



#pacdigi

Sharing Data

Standards and Practice

Deborah Paul, @idbdeb, @iDigBio

Biological Digitization in the Pacific, March 23-27, 2014

East--West Center, Bishop Museum, University of Hawaii,

and the Pacific Science Association



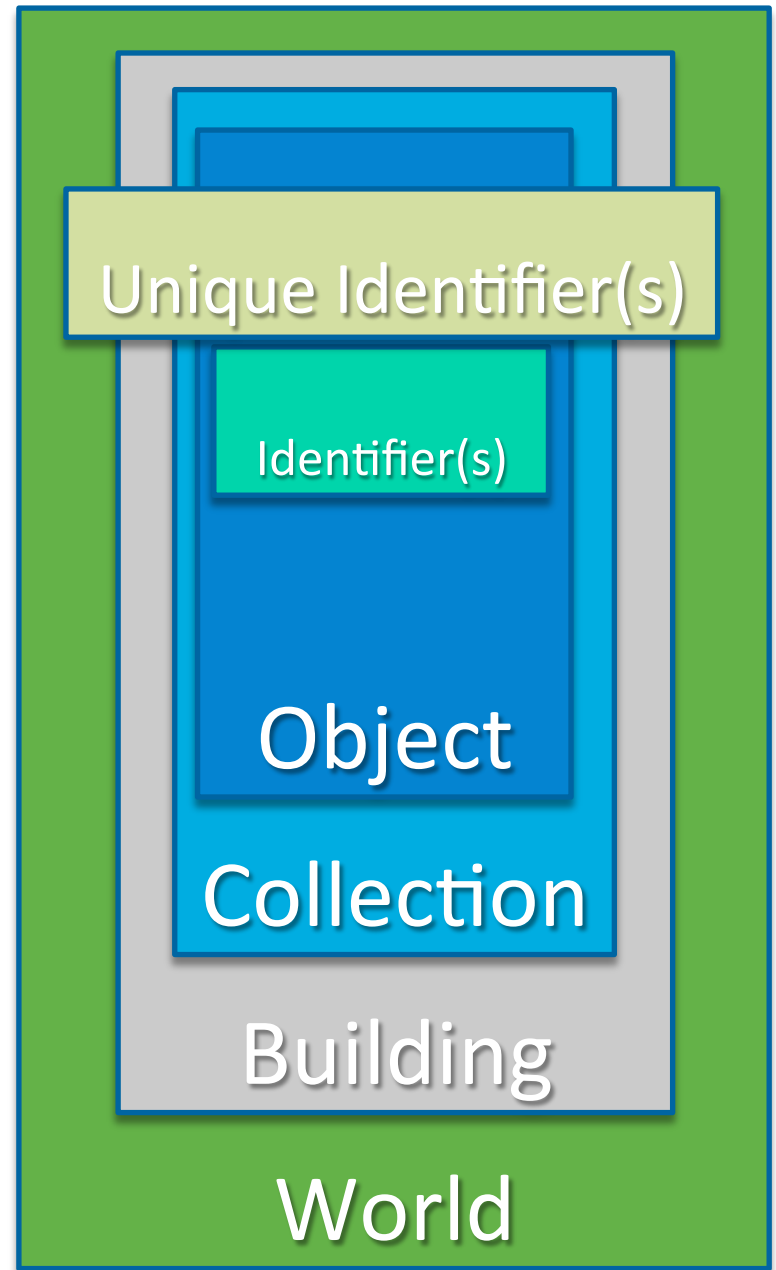
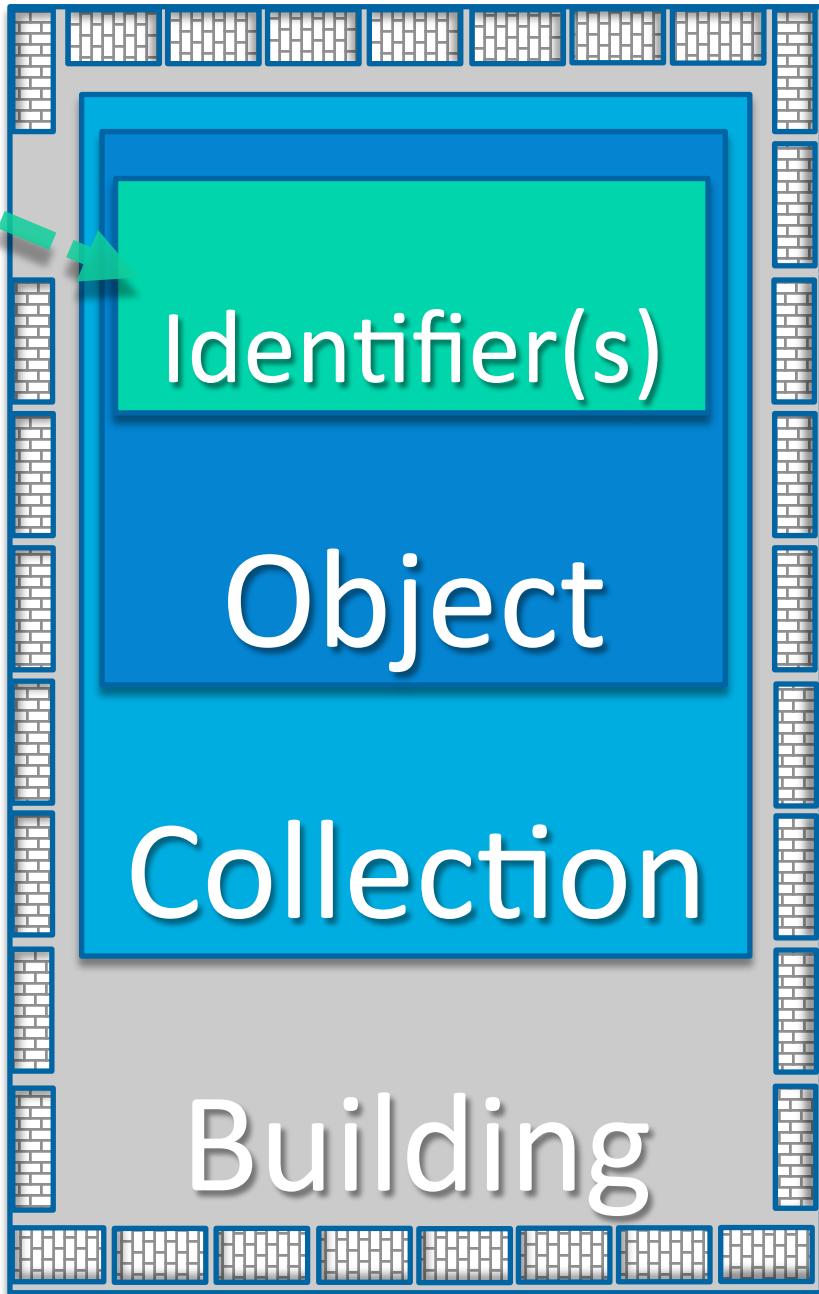
iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect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




