Expanding the Biodiversity Heritage Library

William Ulate R.
Global BHL Coordinator & BHL US/UK Technical Director
Center for Biodiversity Informatics (CBI), Missouri Botanical Garden

Florence, Italy
October 29, 2013
What is BHL?

The Biodiversity Heritage Library is a consortium of natural history and botanical libraries that cooperate to digitize and make accessible the legacy literature of biodiversity held in their collections and to make that literature available for open access and responsible use as a part of a global “biodiversity commons.”
New Partners and Geographies
Serving OLEF format in OAI/PMH

• OLEF: A format defined to facilitate metadata harmonization among BHL Partners by colleagues from BHL-Europe.

• http://www.bhle.eu/bhl-schema/v1/

• BHL-US/UK now serving OLEF format
Welcome to the Biodiversity Heritage Library, a consortium of natural history and botanical libraries that cooperate to digitize and make accessible the legacy literature of biodiversity held in their collections and to make that literature available for open access and responsible use as a part of a global “biodiversity commons.” BHL also serves as the foundational literature component of the Encyclopedia of Life (EOL).

Today's Picks View additional images on BHL's Flickr stream
Generate My PDF

If you are generating a PDF containing the text of a single journal article or book chapter, please provide title and author information.

BHL stores this information to allow these PDFs to be indexed, searched and retrieved by other users. If you download an article but do not provide title or author information, these articles will be lost.

Email Address (required)
william_ulate_r@yahoo.com

Article/Chapter Title
New Illustrations of the sexual system of C

Author(s)
Carl von Linnaeus

Optional

Example: Charles Darwin, Carl Linnaeus

Subject(s)
Flowers, Illustration

Example: Birds, Classification, Mammals

Review My PDF Finish
New Types of Content
New Types of Content
Scientific Name Extraction

• *TaxonFinder* algorithm in production since 2008
  – More than **100 million** candidate name strings
  – More than **1.5 million** unique, verified names
  – Available through UI, APIs, Data Exports & Internet Archive

• New collaboration with Global Names project
  – Improved algorithm, better precision & recall
  – More data with *TaxonFinder* and *Neti Neti*!
The fishes of Malabar / by Francis Day.

Scientific Names on this Page

Page 256

Cravacration fluviatilis Hamilton, 1822

Cravacration

Brachycephalus

Tetraodon

Tetraodon simulans Cantor, 1849

Tetraodon fluviatilis Hamilton, 1822

Contributed by Harvard University, MCZ, Ernst Mayr Library
<table>
<thead>
<tr>
<th>Type</th>
<th>Title</th>
<th>Authors</th>
<th>Volume</th>
<th>Date</th>
<th>Page #</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monograph/Item</td>
<td>The fishes of India:</td>
<td>Day, Francis, Achilles, C., Ford, G. H.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Griesbach, Carl Ludolf, King, J. R.,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mintern, R., Suzini, Mintern Bros.,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monograph/Item</td>
<td>The fishes of India:</td>
<td>Day, Francis, Achilles, C., Ford, G. H.</td>
<td>v. 1</td>
<td>1875</td>
<td>(go to page)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Griesbach, Carl Ludolf, King, J. R.,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mintern, R., Suzini, Mintern Bros.,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monograph/Item</td>
<td>The fishes of Malabar /</td>
<td>Day, Francis,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(go to volume) 1865</td>
<td>Page 256</td>
<td></td>
</tr>
<tr>
<td>Monograph/Item</td>
<td>The fishes of Malabar /</td>
<td>Day, Francis,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(go to volume) 1865</td>
<td>Page 256</td>
<td></td>
</tr>
<tr>
<td>Serial</td>
<td>Memoirs of the Indian Museum.</td>
<td>Indian Museum.</td>
<td>v. 5 no. 1-6</td>
<td>1907</td>
<td>(go to page)</td>
</tr>
<tr>
<td>Serial</td>
<td>Mitteilungen aus dem Naturhistorischen M</td>
<td>Naturhistorisches Museum in Hamburg.</td>
<td>bd. 21 (1903)</td>
<td>1884</td>
<td>Page 191</td>
</tr>
<tr>
<td>Serial</td>
<td>Mitteilungen aus dem Naturhistorischen M</td>
<td>Naturhistorisches Museum in Hamburg.</td>
<td>bd. 21-22 (190)</td>
<td>1884</td>
<td>Page 191</td>
</tr>
<tr>
<td>Serial</td>
<td>Sitzungsberichte der Kaiserlichen Akademie</td>
<td>Kaiserl. Akademie der Wissenschaften in W</td>
<td>1-64 index (1861)</td>
<td>Page 82</td>
<td></td>
</tr>
</tbody>
</table>
Find In BHL

Browse the bibliography for "Crayracion fluviatilis (Hamilton, 1822)".

