Computer Supported Workflow for Cataloging and Management in Digital Libraries

Weimao Ke, Javed Mostafa, and Gayathri S. Athreya
Laboratory of Applied Informatics Research
University of North Carolina at Chapel Hill, {wke@unc.edu, jm@unc.edu, gathreya@lanl.gov}

Service integration is important in digital library implementation. In this paper, we present a project called Extensible Networked Association-based Bioinformatics Learning Environment, or ENABLE. It computerizes a cataloging workflow by integrating backend services and facilitates interleaving between automatic cataloging & human curation activities. To the user end, it provides various interaction means that support bioinformatics learning. Features of the current implementation include automatic bioinformatics resource harvesting, basic information retrieval operations, resource cataloging and management, visualization and learning tools.

Introduction

With the advancement of the Internet and Digital Library technologies, more and more bioinformatics resources have become available online. Bioinformatics researchers and learners alike have immediate access to the immense database of Web resources through hyperlinks and search applications. However, novices are likely to be overwhelmed by the huge amount of information and its heterogeneity. The lack of tools aimed at novices is a gap in current Web information delivery services and hinders learners who wish to access relevant and useful information. One way to facilitate the availability of novice or learner-oriented resources is by delivering tools and processes for creating them. Integration of existing resources is important in digital libraries, especially for learning environments (Oldenettel, Malachinski, & Reil, 2003). Digital libraries of this kind should not only pull together resources but also provide an efficient way of cataloging and managing these resources. They should seek to improve the overall workflow by increasing efficiencies and adding values (Tuai, 2006). In ENABLE, we have developed a cataloging workflow based on standardized backend services and integrated them to provide various interaction means for bioinformatics learners.

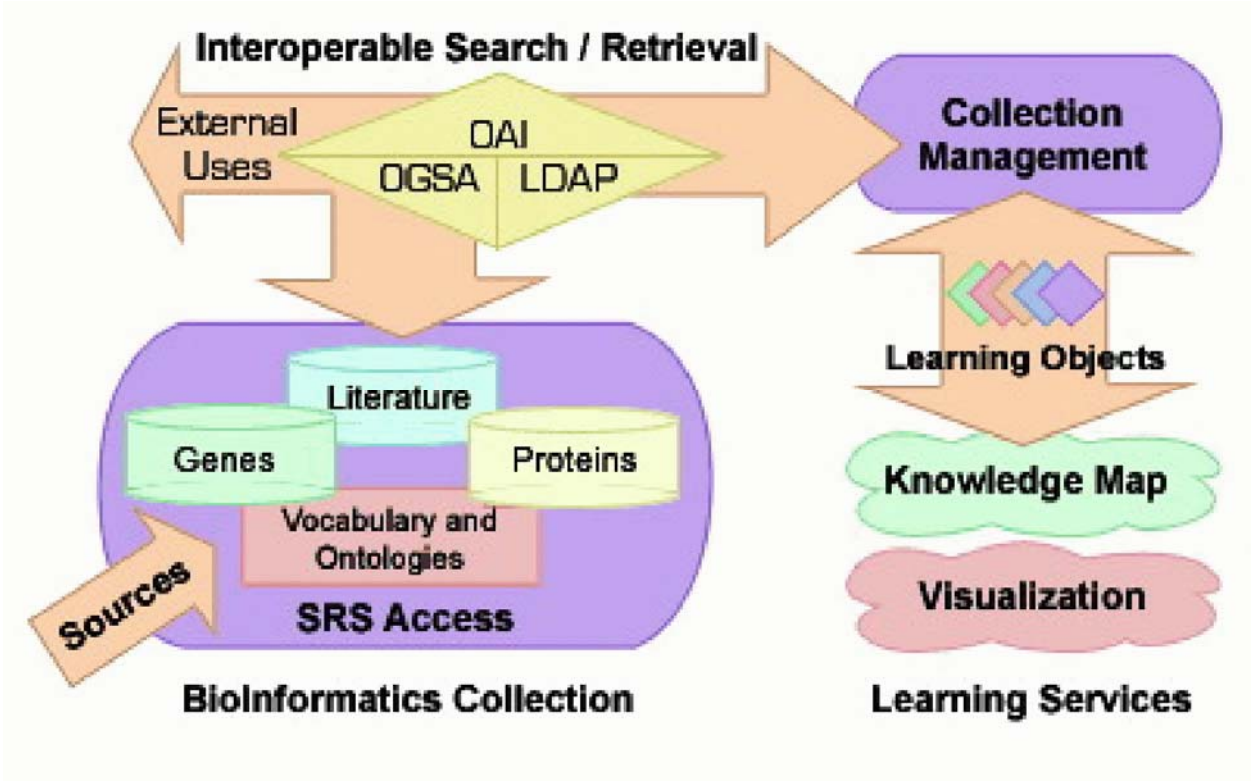


Figure 1: ENABLE Architecture

The ENABLE Project

ENABLE is a multidisciplinary project funded by the NSF's National Science Digital Libraries program. The major goal of this project is to apply advances in digital library technologies to the emerging domain of bioinformatics, and to develop novel interaction means that support learning based on identifying and visualizing associations among key dimensions of bioinformatics resources. As shown in Figure 1, ENABLE consists of three major components: 1) the bioinformatics collection; 2) resource management; and 3) learning services. In addition, interoperable search/retrieval services based on standard protocols are implemented to integrate fundamental digital library functions.

