



#idigtorch

Sharing Data

Data Extraction and Identifiers

24 May 2014

TORCH VIII + iDigBio Digitization Workshop

Deborah Paul, on Twitter @idbdeb @idigbio

Sul Ross State University, Alpine, Texas



iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.



Care and feeding of (clean) data

- (Getting data into your database)
- Getting the data out of your database
- “Mapping” your data to standard terms
- Your objects need identifiers
- Semi-automated data-sharing
- What to expect – data feedback
 - data quality issues
 - data enhancement
- It’s a partnership
 - we are all custodians of these new digital resources
 - we are care-takers of the data, stewards
 - integral to care of the physical specimens

Fit-for-research-use-data

R. K. GODFREY HERBARIUM
202973
FLORIDA STATE UNIVERSITY

Valdosta

Collecting Data

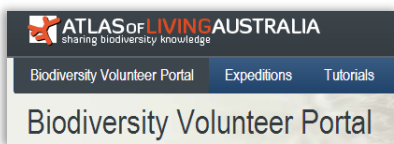
Identifiers are like Elvis, ...
or *Drosophila melanogaster*

Specify 6

FilteredPUSH



Encyclopedia of Life



GenBank



http://morphbank.net

Identifiers

Keep

IFE

Symbiot

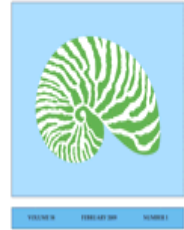
Systematic Biology

ZooKeys



VerNe

GBIF



Silver COLLECTION

Cladistics®

BMC Bioinformatics

vizzuality PhytoKeys



Identifier types

- Content-rich identifiers contain information
 - Simple identifier, attached to specimen
 - Number stamped on band: **1154**
 - **Catalog number**
 - Compound identifier: globally unique
 - **Darwin core triple** (institution, collection, catalog number)
 - (**BOTF, finch, 1154**)
 - URI (uniform resource identifier): internet friendly
 - **<http://beaks.org/finch/1154>**
- **Content-free identifiers**
 - Cannot be parsed, remembered or typed
 - **urn:uuid:40c842c9-c04c-489a-b20e-d84bfc16dedd6**

YPM ENT. No.

815664

Entomology Division
Peabody Museum

USA: Alaska, woods
near Kenai National
Wildlife Refuge head-
quarters building
60.4618°N 151.0806°W
02.Sep.2010. Matt
Bowser. KNWR: Ento: 10036

KNWR1254



AM_ENT



AMNH_PBI 00388325



SEMC0993403

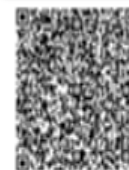
KUNHM-ENT

MCZ-ENT

00107550



Nymphalidae: Nymphalinae:
Epiphilini
Epiphile iblis plusios
Godman & Salvin, [1883]
56.22



{ "f": "Nymphalidae", "b": "Nymphalinae", "t": "Epiphilini",
"g": "Epiphile", "s": "iblis", "u": "plusios", "i": "null", "r": "null",
"a": "Godman & Salvin, [1883]", "d": "56.22" }

Identifying Objects



ID	1565	TSN	176580		
ACQUO			228.0	IDORSAL	<input type="checkbox"/>
CATNO				VENTRAL	<input type="checkbox"/>
SCIENTNAME	Scolopax minor				
Accepted	Scolopax minor				
COMMONNAME	American Woodcock				
Accepted	American Woodcock				
SEX		Subspecies			
MONTH	01				
DAY	04				
YEAR	1962				
COLLNAME	Stoddard, Sr.				

Record: 14 of 361 of 3945

<urn:uuid:b1495230-ac34-42ea-b6b7-7af8b9f1b212>

- Add column to data record for a globally unique, persistent identifier.



- UUID or GUID does not have to appear on the specimen itself.

