Digitization in a global environment

Donald Hobern
GBIF Director
Global Biodiversity Information Facility (GBIF)

Biological Collections Digitization in the Pacific Workshop

Honolulu, 25 March 2014
GBIF: origins and principles

Established in 2001
Response to OECD recommendation
Open to all countries
Voluntary memorandum of understanding (MoU)

Vision - a world in which:

Biodiversity information is freely and universally available for science, society and a sustainable future
12 years – 417 million records
Weaknesses with distributed data

Id: CURC-00015436
Species: Acalles dubius
Location: Berlin, Germany
Latitude: 13.41
Longitude: 52.52

Curator

Digitises
Weaknesses with distributed data

- **Distribution Modeller**: Uses
- **Environment Agency**: Rejects
- **Climate Change Researcher**: Corrects

**Curator**
- Digitises
  - Id: CURC-00015436
  - Species: *Acalles dubius*
  - Location: Berlin, Germany
  - Latitude: 13.41
  - Longitude: 52.52

**Id: CURC-00015436**
- Species: *Acalles dubius*
- Location: Berlin, Germany
- Latitude: 13.41
- Longitude: 52.52

**Id: CURC-00015436**
- Species: *Acalles dubius*
- Location: Berlin, Germany
- Latitude: 13.41
- Longitude: 52.52

**Id: CURC-00015436**
- Species: *Acalles dubius*
- Location: Berlin, Germany
- Latitude: 13.41
- Longitude: 52.52
Weaknesses with distributed data

- **Distribution Modeller**
  - Uses
  - Id: CURC-00015436
  - Species: *Acalles dubius*
  - Location: Berlin, Germany
  - Latitude: 13.41
  - Longitude: 52.52

- **Environment Agency**
  - Rejects
  - Id: CURC-00015436
  - Species: *Acalles dubius*
  - Location: Berlin, Germany
  - Latitude: 13.41
  - Longitude: 52.52

- **Climate Change Researcher**
  - Corrects
  - Id: CURC-00015436
  - Species: *Acalles dubius*
  - Location: Berlin, Germany
  - Latitude: 13.41
  - Longitude: 52.52

- **Curator**
  - Digitises
  - Id: CURC-00015436
  - Species: *Acalles dubius*
  - Species: *Acalles camelus*
  - Location: Berlin, Germany
  - Latitude: 13.41
  - Longitude: 52.52

- **Taxonomist**
  - Reidentifies
  - Id: CURC-00015436
  - Species: *Acalles dubius*
  - Location: Berlin, Germany
  - Latitude: 13.41
  - Longitude: 52.52
Toward a shared knowledgebase

Id: CURC-00015436
Version: 1
Species: Acalles dubius
Location: Berlin, Germany
Latitude: 13.41
Longitude: 52.52

Global Biodiversity Knowledgebase
Toward a shared knowledgebase

Global Biodiversity Knowledgebase

Id: CURC-00015436
Version: 2
Species: Acalles dubius
Location: Berlin, Germany
Latitude: 13.41
Longitude: 52.52

Issue detected: Locality

Distribution Modeller
Uses

Environment Agency

Flags issue

Climate Change Researcher

Digitises

Curator

Taxonomist

www.gbif.org
Toward a shared knowledgebase

- Distribution Modeller
  - Uses
  - Digitises

- Curator
  - Digitises

- Environment Agency
  - Flags issue

- Climate Change Researcher
  - Corrects

Global Biodiversity Knowledgebase

Id: CURC-00015436
Version: 3
Species: Acalles dubius
Location: Berlin, Germany
Latitude: 52.52
Longitude: 13.41
Toward a shared knowledgebase

- **Distribution Modeller**: Uses
- **Environment Agency**: Flags issue
- **Climate Change Researcher**: Corrects
- **Curator**: Reidentifies
- **Digitises**: **Taxonomist**

**Global Biodiversity Knowledgebase**

Id: CURC-00015436
Version: 4
Species: *Acalles camelus*
Location: Berlin, Germany
Latitude: 52.52
Longitude: 13.41
Toward a shared knowledgebase

