Submitted on: 9/2/2014

The Art of Life: Merging the Worlds of Art and Science

Trish Rose-Sandler

Center for Biodiversity Informatics, Missouri Botanical Garden, St. Louis, MO, USA
trish.rose-sandler@mobot.org

Nancy E. Gwinn

Smithsonian Libraries, Washington, DC, USA
gwinnn@si.edu

Constance Rinaldo

Ernst Mayr Library, Museum of Comparative Zoology, Harvard University,
Cambridge, MA, USA
crinaldo@oeb.harvard.edu

Abstract:

This paper shows how a digital library, created primarily for the use of biologists, is reaching a broad audience of artists, art historians, exhibition and graphic designers, publishers, and others in humanities fields, thus merging the worlds of science and art.

Created by a global consortium of natural history, botany, agricultural, university and national libraries, the Biodiversity Heritage Library (BHL) is a rich domain repository of historic biodiversity literature. Providing open access to over 43 million pages of text (approximatly140,000 volumes) via its portal (http://www.biodiversitylibrary.org/), the BHL has developed into an essential research tool for biologists around the world.

Within these texts, but not easily accessible, are millions of visual resources, plates, figures, maps, and photographs, many produced by the finest botanical and zoological illustrators in the world, such as John James Audubon, George Dionysus Ehret, and Pierre Redouté. When BHL staff began to duplicate the beautiful plates in these volumes to a Flickr site, now containing 90,000 images, its popularity made clear that the BHL work could attract an audience far outside the scientific world. There was also a need to automate the identification process and create relevant metadata, a laborious procedure that currently requires significant staff time.

In 2012, the Missouri Botanical Garden embarked on an ambitious project to automatically identify and describe all natural history illustrations in BHL texts, not just the plates, in order to make them more easily accessible and able to be shared with other repositories, such as ARTstor, Encyclopedia of Life, and the Digital Public Library of America. This paper describes the project and shows how scholars, educators, designers—and image lovers—will be able to find and view a wealth of illustrations of plant and animal life from which to make connections between science, art, culture, and history.

BHL, biodiversity, "biodiversity heritage library", "natural history illustrations", EOL

Introduction

The Biodiversity Heritage Library (BHL) is both a project and a product. As a project, it is a global collaboration of natural history museums, botanical gardens, agricultural, university, biological research libraries, and like institutions (Gwinn & Rinaldo 2009). Their combined purpose is to improve and make more efficient the methodology of research in biodiversity studies by making biodiversity literature openly and freely available to the world as part of a global biodiversity community--a laudable purpose, as research in biodiversity includes such topics as the identification and classification of species, as well as their impact on the environment (and the environment and on them) and their evolution. Biodiversity research helps us to understand the world around us and relates to issues of land management, migration, agriculture, and other topics of utmost importance today.

As a product, the Biodiversity Heritage Library is a constantly expanding digital repository, currently containing 140,000 books, journals, and articles, totalling to over 43 million pages of text and illustrations. Almost all of the content is in the public domain and 133 publishers have given permission for their copyrighted content to be included. Many rare and valuable volumes, dating from the 18th century, also appear in the library. Recently BHL Members began to add archival material in the form of field notebooks, diaries that scientists and explorers kept while on their expeditions, which describe and locate the specimens collected, the landscape, the weather, and other information invaluable to document specimen collections. In addition, they contain personal observations, perhaps drawings or maps, travel information, and the journeys undertaken. Field notebooks are a largely unplumbed resource for historians as well as scientists and are generally housed in museum departments or archives -- perhaps drawers and closets -- of the institutions that sponsored the expeditions and fieldwork.

Searching the BHL

While all BHL content is available through the Internet Archive (https://archive.org/), a more specialized portal was developed to serve the needs of scientists at http://www.biodiversitylibrary.org/. The portal was primarily built for scientists with an interest in systematic biology and taxonomy and the primary mode of entry to the database was by species name, book or journal title, and book author. The search capability has recently evolved so that content can be searched at a more granular level such as chapter or article level (also known as segments).

General users or users new to BHL can perform a simple keyword search or can browse the content by Title, Author, Date or Collection. More advanced users will want to take advantage of the advanced search features where specific fields can be searched.

