Creating a Digital Library of Original Taxonomic Literature

Susan K. Kendall
Michigan State University Libraries

Anne Marie Karle-Zenith
Michigan State University Libraries

Elizabeth J. Bollinger
Michigan State University Libraries

ABSTRACT

The Michigan State University Libraries are collaborating with a professor in the Plant Biology Department of Michigan State University on the Euglenoid Project. The National Science Foundation is funding the project as part of their recent program to enhance expertise in taxonomy, a field critically low in researchers needed to address modern problems in biodiversity and environmental biology. The euglenoids are a group of algae with both ecological and evolutionary importance; however, few scientists in the world possess expert knowledge about them. The project involves the creation of a freely available, web-accessible illustrated treatise on taxonomy of the euglenoids that would be an example for taxonomists of other organisms, and serve to pass down knowledge to the next generation. The Library's role in the grant is to digitize and make accessible the original, mostly German, 19th and early 20th century defining euglenoid literature. These works, consisting of both monographs and journal articles, have continuing importance for taxonomists but are not readily available to many of the scientists who need to study and cite them.

Our paper will center on the challenges we faced in undertaking this project and the many decisions that had to be made in the process. An initial challenge was locating the literature, due to the idiosyncrasies of 19th century scientific publication and the vagaries of citations. Once the materials were located and obtained from various sources, they were scanned to create high-resolution, preservation-quality images. We decided to provide several formats for presentation to the public on the Web even though that would require more server space. For fast loading and printing for reading copies, we created lower resolution PDF files. For closer study of the figures of organisms, we are creating higher resolution PDF and jpeg files. Other choices we faced were about how to index the literature and make it searchable, and how to provide bibliographic access to the digitized materials. We considered automated versus manual indexing and different metadata schemes. The granularity at which we should provide access to the digital objects via our online catalog was a concern. Finally, we will discuss the issue of making

digital items available beyond our own web site so that researchers and librarians searching the Web more easily discover them. This involves making the information harvestable through a standard such as the Open Archives Initiative, an idea that is gaining in importance as the number of digital items like these available on the Web continues to rise.

INTRODUCTION

The euglenoids are a group of free-living, eukaryotic, unicellular organisms. The prototype genus of this group was discovered in 1786 and named *Euglena* in 1830 by C. G. Ehrenberg. Their taxonomy is uncertain and interesting because in the past some scientists classified them as animals and some scientists classified them as plants as some of the species contain chloroplasts, cellular organelles that contain the green photosynthetic pigment chlorophyll. Even today, two different current classification schemes exist for these organisms, and the classification scheme has changed over time since their first discovery. Because of these taxonomic disputes, this group of organisms is often referred to as euglenoids, a common term, rather than by an official taxonomic Class or Order name.

Both freshwater and marine euglenoids can be found all over the world, and they have both evolutionary and ecological importance. However, few scientists in the world are currently studying these organisms, and there is concern that important knowledge about them will be lost. The lack of scientists studying the euglenoids is just part of a larger problem that few scientists are training to become taxonomists who study the systematics and phylogeny of diverse species. Without taxonomists with expertise in diverse species, there will not be enough scientists to address modern environmental problems such as biodiversity and species extinction. In order to encourage more scientists to consider taxonomic studies, the National Science Foundation has developed and awarded grants under the Partnerships for Enhancing Expertise in Taxonomy (PEET) program to fund taxonomic studies, train new taxonomists, and make the research accessible on the Web (Jacobson, 2001). Dr. Richard Triemer, chairman of the Plant Biology Department at Michigan State University, who focuses his taxonomic studies on the euglenoids, is a recipient of one of these PEET grants for the *Euglenoid Project*. Besides training new taxonomists and euglenoid specialists, the project also involves creation of a Web-accessible illustrated treatise on the taxonomy of the euglenoids that would pass down knowledge and be an example for taxonomists of other organisms.