Resolver

<http://www.talltimbers.org/museum.html#Birds:279>
<urn:uuid:3Ab1495230-ac34-42ea-b6b7-7af8b9f1b212>

Maintaining and Sharing Identifiers

- apply a given id to only one object ever
- if something happens and that object no longer exists in the physical collection –
 - never reassign the identifier to another object in the collection
- missing numbers do not matter
- (share the ones you know about)

- Now that we have identifiers ... what next?

Sharing requires standard terms



My field notes?

Your field notes?

map to a standard!

Sharing Requires Standard Terms

Darwin Core Location Terms

- higherGeography
- waterbody, island, islandGroup
- continent, country, countryCode, stateProvince, county, municipality
- locality
- minimumElevationInMeters, maximumElevationInMeters, minimumDepthInMeters, maximumDepthInMeters

Darwin Core Event Terms

- habitat

Darwin Core Geological Context

- group, formation, member, bed, ...

Darwin Core Standard

- Darwin Core (often abbreviated to DwC) is a body of data standards which function as an extension of [Dublin Core](#) for [biodiversity informatics](#) applications, **establishing a vocabulary of terms to facilitate the discovery, retrieval, and integration of information about organisms, their spatiotemporal occurrence, and supporting evidence housed in biological collections.** It is meant to provide a stable standard reference for sharing information on biological diversity^[1]
- Does Darwin Core cover every field possible? – No
- Don't panic! There are extensions and other standards.

mapping

WAKULLA CO.: St. Marks Nat'l Wildlife Refuge (Panacea Unit). Frequent in moist roadside depression, less so in drying sand of burned, open longleaf pine along W side Rte 372, just N of Rd 401 and 1.4 mi drive from Hwy 98.

field notes / Excel

- 41 05 54S
- 121 05 34W
- WGS84
- 2 mi. NE Tlh. on Ctrville Rd.
- Tallahassee, 2.5 miles NE on Centerville Road.
- frequent
- Wakulla.
- in moist roadside depression, ...

your database field

- **lat** or **latitude**
- **lon** or **long** or **longitude**
- **datum** or **notes** or ...
- **loc** or **location** or **collectorLocality** or ...
- **abundance**
- **county**
- **hab** or **habitatDescription** or ...

darwin core

- verbatimLatitude
- verbatimLongitude
- verbatimSRS
- verbatimLocality
- locality
- (abundanceAsPercent)
- county
- habitat

Note

Data Mapping & Export

Herbarium A

barcode
collectorNumber
collector

Herbarium B


accessionNumber
collectorNum
collectedBy

Darwin Core

catalogNumber
recordNumber
recordedBy

- All mapped up and ready to go – now what?

Data Export Example.

- How do you get your data out of your  database?
 - Schema Mapper tool
 - Data Exporter tool > creates a temporary table in your database
 - Data Exporter > tab-delimited text file for import into IPT
 - Install IPT, Register at GBIF using the IPT
 - Use the text file with the IPT for upload to GBIF, some mapping may be required
 - Publish your data
- Extensions for more data types: e.g. Audubon Core for Media files

Data Export



Symbiota

- General users download occurrence data from search page as Darwin Core CSV files or raw Symbiota
- Data managers
 - create backup file as a compressed set of Symbiota CSV files (occurrences, determination history, and image links)
- IPT instances are set up for the portals on the Symbiota servers (Lichens, Bryophytes, SCAN, MycoPortal, SCNet).
 - each collection can choose to send data to GBIF themselves or
 - via the portal.
- Future: Symbiota
 - automated packaging of data as Darwin Core archive files.
 - Control panel, collection managers refresh the DwC archive whenever they wish.
 - the ability to turn on or off publishing.

Data Export



- Each NHM client
 - initial mapping process with EMu staff
 - mapping to DwC 1.2 (aka v2)
- use Automated Export to create desired file
 - CSV
 - text
 - Crystal Report
- use DwC CSV file with IPT to create DwC-A file
- DwC-A file is shared with GBIF
- GBIF – harvests periodically

One Pathway to sharing?