For Books and Journals, the searchable fields are: Title, Author Last Name, Volume, Edition, Year, Subject, Language and Collection (e.g. Book of the Week or Darwin's Library). For Articles and Chapters, the searchable fields are Article, Journal or Book title, Author Last Name and Year. Not all articles and chapters are indexed in the BHL yet. If an article is not found with this method, a user can search for the title of the journal within which the article occurs and then navigate to the specific page where the article or chapter is expected to begin. The majority of segments are identified and indexed via BioStor.org. Since segment identification is incomplete, a user should not rely on a segment search for comprehensive coverage. A subject search is based on the Library of Congress Subject Headings and may not be intuitive to some users.

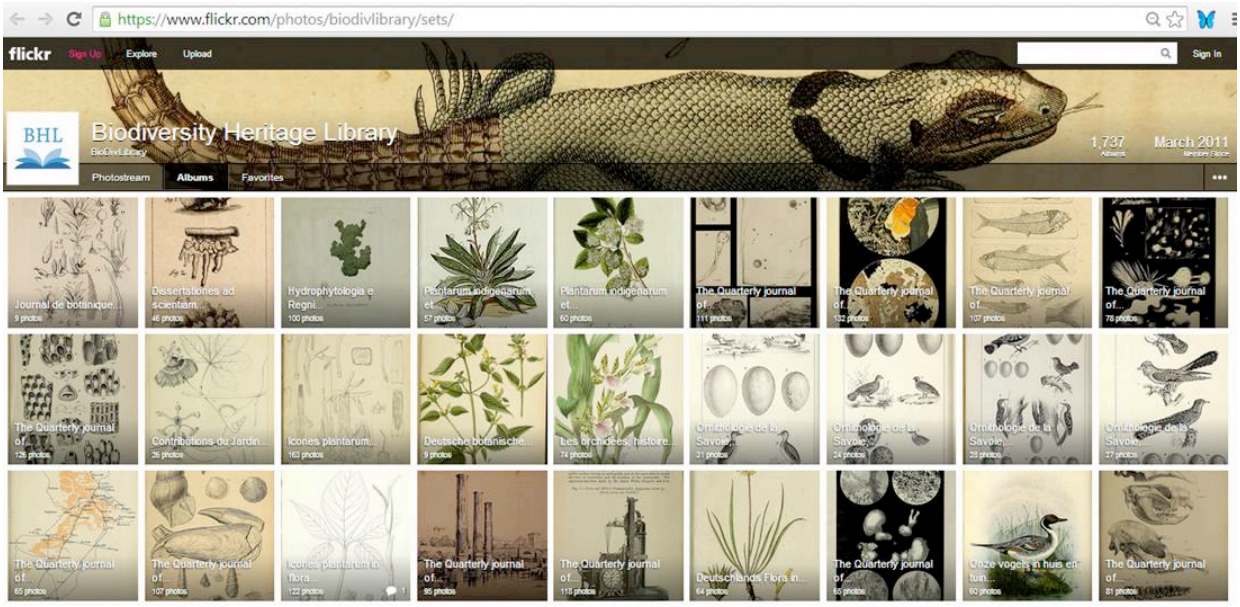
Scientific names are extracted from the OCR'd text and matched to controlled terms via a variety of online taxonomic name services. As noted above, the portal was originally built to answer the needs of systematic biologists and taxonomists who often search for a species name, such as *Microtus pennsylvanicus*. If a scientific name is known, it is easy to create a bibliography by using http://www.biodiversitylibrary.org/name/Scientific_name where "Scientific name" is any uninomial, binomial or trinomial taxonomic category.

BHL bibliographic records are also aggregated and discoverable in other repositories such as Europeana (http://www.europeana.eu/) and the relatively new Digital Public Library of America (http://dp.la/). A researcher can also find relevant literature from BHL through the Encyclopedia of Life (EOL http://eol.org/), a global initiative that is producing a web page for every identified species.

The texts in the Biodiversity Heritage Library are critical, but of equal value both to scientists and a broader audience, are the illustrations. They range from full, folio page-sized, hand-colored artworks to line drawings within the text, maps, charts, graphs, and diagrams. The balance of this paper will focus on these illustrations and the use that they have far beyond the scientific world.