Global Biodiversity Knowledgebase

Id: CURC-00015436
Version: 6
Species: *Acalles camelus*
Location: Berlin, Germany
Latitude: 52.52
Longitude: 13.41
Sequence: ATTGCA...
Image: 15436.jpg
Toward a shared knowledgebase

Global Biodiversity Knowledgebase

- Distribution Modeller
  - Flags issue
  - Corrects

- Environment Agency
  - Flags issue
  - Corrects

- Climate Change Researcher
  - Corrects

- Technician
  - Sequences
  - Publishes model

- Curator
  - Digitises

- Taxonomist
  - Reidentifies

- Photographer
  - Images

Distribution model

Acalles camelus

Id: CURC-00015436
Version: 6
Species: Acalles camelus
Location: Berlin, Germany
Latitude: 52.52
Longitude: 13.41
Sequence: ATTGCA...
Image: 15436.jpg
Global Biodiversity Informatics Outlook
Focus Area: Culture

- The context for sharing digital knowledge
  - Data must be **available for reuse**
  - Data must **follow standards** to support discovery and use
  - Data must be **preserved for future uses**
  - **Policies and practices** must reinforce open use
  - The whole community should collaborate to **curate data**

- Issues shared in common with all research domains

- Investments here will multiply value of other components
Focus Area: Data

The streams of primary biodiversity data
- Literature and journals
- Natural history collections
- Professional and amateur field observations and surveys
- Molecular sequencing
- Remote sensing (including camera traps, acoustic monitoring, etc.)

All deliver fundamental observations and measurements of biodiversity
Foundations for analysis and understanding
Focus Area: Evidence

- Organised views of biodiversity data
  - Consistent assessment of quality and fitness-for-use
  - Comprehensive digital nomenclature and taxonomy
  - Access to all evidence for recorded species occurrence
  - Access to species traits, measurements and interactions
  - Services and interfaces to access data as needed
- Provide comprehensive organised views of all relevant data
- Act as a “lens” into primary data
Focus Area: Understanding

• The application of data to address questions
  – Integrate data into spatial models
  – Develop temporal analyses
  – Incorporate biological reality into models
  – Present compelling representations of biodiversity
  – Optimise future investment in biodiversity informatics

• Data-driven models for science and planning
• Integrate biodiversity with other research and data domains
Example for occurrence data

**Data**
Dataset A asserts that species X was recorded at a given locality on a given date

**Evidence**
Community assessment concludes that species X was recorded at a given set of localities on given dates

**Understanding**
The best available model presents the probability that species X was present at any locality on any date
## Collections and biodiversity data

<table>
<thead>
<tr>
<th>Virtual natural history collection</th>
<th>Ecoinformatics resource</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Focus:</strong> collection objects</td>
<td><strong>Focus:</strong> geospatial data</td>
</tr>
<tr>
<td><strong>Goal:</strong> integrated access to all collection materials for any species</td>
<td><strong>Goal:</strong> maximise reliable evidence for species occurrence in time and space (collections are core data)</td>
</tr>
<tr>
<td><strong>Primary data elements:</strong></td>
<td><strong>Primary data elements:</strong></td>
</tr>
<tr>
<td>• Identification history</td>
<td>• Scientific name</td>
</tr>
<tr>
<td>• Collection identifiers</td>
<td>• Coordinates</td>
</tr>
<tr>
<td>• Locality</td>
<td>• Date</td>
</tr>
<tr>
<td>• Images, sequences</td>
<td>• Confidence/evidence level</td>
</tr>
<tr>
<td>• Morphometric data</td>
<td></td>
</tr>
<tr>
<td><strong>Critical linkages:</strong></td>
<td><strong>Critical linkages:</strong></td>
</tr>
<tr>
<td>• Nomenclators</td>
<td>• Climate and environment</td>
</tr>
<tr>
<td>• Biodiversity Heritage Library</td>
<td>• Politics and land use</td>
</tr>
<tr>
<td>• Barcode of Life Database</td>
<td>• Stable classification</td>
</tr>
<tr>
<td>• Phylogeny data</td>
<td></td>
</tr>
</tbody>
</table>