Data Sources

EOL

Name: Crayracion fluviatilis (Hamilton, 1822)
Local ID: 220448

GBIF Taxonomic Backbone

Name: Crayracion fluviatilis (Hamilton, 1822)
Classification:
- Animalia (kingdom)
- Chordata (phylum)
- Actinopterygii (class)
- Tetraodontiformes (order)
- Tetraodontidae (family)
- Tetraodon (genus)
- Tetraodon fluviatilis (species)

ITIS

Name: Crayracion fluviatilis (Hamilton, 1822)
Classification: Animalia (kingdom)
Article-level metadata
Chapter-level metadata
Treatment-level metadata
Part-level metadata
Part-level metadata

• Disambiguating and *locating* structural components in the corpus
• Done by automated *and* crowdsourced means
  – Thanks Rod Page! Welcome others!
• Greatly increases semantic value of the dataset
• Addressing important – makes data *addressable*
  and thus *linkable*
Deduplication

Comparing Documents with Bayes Classification, Term Frequency–Inverse Document Frequency, and Levenshtein Distance Algorithms

• Mike Lichtenberg’s solution to automatically deduplicate titles

• Test titles for BioStor Article Titles and titles for BHL Titles

• http://bit.ly/1gXeLhX
See also:
# Related Titles

<table>
<thead>
<tr>
<th>Title</th>
<th>Details</th>
<th>MODS</th>
<th>BibTeX</th>
<th>Endnote</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>The auditory region of an Upper Pliocene typotherid /</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Series: Fieldiana. Geology ; v. 6, no. 5.

Series: Publication (Field Museum of Natural History : 1909) ; 331.

Series: Publication (Field Museum of Natural History : 1909). Geological series ; v. 6, no. 5

By

Patterson, Bryan.

Genre

Book

Publication info
Support citation reconciliation

Linnaeus, C. *Species Plantarum*, vol. 2 p. 971. 1753


Caroli Linnaei, *Species Plantarum exhibitentes plantas rite cognitas, ad genera relatas, cum Differentis Specificis, Nominibus Trivialibus, Synonymis Selectis, Locis Natalibus, secundum SYSTEMA SEXUALE digestas..* 2:971. 1753

L. *Sp. Pl.* 2: 971. 1753

*Zea mays*
Macaw

Macaw Metadata Collection and Workflow System
Version 1.6.358 / April 30, 2012

Based on the Paginator originally created at the Missouri Botanical Garden.

Viewing Activity

<table>
<thead>
<tr>
<th>Barcode</th>
<th>Title</th>
<th>Author</th>
<th>Organization</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>39088011499621</td>
<td>Icones omithopterorum</td>
<td>Rippon, Robert H. F. (Robert Henry Fernando)</td>
<td>Default</td>
<td>reviewing</td>
</tr>
<tr>
<td>Test-Book-Identifier</td>
<td>Testing title of the volume</td>
<td></td>
<td>Default</td>
<td>new</td>
</tr>
</tbody>
</table>
Loading Activity
Uploading images via browser
Uploading images via browser

Instructions

1. Scan all of your pages.

2. Copy the scans to this network location:
   
   /var/www/joeirichard.com/macaw/docs/incoming/Test-Book-identifier

3. Start the import:
   
   Start Import  Skip Import

While the scans will be copied to the server, you can monitor the progress on the right.
Reviewing Metadata
Reviewing Metadata
Uploading to the Internet Archive

• Need to get set up with an account at IA first

• Account at IA needs access to the biodiversity collection

• Uploading of completed items is done via scheduled job or the command line
Currently working on...

- RefBank installation to share BHL citations.

- Solr installation to enable Full-text search.

- As part of the Global Names Architecture project, installing a local Global Name Usage Bank (GNUB) instance to help in deduplication.
What we’d like to do

http://biodivlib.wikispaces.com/BHL+and+Gaming

Researching candidate Scientific Names

Image identification & extraction

– http://biodivlib.wikispaces.com/Art+of+Life

– Currently funded by NEH

• Improve OCR

• Rekeying Tables of Contents
Challenges with OCR of Older material

- Great deal of material is pre-1923
- Irregular fonts – blackletter
- Multiple languages on same page – English text with Latin scientific names
- Changes in geographic names
- Changes in scientific names
LECTORI.

diff Nob. Collinian, Torèn, Braud, aliisque. Clifford de
mensquas in duplo'habiliit: Lagerfron indiae orientalis
plurimas: Gnemovius pleraque Virginitas & Gmelinus
Sibericas facile omnes, at S. auv age /i us integrum
ulnerarum plantarum aedit, exemplo raro & inaudit-
to, unde factum, ut copiam plantarum obtinuern
mediocrem.