The Art of Life

How can a user find illustrations? The Art of Life project grew out of a BHL problem statement that resulted from this question. BHL has a critical mass of textual content online (over 43 million pages). BHL users knew there were amazing images within the BHL pages, but there was no easy way to find them, other than clicking on a single book or volume in the BHL portal and scrolling through page by page. Also, there is often no descriptive metadata attached to the illustration that would tell you its content, such as, the species depicted, the date when the illustration was created or who was involved in its creation. From the beginning, requests for illustrations, both general and specific, came in from scientists, artists, exhibitors, educators, designers, and students, and those who just loved images. While some volumes have detailed metadata that point to illustrations, it was partially or wholly derived manually, and clearly extremely laborious. So, staff began to pull out the luscious full-page illustrations of plants and animals and collect them on a Flickr site (https://www.flickr.com/photos/biodivlibrary/sets/) for easier access.



BHL Flickr stream containing over 90,000 images.

Images from the BHL Flickr site are re-used in BHL's Facebook page (https://www.facebook.com/BioDivLibrary) as quizzes and attractive posts and pinned on BHL's Pinterest boards (http://www.pinterest.com/biodivlibrary/bhl-images/). The Flickr site now contains nearly 90,000 images, which are linked back into BHL as individual images and to the volume that contains them. By tagging the images in Flickr with scientific names ("taxonomy:binomial=Genus species"), they can be automatically ingested into the EOL Flickr site (https://www.flickr.com/groups/encyclopedia_of_life/) and displayed on the corresponding EOL species pages, thus facilitating re-use of BHL images and exposing them to an even broader audience. More than 11,000 of the BHL images have been tagged and re-used in EOL. Use of these images made clear that the BHL work could attract an audience far outside the scientific world, an exciting prospect, as shown by the more than 4.2 million views of BHL Flickr content in 2013. BHL members wanted to expand its content to new audiences and domains; illustrations presented in social media venues were the obvious pathway to reach users in the arts and humanities, as well as more general audiences. However, it was clear that a more automated way to identify, extract, and provide detailed metadata for individual images was needed to scale the process to the millions of images available in BHL.

NEH Proposal

In the fall of 2011 the Missouri Botanical Garden's (MOBOT) Center for Biodiversity Informatics developed a project proposal that sought to liberate natural history illustrations from the digitized books and journals in the online Biodiversity Heritage Library (BHL) through development of software tools for automated identification and crowd sourced description of visual resources. MOBOT applied for a grant to support this effort through the National Endowment for the Humanities' (NEH) Division of Preservation and Access under the Humanities Collection and References Resources (HCRR) program, which supports projects that "provide an essential underpinning for scholarship, education, and public programming in the humanities." (http://www.neh.gov/grants/preservation/humanities-collections-and-reference-resources) A panellist from the NEH grant review committee stated the following in deciding to back this project - "They are proposing something that is pretty useful here. They wish to realize (using contemporary technologies and media) what these

18th to early 20th century naturalists originally intended – a searchable, visual inventory of all things in the natural world (here in the form of botanic illustrations)."

NEH recognized the value of this content for the humanities community and the need to make it more widely accessible and therefore granted MOBOT $260,000. The project, named *The Art of Life: Data Mining and Crowdsourcing the Identification and Description of Natural History Illustrations from the Biodiversity Heritage Library,* began in May of 2012 and was scheduled to run for two years. More recently, NEH agreed to an extension for the project until April 2015.

Objectives

The Art of Life project has five primary objectives:

1. Define an appropriate metadata schema for natural history illustrations
2. Build software tools to automatically identify illustrations in the BHL corpus
3. Enhance existing tools to enable the classification of the illustrations
4. Push them into crowd sourcing applications to enable a community of users to add descriptive metadata for the illustrations
5. Integrate the descriptive metadata generated by users back into BHL portal