The *Euglenoid Project* includes a digital library component for which the Michigan State University Libraries is a partner. As part of a larger database-driven Web site, which will contain current information about and images of various euglenoid species, the digital library component will compile and make available in electronic format the known taxonomic literature about the euglenoids, which is no longer under copyright protection. Literature describing euglenoid species and their taxonomy begins in the 1830s with largely German publications. Most universities only have a portion of these materials. While contemporary taxonomists studying the euglenoid species usually have ready access to more recent scientific articles, finding some of this older literature can be

challenging and time-consuming. Information found in these older studies can be relevant and enlightening to modern studies, but the difficulty of access has often led scientists to cite some of these defining studies second-hand without ever reading them or viewing the figures for themselves. Providing free digital copies of this literature will remove barriers for scientists to read and use the original literature.

DIGITIZING THE LITERATURE

The *Euglenoid Project* digital library of original taxonomic literature will number approximately 75 individual items. These range from whole monographs to journal articles, and, in some instances, only parts of larger works. We began with a bibliography compiled by Dr. Triemer of predominately German works, but also works in French, English, Russian, Polish, Spanish, and Latin. An initial challenge was to determine correct citations for each work since many citations were from the older literature and were incomplete. Several publications also were found to carry more than one title in different languages, and library cataloging of 19th century journals, particularly, is not consistent. Holdings are not reported or can be unclear so that in several cases libraries were contacted to find actual items on shelf to verify a citation or holdings. A little less than half of the items to be digitized were available at Michigan State University Libraries, so the remaining materials to be digitized were usually borrowed through interlibrary loan. In a few instances, we purchased digital images of materials from libraries that were unwilling to loan the originals.

In planning this digital project we consulted *A Framework of Guidance for Building Good Digital Collections*, 2nd edition, 2004, a document produced by the National Information Standards Organization (NISO) (NISO Framework Advisory Group, 2004). To determine digital capture specifications, we also considered the recommendations of the Research Libraries Group's *Guide to Planning an Imaging Project*. (Colet, 2000, section 2.3) The *Guide* suggests a use-neutral rather than a use-specific approach, digitizing the images at high resolution, but with neutral standards. This insures that the object is only digitized once. It can then be manipulated after the fact to conform to the needs of specific projects and applications (http://www.rlg.org/visguides/visguide1.html#2.3)

Based on this approach, preservation-quality TIFF images were produced in 24-bit color at 400 dpi. Items were digitized using either a HP scanjet 7400C color flatbed or Bookeye overhead color scanner. For presentation on the Web, quality of images was of primary concern because much of this literature contains either black and white or color figures (mostly drawings) of the organisms. These figures are not merely ornamental but will be studied by scientists, and therefore availability of high-resolution images is crucial. Therefore, we created several types of derivative files from the TIFF images. Lower-resolution PDF files were created for fast loading and printing of reading copies. Higher resolution PDF files and jpeg images were also created to allow close viewing and magnification of the figures.

Various factors such as condition and size of the books determined when we used the overhead instead of the flatbed scanner. Some of the challenges we faced with the overhead color scanner centered on its sensitivity and the best environment was only found after a series of trials and errors. We found that we needed to create an enclosed environment for the scanner because natural light from windows or light from overhead fluorescents caused artifacting or color differences that affected the quality of the scans. Scanning workflow was also an issue we addressed. Because much of our labor involves student workers, we created a workflow that would allow many different people to handle the scanning and processing without too much documentation. We chose the Opus scanning system (from Image Access) because of the workflow management. The Opus system allows us to create a scanable barcode for each item so that files are never lost or confused while we are working on multiple projects at the same time. The Opus software also allows batch processing of files to ready many different images for OCR and PDF processing. This has been a fantastic time saver for us since in the past those were processes which were done by hand.

Optical Character Recognition (OCR) was initially not planned due to accuracy problems and the variety of languages involved in the project. However, OCR accuracy rates have improved and the numbers of languages and alphabets OCR can recognize has increased. Generating OCR files will allow us to provide full text searching capabilities within a particular PDF text file. To generate the OCR files, we are using the OCR engine that is part of the OPUS software package, Readiris Pro 10. This software allows for batch processing of the OCR/PDF creation, has better compression for the PDF files, and allows us to load dictionaries to recognize terms that don't already exist in their lexicon. With this OCR engine we have noticed a significant increase in OCR accuracy from other programs we have used in the past. Readiris has some limitations with the Russian typefaces but it has recognized the Latin alphabet languages with no problems.

