A unified resource for post-secondary program information

Wendy Osborn
University of Lethbridge
Lethbridge, Alberta, Canada
wendy.osborn@uleth.ca

Steve Fox
University of Lethbridge
Lethbridge, Alberta, Canada
steve.fox@uleth.ca

Seamus O'Shea
University of Lethbridge
Lethbridge, Alberta, Canada
oshea@uleth.ca

Abstract
We present a prototype for a post-secondary information resource that is maintained as a digital library. This information resource provides students with a one-stop-shop for all programs from participating institutions. The content is obtained from the World Wide Web, is style-unified, is organized for searching and browsing from different angles, and is updated nightly to keep information current. We present the architecture and functionality of the collection building process. We also present some current issues we are working on.

Keywords
Digital Libraries, Greenstone, HTML page unification, Automatic collection building.

1. Introduction
In this paper, we present a prototype for a post-secondary information resource that is maintained as a digital library collection. The post-secondary resource maintains information on the degree and diploma programs offered by participating post-secondary institutions. The purpose is to provide a one-stop-shop for students to browse and search for information on programs that they may consider pursuing at the post-secondary level.

This resource has the following features. First, it uses the services of the Greenstone digital library software for retrieving information from the World Wide Web, building the resource and its required searching and browsing mechanisms, and providing the interface for accessing the resource. Second, formatting modules are added to Greenstone to unify HTML pages and to reject pages that are outside the scope of the resource. Third, searching and browsing are customized to facilitate information access from different angles. Finally, the resource is updated nightly in order to keep its information current.

We focus our presentation on the architecture and functionality of the post-secondary resource. We begin with some required background information on the Greenstone digital library software. Then, the system architecture is present, followed by the functionality of the post-secondary resource building process and the automated updating approach. Finally, we conclude with current work.

2. Greenstone

Greenstone (Witten & Bainbridge, 2002) is a software suite for constructing and disseminating digital library collections. A collection is available to users on a local machine or over the Internet. It has many properties that make it a desirable choice for a post-secondary information resource. Greenstone is customizable in many ways. Different searching and browsing criteria can be created for the same collection. For example, the same collection can be set up for browsing by author, document type, and title. Also, the overall style (or, look-and-feel) of a collection is customizable. For example, if an organization has a specific web page style, this can be incorporated easily into a Greenstone collection.

In addition, Greenstone contains many modules, which can be selected and applied to a specific problem domain. Two modules that are specific to Greenstone and applicable to our work are Import and Build. Import takes a set of documents and adds them to a collection, while Build creates the indexes and classifiers for the collection. Two additional modules that are included with Greenstone are Wget (Niksic, 2007) and PDFtoHTML (Ovtcharov & Dorsch, 2007). Wget is a program that retrieves website content. Starting with a root page, Wget recursively visits each link in the page to retrieve other linked pages. PDFtoHTML, as the name implies, takes a document in PDF format and converts it to HTML format. After conversion, the document can be treated in the same manner as other HTML documents.

3. System Architecture

The architecture of our proposed system incorporates existing modules from Greenstone with additional modules that handle the unification of HTML pages and maintain the list of program websites from participating institutions. Figure 1 depicts the overall architecture.

The Web Document Retriever (Wget) module retrieves HTML and PDF content from, each participating website in the Post-Secondary Site List. Because the documents that are retrieved depend on the links obtained from HTML pages, some documents that are external to the main site may also be retrieved. All documents are passed to the HTML Unifier module. This unifier module is responsible for performing the following two-phase process:

- Unifying the appearance of HTML pages, and;
- Removing irrelevant HTML pages

Any documents that are in PDF format are passed to the PDF to HTML converter (PDFtoHTML) module to convert them to HTML before they are considered for selection. Finally, the unified pages are passed to the Collection Builder module, which contains the Import and Build modules, so the collection can be created.