Metadata schema. The project sought to develop a metadata schema that could fully describe natural history illustrations. The schema needed to serve three purposes. First it needed to enable the discovery, description and use of the identified images by artists, biologists, humanities scholars, librarians, and educators. Second, it needed to make BHL's metadata and images available to other platforms. Third, it needed to support the import of crowd sourced metadata generated in other platforms back into BHL. Project team members from MOBOT worked on developing the schema with project partners from the Ecology & Evolutionary Biology Department at the University of Colorado at Boulder (http://ebio.colorado.edu/), who were chosen for their expertise in biodiversity data standards and their participation in Wikimedia Commons. A landscape review was conducted of existing metadata standards suitable for the needs of natural history illustrations. The team focused in on five element sets: VRA Core, LIDO, Dublin Core, Darwin Core, and Audubon Core. VRA Core was found to be the best fit for the natural history illustrations because of its role as a standard for the "description of works of visual culture as well as the images that document them." Its elements and attributes were mostly closely aligned with the types of information we wanted users to record and it could express the many to one relationship, which parallel a book structure's multiple illustrations on a single page. Lacking in VRA Core was a way to record an accepted name and common name for a species - while it has a subject field with an attribute type for scientific name, taxonomists need more specificity. Therefore, staff borrowed two elements from the Darwin Core standard to fulfill this need.

Example of illustration using Art of Life schema.

Automatic identification of illustrations. Algorithms needed to be developed to locate the BHL pages containing illustrations. For this task, MOBOT enlisted project partners from the Indianapolis Museum of Art (IMA) Lab (http://lab.imamuseum.org/) who were chosen for their expertise with building open source applications for cultural heritage collections in particular, their creation of a tagging tool for the Steve.museum project (http://www.steve.museum/). For the Art of Life project, the IMA Lab developed four algorithms based on several characteristics of a digitized page: 1) Picture blocks, 2) Contrast, 3) Color, and 4) Compression. Picture blocks were approximate image coordinates generated by OCR software and Contrast involved comparing the contrast ratio of images vs. text on a page. In testing, both picture blocks and contrast characteristics were found to be much more effective at identifying pages containing illustrations (accuracy rates of between 87-88%) than were color and compression characteristics (accuracy rates between .09%-9%) Therefore, the latter two were dropped when running across the full BHL corpus.

Classification of illustrations. Team members from MOBOT worked with team members from Smithsonian Libraries (SIL http://library.si.edu/) to refine an existing tool they had developed called Macaw. SIL developed Macaw for adding page level metadata (e.g. numbering, volume info, etc) to digitized page scans as they were being uploaded into the BHL portal. SIL modified Macaw to meet the needs of the Art of Life project so that BHL staff could easily view, sort and classify the pages that were identified by the algorithms as having images. Each page was classified as belonging to one or more of five classes: drawing/painting/diagram; chart/table; photograph; map; or bookplate. Those classes allow Art of Life staff to have a better sense of the different types of illustrated content found on BHL pages and to determine which are most appropriate for further description. Classifiers can also indicate if a page does not contain an image, thereby providing a further verification of the accuracy or inaccuracy of the algorithmic output.

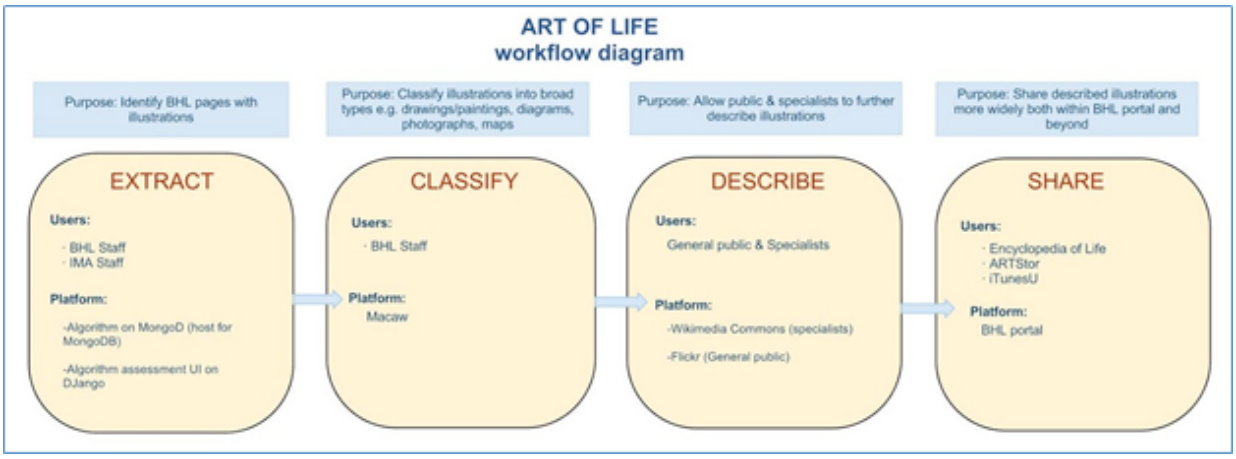
Crowdsourcing descriptions. Once classified, illustrated pages are then pushed into crowdsourcing environments where their content can be tagged and described. These environments include Flickr, where BHL has already manually identified and shared over 90k images (https://www.flickr.com/photos/biodivlibrary/sets/) and Wikimedia Commons http://commons.wikimedia.org/wiki/Commons:Biodiversity_Heritage_Library), which will