Bioinformatics Collection

The bioinformatics collection component automatically collects and updates information of online educational resources for bioinformatics. To run the crawler, the width and depth of the crawl and seed URLs need to be specified. The crawler can either automatically categorize the pages or can let the editors do it afterwards. Once the crawler is triggered, it estimates the completion time and runs on the server side. So as not to cause too much network traffic and possible DoS attacks, the crawler's delay and time-out options should be defined properly in advance.

Collection Management

After the crawler is finished, the collected items are available for authorized editors to edit and catalog. The collection management interface, shown in Figure 2, allows one to update URL seeds, edit collected resources, check collected links, and categorize them.

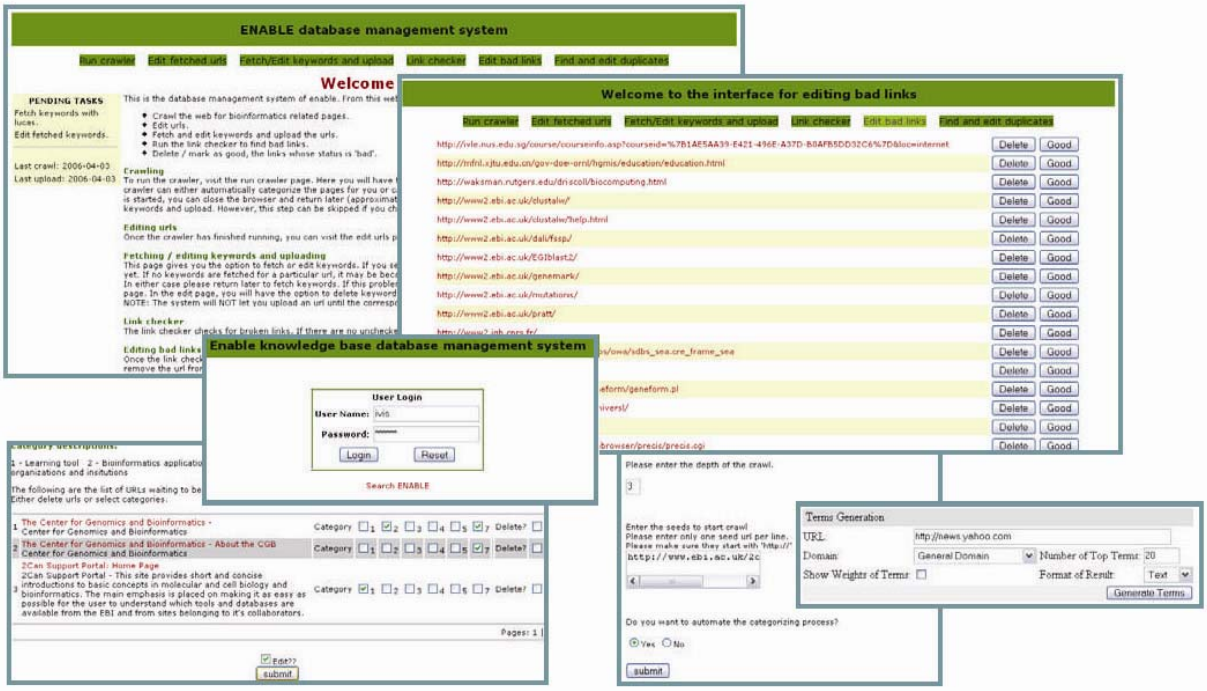


Figure 2: Management Interface

Editing and Cataloging

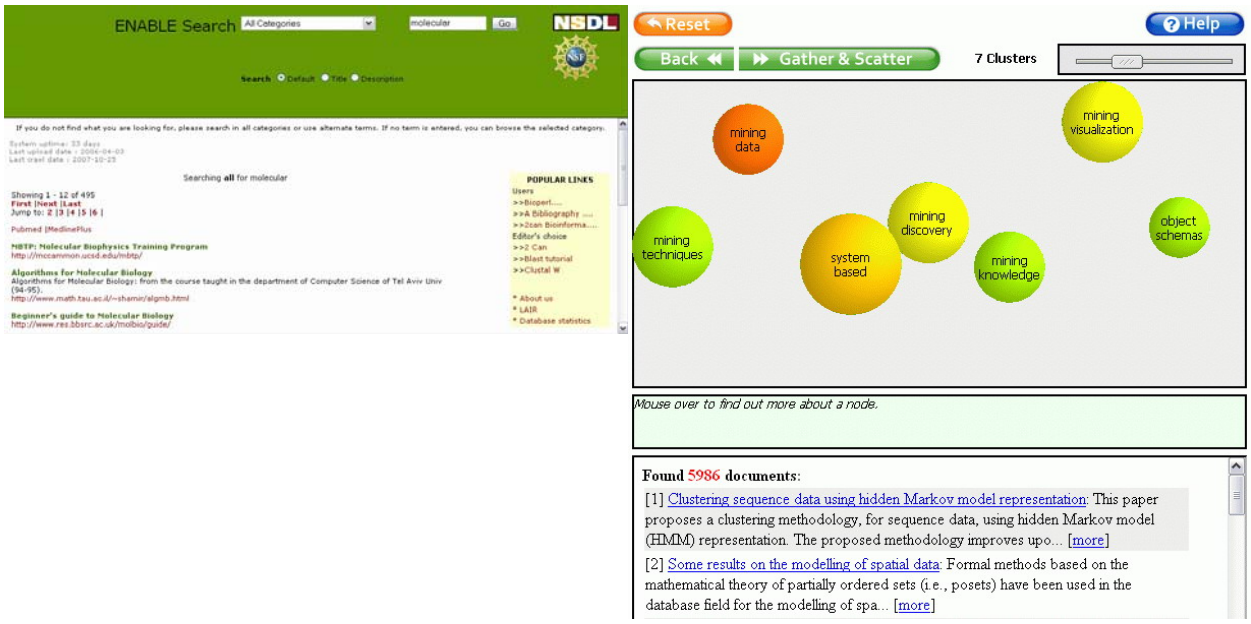
An editor has the options to fetch, edit, and upload keywords for the crawled resources. This is done by integrating a web service library called LUCAS II (Ke, Fu, & Mostafa, 2005), which is a part of the interoperable services shown in Figure 1. LUCAS generates keywords for the URLs that do not have keywords yet. A bioinformatics domain is used to weight terms within the knowledge domain and to improve the quality of term generation. Editors can add or delete keywords manually, if the generated keywords are not satisfactory. They can also assign categories to the resources. The resources are uploaded to a production database after being properly edited and cataloged.

Checking

Accessibility of the resources in a consistent and predictable way is very important. A link checker module is implemented for editors to check broken links periodically. When the checker is done, a list of bad links are provided. Editors can either delete a URL or change its status. Deleting the URL does not remove the URL from the database, but merely changes its status to “broken” and prevents it from being shown to users. The deletion is revokable in case the resources become accessible later. Another module is also provided to check duplicate items.

Learning Services

Integration of the backend components facilitates the access to heterogeneous resources available on the Web. Based on the collected and cataloged resources, ENABLE provides several educational tools for bioinformatics learners. These tools provide interaction by means of search, interactive browsing, and visualization of associations, as shown in Figure 3 (a) and (b).



(a) Web search (b) Scatter/Gather browser

Figure 3: Learning Services

Web Search

