Digital Libraries for Biodiversity and Natural History Collections

Authors

Miguel Ruiz
University of North Texas, Department of Library and Information Sciences
1155 Union Circle 311068. Denton, TX 76203-1068
Email: Miguel.Ruiz@unt.edu

Jacob Kramer-Duffield
University of North Carolina, School of Information and Library Science
216 Lenoir Drive CB#3360, 100 Manning Hall, Chapel Hill, NC 27599-3360
Email: jkd@email.unc.edu

Jane Greenberg
University of North Carolina, School of Information and Library Science
216 Lenoir Drive CB#3360, 100 Manning Hall, Chapel Hill, NC 27599-3360
Email: janeg@email.unc.edu

Nathan Hall
University of North Texas Libraries
1155 Union Circle #305190. Denton , TX 76203-5017
Email: Nathan.Hall@unt.edu

This panel aims to discuss the importance of creating digital libraries for biodiversity and natural history collections, the state of the art in terms of standards, best practices and the challenges that natural history museums and herbaria face when trying to digitize their collections , and the creation and management of personal digital libraries for botanical learning.

Overview

Natural History collections have been created by museums and other institutions for

preserving plant and animal specimens collected for scientific use. These institutions hold a large number of specimens and artifacts that represents the biodiversity of our planet and contain important information about the existence of species that in some cases are now extinct. For example, approximately 3000 Herbaria, which are special natural history collections that collect plant specimens, have been created around the world in 145 countries and contain approximately 300 million specimens (Holmgren et al. 1990). Herbaria specimens are pressed flat and dried, and mounted on individual sheets of paper of standard size. Each specimen includes a label that contains metadata such as the scientific name, place where the specimen was collected, name of the person who collected it, and name of the person who did the identification. The specimen may also include other data such as the name of the institution or collection, and other information added during the curation process such as geo-coordinates and measures of data quality (Morris, 2005).

There has been a significant interest in the past 10 years to create and expand digital repositories for natural history collections. Examples of these initiatives include the National Biological Information Infrastructure (www.nbii.gov/portal/server.pt) and the Global Biodiversity Information Facility (www.gbif.org). Complementing these developments has been research using Web2.0 technologies connecting to personal digital libraries for describing, identifying and keeping a store of digital plan specimens. The advantages of creating digital collections include the increased use and exposure of collections to a wider audience, the reduction of physical manipulation of the specimens, and the creation of a permanent digital archive with associated metadata that facilities optimal organization, retrieval and exchange of the information that can benefit the scientific community as well as educators and the general public.

Broadening the potential user base for these systems means incorporating new vocabularies, and investigations concerning the use of unstructured free-text tagging systems for plant description have found these approaches hold promise as a way for entry-level botany students to begin mapping the domain (Shoffner et al. 2008; Daniel et al. 2008). These approaches lower the hurdle of mapping and understanding the botanical taxonomic system and its key-based identification logics, allowing for a common-sense description of plants as a legitimate approach to identification.

This panel will present an overview of the state of the art in the creation of digital repositories for natural history collections. We will discuss current metadata standards (e.g., DarwinCore, and Access to Biological Collection Data (ABCD)), computerized methods available to generate metadata using optical Character recognition (OCR) and automatic information

extraction, methods for optimizing the usage of machine-assisted and human-assisted procedures that effectively and efficiently convert textual data of the specimen labels to a structured machine readable format, and the use of flexible Web2.0 Memex frameworks for experiential and collaborative learning processes.

Presenter Kramer-Duffield and Dr. Greenberg will present findings on their current work with Bot2.0, an NSF-funded multidisciplinary program that is employing a suite of Web 2.0 technologies, mobile phones and hands-on experience, to introduce and lay the groundwork for greater retention of non-traditional students (especially women and minorities) into botany.

Dr Ruiz will present an overview of state-of-the-art optimal mechanisms to combine machine aided and human-aided digitization process as well as automatic systems such as HERBIS (Heidorn & Wei, 2008) that combine OCR and automatic information extraction for digitization of specimen labels.