allow incorporation of templates such as the Art of Life schema for more in-depth descriptions.

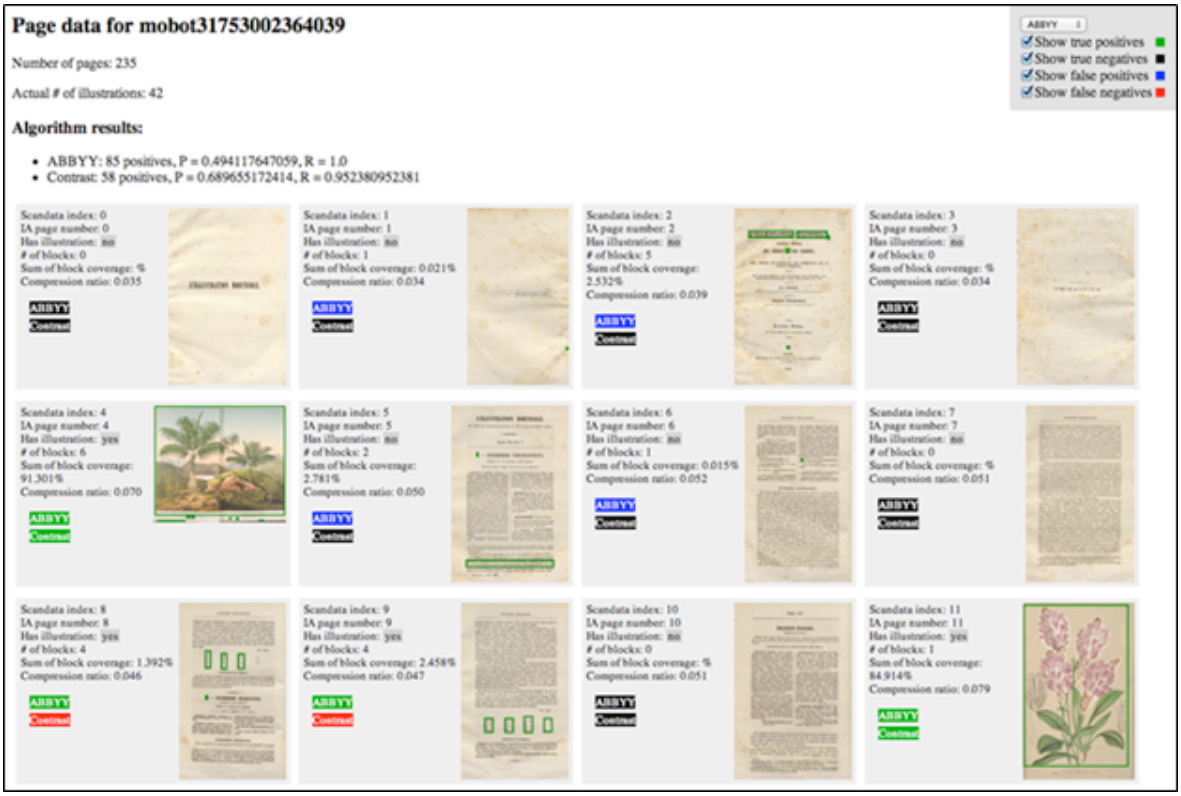
Ingest metadata back into BHL portal. Descriptions generated through the crowdsourcing platforms are ingested back into the BHL portal for access as well as preservation. Once stored in the BHL system, image metadata can be indexed and searched from within the portal. Both images and metadata are then preserved and can be shared more easily with other repositories such as EOL and ARTStor (http://www.artstor.org/).