**Two major applications**
Planning digitization

• Understand content of collections
  • Collection type: wet, dry, pinned, tissue, living, ...
  • Factual: size, taxa, geography, types, curation status, subcollections, ...
  • Relevance: time series, protected areas, threatened species, indicator taxa, ...
  • Planning: digitisation options, costs/benefits

→ Metadata for discovery and publicity
→ Inputs to institutional/national strategy
→ Opportunity for collaborative funding or labour
Publishing data

Darwin Core Archive Components

- TypesSpecimens.txt
- VernacularNames.txt
- Basis of Resource (Taxon or Occurrence)
- speciesProfile.txt
- References.txt
- Distribution.txt

Extensions

- Metadata Document
- DarwinCore "Star Schema"

Darwin Core Archive

- GBIF’s preferred standard for sharing data
- ZIP file with data spreadsheet and metadata
Integrating data

GBIF data for “United States” intersecting USA

All GBIF data for “United States”
Collaborative curation

Great Crested Newt (Denmark)

Observed by Donald Hobern on 14th April 2013

(Added to iSpot on 14th April 2013)

Apparently dried out on gravel path.

Location: 2, Øræhof Alle

Identifications

Smooth Newt (Lissotriton vulgaris) by Donald Hobern at 7:25 pm 14/04/13
Confidence: I'm as sure as I can be.

@ Search Encyclopedia of Life for Lissotriton vulgaris
W Search Wikipedia for Lissotriton vulgaris
View NBN map for Lissotriton vulgaris

Great Crested Newt (Triturus cristatus) by Masked Marvel at 10:23 pm 14/04/13
Confidence: I'm as sure as I can be.
Notes: The black throat, deep orange belly and orange stripe along the bottom the tail are characteristics of a great crested newt.

I agree!

ID agreements (✓): 4 people agree with this identification.
Collaborative curation

Following the Systematics and taxonomy of Australian Birds (and also Clements/eBird), *R. fuliginosa* is confined to New Zealand. All else is now considered *R. albiscapa*. If your observations are from New Zealand, we recommend you reclassify them as the narrower concept of *R. fuliginosa*, otherwise reclassify them as *R. albiscapa*.

Source: Christidis, Les, Boles, Walter (2005). (Citation)
Added by iNaturalist on 2013-04-09 | Committed by iNaturalist on 2013-04-10

**Rhipidura fuliginosa** 8161
6 Obs | LC | Range | Inactive | Flag for curation

**Rhipidura albiscapa** 144737
20 Obs | Range | Active | Flag for curation

**Rhipidura fuliginosa** 244276
3 Obs | Range | Active | Flag for curation

Map of Australia showing distribution of species.
Data Publishers
Biodiversity data in institutional databases
Stable record identifiers

Access Portals
iDigBio, GBIF, BISON
Normalised and verbatim data
Validation and annotation
Integrated access
"EVIDENCE"

Data Repository
Dryad, Data ONE, GBIF IPT
Stable location, digital licence, annotation tools, data papers
Stable dataset identifiers
"DATA"

Spreadsheet Data
(DwC-A / CSV / Excel / XML)
Data and metadata for free and open use
Standards compliance

Data Publishers
Biodiversity data in institutional databases
Stable record identifiers

Darwin Core Archive Datasets (with DOIs)
GBIF in the landscape

National Agencies
Research Community
Intergovernmental Mechanisms
Other Biodiversity Informatics Organisations
(Encyclopedia of Life, Catalogue of Life, Barcode of Life, etc.)

GBIF
National Biodiversity Information Facilities
Thank you

Donald Hobern
GBIF Director
Global Biodiversity Information Facility (GBIF)

Biological Collections Digitization in the Pacific Workshop

Honolulu, 25 March 2014