Nathan Hall will talk about the University of North Texas Libraries initiative on developing an international digital library for environmental science and policy. This initiative will include biodiversity and natural history collections for use in research and education. Two specific digital collections will allow further research into the impact of human activity on biodiversity in freshwater systems. The first collection will be derived from the End of Term Web Harvest dataset. The second collection will come from the digitization of 6000 freshwater mussel specimens. UNT Libraries partnered with the Library of Congress, the Internet Archive, the California Digital Library, and the Government Printing Office to conduct a harvest of all U.S. government web content before the 2008 election, before the 2009 inauguration, and after the 2009 inauguration. The resulting twenty five terabyte dataset documents all policy changes between the Bush administration and the Obama administration. Through link structure analysis, natural language processing, and tools for web archive indexing, UNT will parse and extract environmental policy and science documents from the dataset and make them available in the environmental science digital library.

For the other collection mentioned above, UNT Elm Fork Natural Heritage Museum plans to digitize its collection of 6000 North American freshwater mussels and make the collection available through the UNT Libraries Digital Collections. Information about the mussels may be useful for evaluating environmental policies. As filter feeders, mussels are particularly sensitive to changes in their environment, and factors such as toxins and sediment runoff can affect their prevalence in an area. As a result, scientists can use changes in freshwater

mussel populations to track water pollution (Howells et al. 1996). As a part of their life cycle, freshwater mussels in the larval stage (glochidia) attach themselves to certain fish (Bauer, 2001) for a period of time and later drop off, potentially in a new habitat. Because the mussels require certain species of fish to act as hosts, fluctuation in fish populations also affect the prevalence of mussels. The collection metadata will include details about where and when each specimen was found, as well as the general habitat and range of each species.

By linking together the data of these two discrete collections, UNT hopes to allow researchers to determine whether changes in biodiversity can be a useful measure of the effectiveness of environmental regulations as an aspect of human activity. The work explained in this paper represents just the beginning of collection building for the environmental science digital library, and each additional collection will extend its utility as a research tool and source of data for scientists.

References

Bauer, G. 2001. Characterization of the Unionoida (Naiads). In: Caldwell, M.M., Heldmaier, G., Lange, O.L., Mooney, H.A., Schulze, E.D., Sommer, U. (Eds.), Ecology and Evolution of the Freshwater Mussels Unionoida. Springer, pp. 3-4.

Greenberg, J., Daniel, E., Kramer-Duffield. J., Seiberling, S., Weakly, A., and Woodbury, D. (2008). BOT 2.0 – Botany through Web 2.0, the Memex and Social Learning. Inventions and Impact 2: Building Excellence in Undergraduate Science, Technology, Engineering, and Mathematics (STEM) Education: A Conference of Course, Curriculum, and Laboratory Improvement (CCLI) Program National Science Foundation, Division of Undergraduate Education, 13–15 August 2008, Washington DC, pp. A32.

Heidorn, P.B & Wei, Q. (2008) Automatic metadata extraction from museum specimen labels. In Proceedings of the International Conference on Dublin Core Metadata Applications.

Holmgren, P.K., N.H. Holmgren and L.C. Barnett. 1990. Index herbariorum. Part I: The herbaria of the world.
8th edition. New York Botanical Garden. 693 pp.

Howells, R.G., Neck R.W., Murray, H.D. 1996. Freshwater Mussels of Texas. Texas Parks and Wildlife Department, Austin

Morris. P.J., 2005. Relational database design and implementation for Biodiversity Informatics.
Phyloinformatics 7:1-63. (http://www.athro.com/general/Phyloinformatics_7_85x11.pdf)

Shoffner, M., Greenberg, J., Kramer-Duffield, J. and Woodbury, D. (2008) Web 2.0 Semantic Systems – Collaborative Learning in Science. International Conference on Dublin Core and Metadata Applications – Metadata for Semantic and Social Applications 22 - 26 September 2008, Berlin (DC-2008)