DIFFERENTIAS Specificas antheac Plantis non pa-
cis impofui in Flora Lapponica, Suecica, Leylanca, in
Horto Cliffortiano, Uplandia, lisdem principis in
fiteri praeconifimi Botanici Gmelinius, Royen, H
thenoff, Gatter, B. Juifinus, Le Montier, Gauffard,
Dalhard, Siliesius, Cohnen, Hill, inque non paucis
Hallerus, Gmelinii, aliique, quorum opera plurima
Species evadere rase & evidentiae.

NOMINA hanc paffim difperfa in gratiam studioforum
conftantium duxi colligere, acqüitias dein plantas ade-
dere, inque unum, Systema redcre, at vero pluribus
Species observatis, praefiantibus notis detectis,
aptioribus Terminus cuitus, debui interdum differentias
emendare, antea quamvis optimas.

Charafters essentielle pro NOMINE SPECIFICO
conftituere non levidentis opus eft, requirit enim Spe-
rum plurimis accufatam cognitionem, harum atten-
tifimam partium indagationem, differentiam fecleto-
nem, Terminorum denique artis propriam applicationem,
vit evadant comphendiufisti tumfulgium.

NON VISAS plantas heic omifit, toties elufus ab au-
toribus, ne dubia certiffms miscrener, fi vero al-
qundo conterger non sufficiemer infepfes plantam,
vel fpecimen imperfectum obtinuueris, figno hoc no-
tavi, ut ali candem accuratius examiner.

Non
OCR Improvements

• Transcription
• Purposeful Gaming

• Looking at...
  – Crowdsourcing Markup
Purposeful Gaming

Zooniverse
Real Science Online

Space
- Galaxy Zoo
- Moon Zoo
- Solar Stormwatch
- Planethunters.org
- The Milky Way Project
- Planet Four

Climate
- Old Weather
- Cyclone Center

Humanities
- Ancient Lives

Nature
- Bat Detective
- Whalefm
- Seafloor Explorer

Biology
- Cell Slider
Purposeful Gaming

DIGITALKOOT

• Joint project run by the National Library of Finland and Microtask to index the library's enormous archives so that they are searchable on the Internet for easier access to the Finnish cultural heritage.
Purposeful Gaming

DIGITALKOOT

- Launched on Feb 8 2011, nearly 110,000 participants completed over 8 million word fixing tasks by Nov 29 2012
- DigiTalkoot enabled volunteers to participate in this fixing work by playing games.
Purposeful gaming and BHL: engaging the public in improving and enhancing access to digital texts

• IMLS Grant Program: National Leadership Grants for Libraries
• Partners:
  – Missouri Botanical Garden
  – Harvard University
  – Cornell University
  – New York Botanical Garden
• P.I.: Trish Rose-Sandler, Missouri Botanical Garden
• Dates: Dec 2013 – Nov. 2015
Project objectives and benefits

• Test new means of crowdsourcing to support the enhancement of content in BHL

• Demonstrate if digital games are an effective tool for analyzing and improving digital outputs from OCR and transcription

• Benefits of gaming include:
  – improved access to content by providing richer and more accurate data;
  – an extension of limited staff resources; and
  – exposure of library content to communities who may not know about the collections otherwise.
OCR Improvements

German text interpreted by the OCR process as:
“unb auf ben Â©elnrgeren be6 fublic{}en”
Different resulting texts from parsing the phrase: “und auf den Gebirgen des südlichen Deutschlands”  
(“and on the mountains of southern Germany”)
Purposeful Gaming

Workflow Diagram for Processes and Decision Points

Digitize → Generate OCR 1 (Internet Archive) → Automated Quality Assessment

Clean-up BHL corpus

Fine-tune Text Correction Process

Evaluate accuracy scores for a sample of assembled corrected text

Assemble corrected text output

Add to Ground Truth files

Model Correction

Run Game to Gather Corrections

Push Differences to the Game

Acceptable accuracy score?

NO → Discard?

YES → Manual Review

Transcribe (2 outputs)

Generate error matrix

Generate OCR 2 (MOBOT)

> 20% non-alpha characters?

NO

> 1 OCR output?

YES

NO

Manual Review
<table>
<thead>
<tr>
<th>Display text</th>
<th>Species Profile Model category</th>
</tr>
</thead>
<tbody>
<tr>
<td>General/summary</td>
<td>TaxonBiology</td>
</tr>
<tr>
<td>Geographic range</td>
<td>Distribution</td>
</tr>
<tr>
<td>Habitat</td>
<td>Habitat</td>
</tr>
<tr>
<td>Food sources and feeding behavior</td>
<td>TrophicStrategy</td>
</tr>
<tr>
<td>Physical description (general)</td>
<td>Description</td>
</tr>
<tr>
<td>Physical description (detailed morphology)</td>
<td>DiagnosticDescription</td>
</tr>
</tbody>
</table>
Thank you

William Ulate
Global BHL Project Manager / Technical Director
Missouri Botanical Garden
william.ulate@mobot.org
Skype: william_ulate_r