- Discoverability
 - **Identifiers** are key
 - **Metadata** is key
 - for use / re-use / re-purpose
- Data in more than one place
 - + Aids discoverability
 - - Can be a issue to track
 - Identifiers help
- Dataset identifiers too

More Ways to Share Data

- Thematic Collection Networks (TCNs)
 - have data ready to share?
 - fits a current TCN theme?
- Partners to Existing Networks (PENs)
 - join the effort
- Through an existing portal or repository
 - Symbiota
 - Many portals to choose from
 - VertNet
 - Morphbank
 - GBIF
- Help is everywhere!

ARE YOU COMING TO BED?

I CAN'T. THIS
IS IMPORTANT.

WHAT?

SOMEONE IS WRONG
ON THE INTERNET.



Sharing Brings Opportunities, Benefits, Decisions

- Revisualization - discovery
- Unforeseen errors / relationships revealed
 - Recent Morphbank – SERNEC Symbiota Portal example
- Taking it in stride
- Specify's Scatter-Gather-Reconcile (SGR)
 - duplicates found
 - dataset for checking is increasing in size
 - example, at least 50% dupes
 - lending credence to the skeletal dataset concept
- Filtered PUSH (works with Specify, Symbiota and Morphbank)
 - finding dupes
 - the benefits of shared datasets
 - enhancing the skeletal (or short) record
 - finding annotations (determinations, general comments)
 - to import or not to import
- **More Loans, Exposure for your collection, \$**

Where do the standard terms come from?

- **Biodiversity Information Standards (TDWG)**

- formerly known as
 - Taxonomic Databases Working Group (TDWG)
 - began 1985

- **Our Mission**

- Develop, adopt and promote **standards** and **guidelines** for the **recording** and **exchange** of **data about organisms**
- **Promote** the **use of standards** through the most appropriate and effective means and
- Act as a **forum for discussion** through holding meetings and through publications

Identifiers

- Lots of uncertainty in community
 - What form of identifiers, what services to provided, etc.
- We need to
 - Emphasize identification of specimens and other objects
 - Help providers to see value of specimen identifiers
 - Remove obstacles to adoption
 - E.g. validate and advocate standard practice in collections management
 - Move forward in spite of problems
- Current suggestion
 - UUID as basis of identifier
 - URI with embedded UUID
 - urn:uuid:f47ac10b-58cc-4372-a567-0e02b2c3d47
 - ark:/87286/B2/f47ac10b-58cc-4372-a567-0e02b2c3d479

Conclusions

IF we want to provide, and use fit-for-research data, then

- Must have **identifiers** for objects
 - Especially occurrences
- Must have agreement on **properties (standard terms)**
- Must have strategy for **representing relationships**
 - In provider databases
 - In repositories
 - In transit



At iDigBio, we look forward to your continued input.
We need your voices.
Our planet needs the fit-for-use data.



Please Share Yours

DwC-A and the IPT

- DwC-A = Darwin Core Archive – contains 3 or more files
- Identifiers make this possible
- IPT = Integrated Publishing Toolkit creates the DwC-A
 - csv file – e.g. your specimen data records
 - meta.xml – a file that explains the contents of each column in the csv file
 - eml.xml – information about the data provider and what data is provided
- extensions – extending what the IPT can do.
 - image records for the specimens
- <http://tools.gbif.org/>
- <http://tools.gbif.org/dwca-assistant/>

From the researcher into a database (eventually)

- has standard metadata
 - in standard formats
 - standard packaging
 - storage
-
- Who bridges the transition from data collected in the field to transform it, standardize it for sharing, publication, storage, and insures it is discoverable for reuse?