Art of Life workflow and tools

The Art of Life project team developed a diagram which identifies the four processes the illustrations will go through as they move through each stage of the workflow (Extract, Classify, Describe, and Share); the tools or platforms needed at each stage, and who the users of those tools will be.



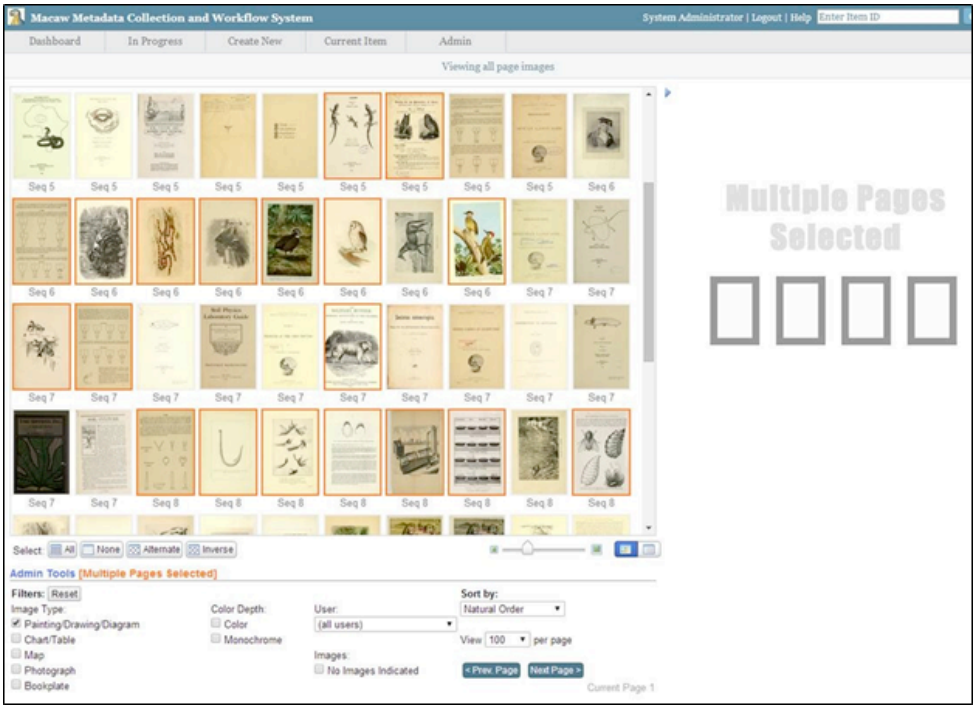
The extraction process identifies which BHL pages contain illustrations. So that BHL and IMA staff could determine which algorithms were most accurate, IMA built a Django web application for analyzing and visualizing the algorithm results.



Django web application for analyzing and visualizing the algorithm results.

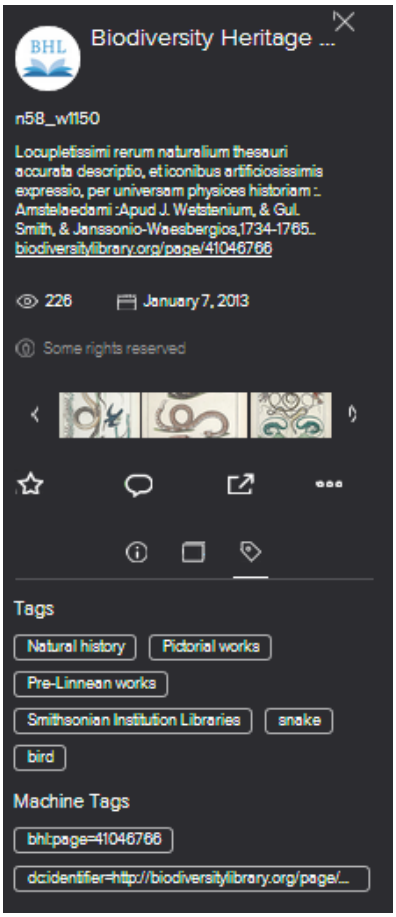
The most successful algorithms were run across a multi server environment and the output is stored in an open source document database called MongoDB. The tools and their documentation are available on github at https://github.com/IMAmuseum/artoflife.

