proxy server, sitting on top of Tomcat, an open source servlet container as the blue line represented in Figure 1. When a user makes a request to access a digital library, he or she is authenticated by a single sign-on (SSO) mechanism. We adopt an open source authentication mechanism, Shibboleth, in order to allow seamless browsing across digital libraries that have also adopted Shibboleth. Once the user is authenticated, this SSO mechanism allows the user to surf among various subscribed digital libraries without going through repetitious logins at each stage. The user's credential information is stored in hash files on the proxy.

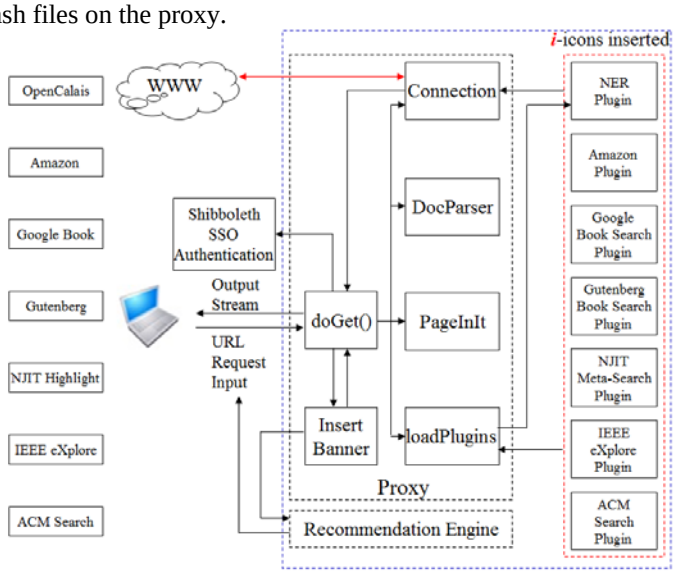

Figure 1: IntegraL System Architecture

All users' requests are handled by the proxy. When a user browses any other subscribed digital library, their request to access other digital library is taken care of by the proxy. If a user requests to use other search engines, the pages of their request are returned to the user untouched. When a user makes a URL request, an IntegraL banner is inserted on top of the digital library. All HTML pages are converted into DOM Documents; all relative URL paths are converted to absolute URL paths. What we innovatively create is the addition of i-icons (Figure 3). The i-icons are inserted whenever a name entity is recognized by OpenCalais, a service that annotates data

with rich semantic metadata. The name entity recognition module, or NER plugin, receives the document requested by the user from the proxy, analyzes the lexical meaning of recognized categories, such as person, location, etc., and then creates rich semantic metadata that serves as recommended search for user's further references.

As illustrated on the right-hand side of the Figure 1, plugins are developed using XPath, which parse the HTML documents and insert i-icons wherever elements of interests have been located. The development of plugins is based on templates (Figure 2). This provides flexible expansion of integrating search engines and digital libraries. IntegraL web-based administrative interface empowers users with ad hoc configuration.

4 Benefits for Book Search

The IntegraL infrastructure holds several benefits to users' book search. First, while the system horizontally links multiple digital libraries, it also provides a broad spectrum of vertical search according to identified name entities. For example, while users search with a particular engine, say Google Book Search, the system, particularly the NER Plugin, automatically identifies other name entities from those search results that interest the users and further give options to dive into other digital libraries from the selected element of interest, such as author, title, language, location, etc. These elements of interest are either automatically discovered by their lexical meaning, or can be manually discovered and identified for further linking using XPath. This benefit leads to the second benefit, which allows users that are interested in the same element of interests from other digital libraries, to systemically get connected to, say Google Book Search engine, and vice versa (Figure 3). Third, the IntegraL infrastructure allows users to turn on or off their preferred digital libraries in a simple administrative menu. This option allows users to specify specific search needs and can initially filter out undesired search. Fourth, this user-programmable function also helps the systems to respond in a timely manner.

```
Document parse (Document doc, HttpServletRequest
req)
{
   for each eoi type found for this page
   {
      if user has this eoi type enabled
      {
         retrieve the xpath for that eoi type from cache;
         run xpath on doc and save matches;

         for each match
         {
            insert eoi node;
         }
      }
   }
   return doc;
}
```

Figure 2: Plugin Template

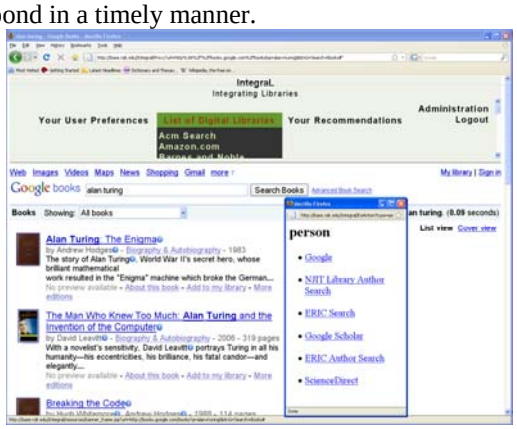

Figure 3: IntegraL links with Google Book Search

5 Conclusion

Without the complication of complete system integration, IntegraL adopts a light-weight approach that links multiple heterogeneous digital libraries and search engines. It allows interoperability among different search results, search engines and digital libraries. The system itself is mostly built on open-source software, which can be reliable, auditable and cost-effective. This approach provides recommendations to users for further deep search based on identified elements of interests among wide ranges of virtual resources.

Acknowledgments. Partial support for this research was provided by the National Science Foundation under grants DUE-0434581 and DUE-0434998, by the Institute for Museum and Library Services under grant LG-02-04-0002-04, and by the New Jersey Institute of Technology.

References

Kazai, G. and Doucet, A. (2008) *Overview of the INEX 2007 Book Search track: BookSearch '07*.

Rao, R. (2004) *Queue: Open Source Grows Up,* **2,** 66-73.

Song, M. and Bieber, M. (2008) In *10th IEEE Conference on E-Commerce Technology and the 5th IEEE Conference on Enterprise computing, E-Commerce and E-Services* IEEE, pp. 369-375.