Data use, data re-use

8.6. Lörttyneiden lep. tiedot, pöytä SU12

Benges	Laji	Ikä	Fukup.	Kyynärvarren Pituus, cm	Paino, g
BA00710	Myöskä	Ad.	N	39,0	8,5
BA00711	Myöskä	Ad.	N	39,0	9,0
BA00712	Myöskä	Ad.	N	38,0	9,5
BA00713	Myöskä	Ad.	N	39,0	9,0

9.6.

Eiisiltana odotimme kirkon nurkalla klo 22.50 aikaen, ja ensimmäinen lepakko lensi ulos klo 23.25. Ehdimme laskea 9 korvayökköä ennen kuin radiolähetin seurattavamme Ruut lähti klo 23.47. Ruut lensi taas suoraan puron ylä kohti Krejansbergetiä, ja me kiiruhdimme perään. Kirskoessani radiovastanottimen antennia läpi tiheästä kuusikosta järjän rinteen puolivälissä, totesin jälleen miten mahdotonta on pysytellä lentävän eläimen perässä jalkaisin, ja me kun vielä valitrimme korvayököt radiolähettimillä tutkittavaksi lajiksi osin siksi, että kirjallisuuden perusteella olemme niiden saalistavan pienellä alueella,

mieluiten puistomaisessa trossa, mikrei päiväpuloa toimivan kerkiäkaisen hautausmaa jaloine lehtipuineen sūs rütä?!

Nyppylän laelle päästyämme emme juuri saaneet hengähtää. Lähettimen signaali heikkeni, Ruut liikkui vauhdilla kohti pohjoista. Hetken luulimme jo kadottaneemme sen kokonaan. Tunnin kuluttua kuitenkin saavutimme sen parin kilometrin päästä, hakkuvauktion laidalta, jossa se pyöri saalirtaen



Benges	Laji	Ikä	Fukup.	Kyynärvarren Pituus, cm	Paino, g
BA00710	Myöskä	Ad.	N	39,0	8,5
BA00711	Myöskä	Ad.	N	39,0	9,0
BA00712	Myöskä	Ad.	N	38,0	9,5
BA00713	Myöskä	Ad.	N	39,0	9,0

Eiisiltana odotimme kirkon nurkalla klo 22.50 aikaen, ja ensimmäinen lepakko lensi ulos klo 23.25. Ehdimme laskea 9 korvayökköä ennen kuin radiolähetin seurattavamme Ruut lähti klo 23.47. Ruut lensi taas suoraan puron ylä kohti Krejansbergetiä, ja me kiiruhdimme perään. Kirskoessani radiovastanottimen antennia läpi tiheästä kuusikosta järjän rinteen puolivälissä, totesin jälleen miten mahdotonta on pysytellä lentävän eläimen perässä jalkaisin, ja me kun vielä valitrimme korvayököt radiolähettimillä tutkittavaksi lajiksi osin siksi, että kirjallisuuden perusteella olemme niiden saalistavan pienellä alueella,

mieluiten puistomaisessa ympäristössä, mikrei päiväpuloa toimivan kerkiäkaisen kivikirkon hautausmaa jaloine lehtipuineen sūs rütä?!

Nyppylän laelle päästyämme emme juuri saaneet hengähtää. Lähettimen signaali heikkeni, Ruut liikkui vauhdilla kohti pohjoista. Hetken luulimme jo kadottaneemme sen kokonaan. Tunnin kuluttua kuitenkin saavutimme sen parin kilometrin päästä, hakkuvauktion laidalta, jossa se pyöri saalirtaen



men

ibility

Relationships as annotations

- A relationship that needs its own properties
 - When, who, why, what evidence
- An annotation (*url1*) is an assertion of properties for objects
 - **On** 4 October 2013, Joe **claims** that specimen (*url2*) is of species *url3* **because** he disagrees with the determination on the label, and for **evidence**, he offers a set of image annotations (*url4*) showing morphological features that can be seen in the photograph (*url5*).
- Many relationships are expressed in the annotation