Using Digital Library Components for Biodiversity Systems

Ricardo da S. Torres[1],
Claudia Bauzer Medeiros[2],
Renata Q. Dividino[3], and
Mauricio A. Figueiredo[4]
IC, University of Campinas, Campinas, SP, Brazil

{[1]rtorres,[2]cmbm}@ic.unicamp.br

{[3]ra993183,[4]ra009423}@ic.unicamp.br

Marcos André Gonçalves,
Edward A. Fox, and
Ryan Richardson
Virginia Tech, Blacksburg, VA, USA

{mgoncalv,fox,ryanr}@vt.edu

ABSTRACT

Biodiversity information systems (BISs) involve all kinds of heterogeneous data, which include ecological and geographical features. However, available information systems offer very limited support for managing such data in an integrated fashion, and such integration is often based on geographic coordinates alone. Furthermore, such systems do not fully support image content management (e.g., photos of landscapes or living organisms), a requirement of many BIS end-users. In order to meet their needs, these users – e.g., biologists, environmental experts – often have to alternate between distinct biodiversity and image information systems to combine information extracted from them. This cumbersome operational procedure is forced on users by lack of interoperability among these systems. This hampers the addition of new data sources, as well as cooperation among scientists. The approach provided in this project to meet these issues is based on taking advantage of advances in Digital Library (DL) innovations to integrate networked collections of heterogeneous data. It focuses on creating the basis for a biodiversity information system under the digital library perspective, combining new techniques of content-based image retrieval and database query processing mechanisms. This approach solves the problem of system switching, and provides users with a flexible platform from which to tailor a BIS to their needs.

The main contributions of this project are the following: (a) a generic architecture for managing heterogeneous collections, based on digital library components, to access heterogeneous biodiversity data sources (text and images), that allows combining text-based and content-based queries in a seamless way; and (b) a new component, for content-based image search, integrated into that architecture.

The proposed architecture has been implemented by using DL components which are mostly new or recently developed [1, 3]. Furthemore, its implementation uses the Open Archives Initiative (OAI) protocol [2] as a basis for interoperability. This architecture is easily extensible, and provides users a considerable degree of flexibility in data management. To illustrate our claim that this architecture can be applied to several domains, we are investigating its application in building a biodiversity information system on

fish species. This solution solves many current problems in this kind of system, allowing handling of images and textual information in an integrated fashion.

A new Content-Based Image Search Component has been developed to support queries on image collections. Since this component is based on the OAI principles, it provides an easy-to-install search engine to query images by content. It can be readily tailored for a particular collection by a trained designer, who carries out a clearly defined set of pilot experiments. It supports the use of different image descriptors, which can be chosen from the pilot experiment, and then easily combined to yield improved effectiveness. In addition, it encapsulates a multidimensional index structure to speed up the search process, that also can be easily configured for different image collections.

Categories and Subject Descriptors

H.3.7 [**Information Systems**]: Information Storage and Retrieval—*Digital Libraries*

General Terms

Design, Experimentation

Keywords

Biodiversity Information System, Content-Based Image Retrieval, OAI

Acknowledgments

1. REFERENCES

[1] ESSEX, 2003.
http://br.endernet.org/~akrowne/elaine/essex/index.html (as of December 2003).

[2] OAI. Open Archives Initiative, 2003.
http://www.openarchives.org (as of December 2003).

[3] H. Suleman. *Open Digital Libraries.* PhD thesis, Computer Science Department, Virginia Tech, Blacksburg, VA, 2002.
http://scholar.lib.vt.edu/theses/available/etd-11222002-155624/unrestricted/odl.pdf.