Linked Open Data and Systematic Taxonomy

A tale of two publications
In three acts

Joel Richard
Smithsonian Libraries
richardjm@si.edu

CENDI/NKOS, December 6, 2012
Who are the Smithsonian Libraries?

- 20 Libraries in the U.S. and Panama
- Supports research of staff and the public
- Strong effort to digitize pre–1923 texts
- Index Animalium and Taxonomic Literature II are two examples
We are still learning.
We are still building.
Act I: The Players

(or, identifying the data with which we are working and their meaning and usefulness to the scientific community.)
Taxonomic Literature II

Essential Reference Tool for Botanists

Botanists/Authors and Publications from 1753–1940

Multiple indexes, “unique identifiers”

It is a “database in book form”

Joel Richard, richardjm@si.edu
<table>
<thead>
<tr>
<th>Author Name</th>
<th>Darwin, Charles Robert (1809-1882)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TL-2 Number</td>
<td>1313</td>
</tr>
<tr>
<td>Title</td>
<td>On the origin of species</td>
</tr>
<tr>
<td>Year(s)</td>
<td>1859</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Author Name</th>
<th>Darwin, Robert Waring (1724-1816)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TL-2 Number</td>
<td>1314</td>
</tr>
<tr>
<td>Title</td>
<td>Principia botanica</td>
</tr>
<tr>
<td>Year(s)</td>
<td>1787</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Author Name</th>
<th>*Darwin, Charles Robert (1809-1882)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TL-2 Number</td>
<td>29,064</td>
</tr>
<tr>
<td>Title</td>
<td>On the two forms, or dimorphic condition, in the species of Primula</td>
</tr>
<tr>
<td>Year(s)</td>
<td>1862</td>
</tr>
</tbody>
</table>
Index Animalium

Genus name, author & citation for 430,000 animals

Covers Publications from 1758–1850

Also a database, but many challenges still exist in the data.
Platius plicatella

Platius M. T. Rothe, 1771, 78—A.
plebeja Hesperia, J. C. Fabr. ex. Syst. Nat. III. (1) 1775, 99—A.
plebeja x Papilio.

plebeja Bathe (Ex.), J. C. Fabr. ex. Syst. Nat. III. (1) 1775, 776—A.
plebeja Melichona, A. G. Olivier, Entom. L. 1799, 32.
plebeja Phalaena (Noctua), Linn., Syst. Nat., ed. 10, L. 1758, 505—A.
plebeja Phalaena (Noctua), Linn. Syst. Nat., ed. 10, L. 1758, 505—A.
plebeja Phalaena (Noctua), Linn. Syst. Nat., ed. 10, L. 1758, 505—A.
plebeja Phalaena (Noctua), Linn. Syst. Nat., ed. 10, L. 1758, 505—A.
plebeja Phalaena (Noctua), Linn. Syst. Nat., ed. 10, L. 1758, 505—A.
plebeja Phalaena (Noctua), Linn. Syst. Nat., ed. 10, L. 1758, 505—A.
plebeja Phalaena (Noctua), Linn. Syst. Nat., ed. 10, L. 1758, 505—A.
plebeja Phalaena (Noctua), Linn. Syst. Nat., ed. 10, L. 1758, 505—A.
plebeja Phalaena (Noctua), Linn. Syst. Nat., ed. 10, L. 1758, 505—A.
plebeja Phalaena (Noctua), Linn. Syst. Nat., ed. 10, L. 1758, 505—A.
plebeja Phalaena (Noctua), Linn. Syst. Nat., ed. 10, L. 1758, 505—A.
plebeja Phalaena (Noctua), Linn. Syst. Nat., ed. 10, L. 1758, 505—A.
plebeja Phalaena (Noctua), Linn. Syst. Nat., ed. 10, L. 1758, 505—A.
plebeja Phalaena (Noctua), Linn. Syst. Nat., ed. 10, L. 1758, 505—A.
plebeja Phalaena (Noctua), Linn. Syst. Nat., ed. 10, L. 1758, 505—A.
Act II: The Linking

(or, identifying those data elements to be linked, inherent challenges of parsing OCR text, and identifying linkable remote data sources)
Linkable Data Elements

Darwin, Charles Robert (1809-1882), British evolutionary biologist. (Darwin).


plexippus Papilio (Danaus) Linnaeus, Syst. Nat. ed. 10, 1758, 471 ed. 12, 1767, 767.
Challenges with Our Data

• Errors in the Corrected OCR
• Challenges in Parsing Citations
• The 80/20 rule: manually making connections unable to be made by automated means
• Finding suitable sources of data to link to. (DBPedia? VIAF? EOL? Others?)
Linked Data Sources

Low-Hanging Fruit:

- DBPedia
- OCLC WorldCat
- Biodiversity Heritage Library
- Virtual International Authority File
- Encyclopedia of Life
- Library of Congress Subject Headings
- GeoNames
- Open Library
Act III: The Sum of the Parts

(or, our goals and desires for this data, what it means to the linked data world and the scientific community in general)
What’s the point?

• This data may already exist online.
• It may also not always be as accurate as needed for science.
• We are in a position to be the authoritative source for this information.
• Linked Data allows it to be easily reused and shared.
Danaus plexippus

Aimeé Antoinette Camus (botanist)
One Example of Reuse

Ryan Schenk
http://synynyms.com/

Cavia porcellus
Cavia aperea porcellus
Cavia cobaya
Thank you!

Joel Richard
RichardJM@si.edu
http://library.si.edu/staff/richardjm
http://slideshare.net/joelrichard