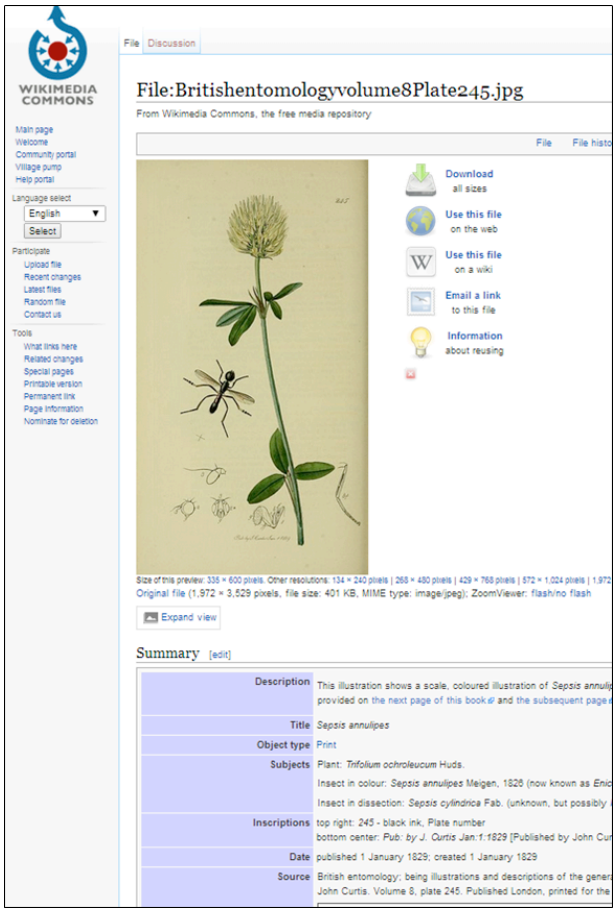
The classification process allows BHL staff to group pages identified by the algorithms into broad categories or types. SIL modified an existing open source tool called Macaw. This tool and its documentation is available on github at https://github.com/cajunjoel



Macaw tool for classifying illustrations.

The description process allows users to tag or describe in detail the content of the illustrations. Flickr and Wikimedia Commons are the two tools chosen for this task because they are crowdsourcing environments where large numbers of users are already sharing and marking up images and media files.





Left: Screen shot from the BHL Flickr site showing tags added to an image. Right: Screen shot from Wikimedia Commons showing use of the Art of Life schema template.

The share process involves bringing the metadata generated within the crowdsourcing platforms back into BHL for access and re-use. BHL is already sharing its images and metadata with biodiversity-related repositories such as EOL and intends to share these images more widely with image repositories such as ARTstor whose heaviest users are humanities scholars.

Current Status

As mentioned previously, the Art of Life project was scheduled to complete by April of 2014 but has been given an extension by NEH until April of 2015. Currently, staff are in the process of running the algorithms across the entire BHL corpus and have begun classifying the resulting pages that were found to have illustrations. Testing has begun to determine how to bulk upload large number of image files to Wikimedia Commons, as well as how to extract metadata back from the tool once the descriptions are crowd sourced. The BHL architecture has already been modified to be able to store and preserve metadata created at each stage of the workflow. The Art of Life schema is being converted to an application profile and will be

made publically available once complete. Interested parties can follow our progress through the project page http://biodivlib.wikispaces.com/Art+of+Life.

Benefits to Humanities and Science Scholars

Natural history is part of the human cultural heritage and is a rich source of knowledge for a broad spectrum of students, educators, the general public and scholars (Rinaldo & Smith 2014). Aside from their aesthetic qualities, detailed illustrations of plants and animals can be important, even today, for biologists tracing the taxonomic history of an organism or as documentation for lost or discarded specimens. Before the advent of photography, botanical and zoological artists were necessary partners for documentation of the natural world. Natural history illustrations provided a window to biodiversity around the world for scientists and the public who could not travel and richly illustrated volumes sold well. Thus artists and their work are integral to natural history publications (Blum 1993).