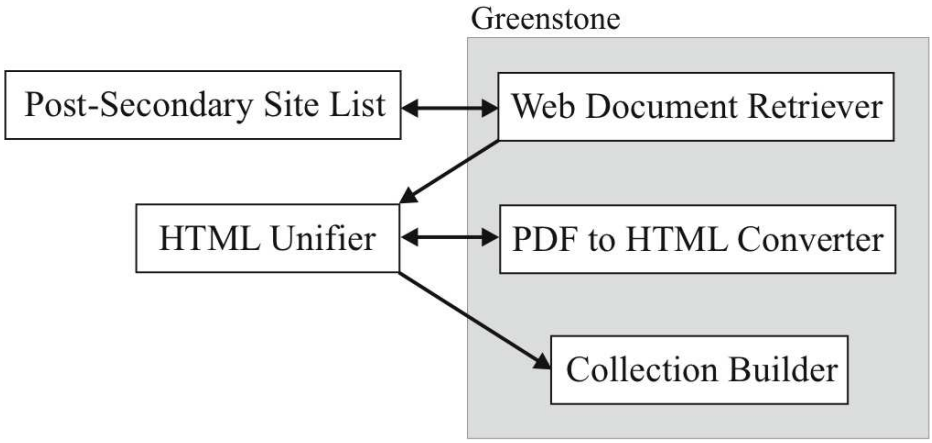

Figure 1: System Architecture

We present the functionality of the HTML Unifier module over the next two sections. First, we present the HTML page unification approach. Then, we present the approach for removing irrelevant documents.

4. Unifying Web Content

Greenstone provides a module to process HTML pages before creating a collection that contains them. We found additional issues that needed to be addressed in order to unify the documents for our requirements:

- **Broken and unwanted links**. The Greenstone HTML module has an option for setting up active links in a page. However, if a link on the original site does not work, then it will not work in a Greenstone collection. In addition, some links lead to information that is outside the scope of the collection. Therefore, to keep our focus on content, we decided to remove all links from the document.
- **Formatting differences**. Greenstone can remove some style formatting and content (such as images). However, there is some unwanted formatting may remain after the style-removal process of Greenstone. For example, if an image was inside a table, the table is not removed when the page is processed. In addition, each participating site has its own page style. Collectively, they are not visually appealing.

Therefore, the first phase of the HTML Unifier module is to remove the majority of formatting information, such as HTML tags and their content, from the documents for all participating sites. The resulting HTML pages contain mostly text information and some HTML formatting that can be processed into metadata. The pages are passed to the filtering phase of the process.

5. Filtering Web Content

As mentioned, some HTML documents that are obtained by the Web Document Retriever module are outside the scope of the resource, and therefore are considered irrelevant. The second phase of the HTML Unifier module is to remove irrelevant pages. Currently, it employs a simple search for specific keywords that are related to post-secondary programs. These keywords are

required to be in the document in order for it to be included in the collection. Any pages not containing the keywords are rejected. The reason why filtering is not performed before formatting is because some phrases that may qualify as potential keywords may be removed during the HTML page unification stage.

6. Automatic Updating

As mentioned, the collection of post-secondary resources must be updated nightly to ensure that its content is kept current. Furthermore, it is desirable to have this update occur automatically. The best approach for this is to use the scheduling program on the host operating system to perform the updating.

To automate this process, a script was created (Osborn & Fox, 2007) to combine the retrieval, unification and construction of the resource. Since the post-secondary resource currently resides on Linux, the cron scheduling program (Nemeth, Snyder & Hein, 2007) can be configured to run the script on a nightly basis.

7. Conclusions

We present a prototype for a unified post-secondary information resource. This resource is updated nightly to keep content current. This work in progress can be viewed by visiting *http://www.sadl.uleth.ca/playbox/cgi-bin/library*, and clicking on "Post-secondary Resource".

Some current issues we are addressing are the following. The first is to add the service of a web crawler to perform focused crawling (Bergmark, 2002). The purpose of the crawl is to identify relevant websites for participation in the Post-Secondary Site List. The second is to incorporate some of the HTML Unifier functionality into the HTMLPlug of Greenstone, so that other HTML collections can use it. The final one is to add a summarizer to the system so that some HTML documents can be condensed further into a minimal amount of relevant information.

References

Witten, I., and Bainbridge, D. (2002). *How to Build a Digital Library*. Morgan-Kaufmann.

Niksic, H. (2007). *GNU Wget*. Retrieved March 12, 2008 from GNU Project Web Site: http://www.gnu.org/software/wget.

Ovtcharov, G. and Dorsch, R. (2007). *PDFtoHTML*. Retrieved March 12, 2008 from Sourceforge Web Site: http://pdftohtml.sourceforge.net.

Osborn. W. and Fox, S. (2007). Automatic and scheduled maintenance of digital library collections. In *Proceedings of the 2nd International Conference on Digital Information Management (ICDIM 2007)*.

Nemeth, E., Snyder, G., and Hein, T.R. (2007). *Linux Administration Handbook*. Prentice-Hall.

Bergmark, D. (2002). Collection Synthesis. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL 2002)*.