INDEXING

Besides providing full text searching within each particular digitized text, we also wanted to provide some searching across the whole digital collection, which required that the texts be indexed. We chose to focus the indexing to allow searching only for names of euglenoid genera and species because those Latin names remain consistent despite the different languages in which each text is written. We are manually indexing each document for every instance of a euglenoid genus and species mentioned in the text and for every instance of a euglenoid genus and species for which there is a figure, along with the page numbers. An XML data structure is used to store the indexing information so that it can be read by the search engine that will be built for the project. A document type definition (DTD) was created to define the XML document structure and the allowed indexing elements.

BIBLIOGRAPHIC ACCESS

To provide bibliographic access for the digitized items in the *Euglenoid Project* we created a metadata record for each resource, as well as a traditional MARC record that will be contributed to WorldCat.

The choice of which metadata scheme to use was informed by a cost-benefit analysis such as described by the NISO *Framework of Guidance for Building Good Digital Collections* (2004). We wanted to use a schema that was compatible with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Use of this protocol allows data providers to make their metadata available for "harvesting" by metadata harvesting services that aggregate the metadata from the institutions they harvest and provide search tools and interfaces to facilitate discovery and retrieval. The MSU Libraries is an OAI provider, and we eventually plan to make the Euglena metadata available to harvesters such as University of Michigan's OAISTER, the National Science Digital Library, and the Committee on Institutional Cooperation (CIC) metadata portal.

We considered three different OAI-compatible metadata schemes for this project: Simple Dublin Core, Qualified Dublin Core and Metadata Object Description Schema (MODS). A few of the resources being digitized for the Euglenoid project are complete books, but most are journal articles, or parts of books, or even parts of articles. In the end we concluded that a robust Qualified Dublin Core set would be the best choice, as it is rich enough to convey the various granularities of the digital objects without having to create four different element sets, yet it is less complex than MODS and therefore would be easier to implement.

Another issue we addressed was how the metadata would be expressed and stored. We chose to use XML so that we could work on the documents from anywhere and move them around easily, allowing self-sufficiency. In addition, XML has the advantage of storing data only in ASCII text, rather than a proprietary database format.

In accordance with Collections Principle #2 of NISO's *Framework* document *(2004)*, we will also create collection-level records for inclusion in registries such as the National Science Digital Library, as well as for OAI harvesting. For our online catalog and for WorldCat, we will also be creating a collection level MARC record representing the entire project.

WEB SITE AND SEARCH ENGINE

As mentioned above, the *Euglenoid Project* involves the creation of a Web-accessible illustrated treatise on the euglenoids. The digital library portion consisting of older original scientific literature will be one part of a larger whole that will include informational pages on all of the euglenoid species and their taxonomy along with contemporary photographs and short movies. The data for the site will be stored in a mySQL database and the site and search engine will be built using Java. There will be several access points for the digital library. First, there will be a browsable page listing the items in the collection by author last name and title. This will allow people looking for a particular article to easily determine if it is included in the collection. Each item in the collection will have its own page from which will be linked a low-resolution PDF image for printing, quick viewing and full-text searching, a high-resolution PDF image for closer viewing of the figures, and jpeg images of each individual page in the document. Second, there will be a keyword search through the metadata of the items in the digital library as well as the data in the rest of the Web site. The combined search will return grouped results, setting aside a section for the literature and a section for the content of the Web site. The keyword search will allow the user to enter a genus or species name to retrieve information about that organism from the Web site and references to various pages in the digitized literature. A particular challenge for the project is that some genera and species have been renamed over time. To address this, we have created a table that associates the older name(s) with the current versions so that when researchers search under either name, they will be taken to the correct information. Lastly, the literature references for each genus and species will appear on the Web site's pages for each genus and species as a browsable item.

REFERENCES

Colet, L. S. (2000). Guide to planning an Imaging Project. Retrieved April 29, 2005 from http://www.rlg.org/visguides/visguide1.html.

Jacobson, J. (2001, September 27). Saving a Dying Field. *The Chronicle of Higher Education: Chronicle Careers*. Retrieved April 29, 2005 from http://chronicle.com/jobs/2001/09/2001092701c.htm.

NISO Framework Advisory Group (2004). A Framework of Guidance for Building Good Digital Collections. 2nd edition. Retrieved April 29, 2005 from http://www.niso.org/framework/framework2.html.