There are stories to be found in natural history illustrations. Natural history artists copied illustrations from the past and perpetuated inaccurate information or even tried to trick others into accepting them as "first" to illustrate a natural object: John James Audubon is one of these tricksters (see detailed discussion of Audubon's attempt to show his Ruffed Grouse illustration pre-dated Alexander Wilson's in Burtt & Davis 2013). Providing easy access to natural history illustrations benefits humanities scholars such as Janet Browne, who is interested in the visual and cultural history of the gorilla and other natural objects. Her research relies in part on historical illustrations along with preserved specimens and other media. Browne suggests that scientists used natural history illustrations as substitutes for specimens, because few scientists or non-scientists could travel and see the living objects (Browne 2011).

For humanities scholars, a significant resource of natural history images will be made openly accessible and reusable by completing the tasks from *The Art of Life*, providing new ways to aggregate previously hidden resources and study the historical and sociological relationships inherent in the collaborations of naturalists and artists (Blum 1993). There will be 1 million plus images and related catalog metadata that will be downloadable for free use and available in already familiar image platforms (e.g. ARTstor, Encyclopedia of Life). The content is cross-disciplinary. The BHL Flickr statistics and testimonials demonstrate that these images appeal to a wide range of audiences including artists, biologists, humanities scholars, particularly historians of science, librarians, education and outreach-- in other words, anyone who uses images in their research, business or teaching. Some of the more unusual uses of BHL images include decorating wedding invitations and greeting cards, as aids to digital collages, and to enhance fashion photography.

For the biodiversity scientist, this work will provide access to content in print or online repositories that has been difficult to discover. Functionality added to the BHL portal will allow searching for images by species name, common names, subjects, and illustrators. Expanded access will benefit users from students and teachers with minimal scientific training to scientists with more sophisticated training. Once the images are available and described in places like Flickr and Wikimedia Commons they will be easily linked to, and available in, other biodiversity and non-biodiversity-related platforms such as Wikispecies, Wikimedia and EOL. Like the textual content in BHL, the majority of the image content will fall under public domain and be freely available for download and re-use, enabling images to be incorporated into research, publications and teaching tools for schoolchildren as well as

scholars in many fields. The algorithm will be made available and can be used on any text collections with OCR output thus providing a more general benefit for other digital libraries. Additionally, the schema can be applied to other image collections that contain a large number of natural history illustrations.

Acknowledgments

The authors are grateful to everyone who contributed to this project: MOBOT (Trish Rose-Sandler, William UIate, Mike Lichtenberg, Mike Blomberg); IMA (Ed Bachta, Kyle Jaebker, Charlie Moad); CU-Boulder (Gaurav Vaidya); Smithsonian Libraries (Martin Kalfatovic, Joel Richard); Washington University (Chris Freeland); Advisory Board - Doug Holland, Director, Missouri Botanical Garden Library; Dr. Hong Cui, Assistant Professor, University of Arizona; Dr. David Kohn, Director and General Editor, Darwin Manuscripts Project, American Museum of Natural History; Charles Miller, Chief Information Officer, Missouri Botanical Garden; Nancy E. Gwinn, Director, Smithsonian Libraries; Robert Guralnick, Associate Professor at the University of Colorado at Boulder; Betty Smocovitis, Professor of Zoology and History at the University of Florida.

References

Blum AS, 1993, *Picturing Nature: American Nineteenth Century Zoological Illustrations*, Princeton University Press, Princeton, NJ.

Browne, J 2011, 'Illustrations as substitute specimens'. Paper presented at the *Life and Literature Conference*, Chicago, IL.

Burtt, EH & WE Davis 2013, *Alexander Wilson: The Scot who Founded American Ornithology*, Harvard University Press, Cambridge, MA

Gwinn, NE & Rinaldo, CA, 2009, 'The Biodiversity Library: Sharing biodiversity with the world', *IFLA Journal,* vol. 35, no. 1, pp. 25-34.

Rinaldo, CA & Smith, J 2014, 'Moving through time and culture with the Biodiversity Heritage Library', in *Migrating Heritage: Experiences of Cultural Networks and Cultural Dialogue in Europe,* ed P Innocenti, Ashgate Publishing Group, Surrey, pp. 95-108.