A Web search interface is available for users to search, browse, and access the resources (Figure 3 (a)). Several features have been implemented to insure accessibility, usability, and data transparency. Users are allowed to browse the categories and do free-text search for titles and descriptions. Source of each listed item, system uptime, and last update time are provided to give users a sense of credibility and freshness of the bioinformatics resourses. In a seperate window, frequently browsed and editor-selected links are also presented to provide guidance to learners.

Scatter/Gather Browser

As an interaction mode, Scatter/Gather is well known for its ease of use and effectiveness in situations where it is difficult to precisely specify a query (Cutting, Karger, Pedersen, & Tukey, 1992). It helps users refine search criteria through iterations of cluster selection and reclustering. In ENABLE, we provide this type of resource navigation through a visualization, shown in Figure 3 (b). Initially the system presents the top-level clusters of the whole collection. Then a user can select her favored clusters and click “Gather & Scatter” to recluster related bioinformatic resources into sub-level clusters. Iteratively, a user can not only narrow down the search results, but can also learn more about concepts and associations in the domain.

Conclusion and Future Work

With all of these tools, ENABLE integrates heterogeneous bioinformatics resources on the Web, provides a computerized workflow for cataloging, and facilitates interleaving between automatic and human archives. It also offers its users various perspectives to view its resources and to learn bioinformatics through associations. Future work will involve integrating the learning services, populating the system with more quality data, and automating the resource categorization process.

Acknowledgments

Authors acknowledge the NSF grant ENABLE #0333623.

References

Cutting, D. R., Karger, D., Pedersen, J. O., & Tukey, J. W. (1992). Scatter/Gather: A cluster-based approach to browsing large document collections. In *SIGIR '92: Proceedings of the 15nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 318–329).

Ke, W., Fu, Y., & Mostafa, J. (2005, June). Advanced information retrieval Web services for digital libraries. *Library Collections, Acquisitions, and Technical Services*, *29*(2), 220–224.

Oldenettel, F., Malachinski, M., & Reil, D. (2003). Integrating digital libraries into learning environments: the LEBONED approach. In *JCDL '03: Proceedings of the 5th ACM/IEEE-CS joint conference on digital libraries* (pp. 280–290).

Tuai, C. K. (2006). Implementing process improvement into electronic reserves: A case study. *Journal of Interlibrary Loan, Document Delivery & Electronic Reserve*, *16*(4), 113–124.