R. K. GODFREY HERBARIUM
202973
FLORIDA STATE UNIVERSITY

Valdosta

Collecting Data





Making data and images of millions of biological specimens available on the web

12,999,602

Specimen Records

1,637,767

Media Records

200

Recordsets

Search the
Portal



Why digitization matters

More about what we do and why

- build an accessible aggregated, integrated, scalable, **vouchered-specimen database** (USA collections)
- facilitate and increase **participation in digitization**
- enable **researchers' access** to and use of **fit-for-use** data
- build **partnerships** to expand and enhance

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, ...

Why darwin core / georeferencing standards?



My field notes?

Your field notes?

map to a standard!

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.

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, ...

Note

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

New

Specify 6

FilteredPUSH



Encyclopedia of Life



GenBank



Identifiers

Keep

IFE

Symbiota

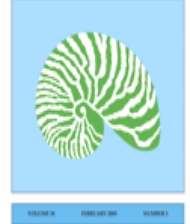
VeriNe

ZooKeys

Systematic Biology

Silver COLLECTION

GBIF



vizzuality

Cladistics®

BMC Bioinformatics

PhytoKeys



ZOONIVERSE REAL SCIENCE ONLINE



Identifiers are key

Maintaining and Sharing Identifiers

1 to many IDs known for a
given object

Store and share the ones you
know about

Specimen RecordID	19537
Specimen Previous Catalog Number	12345
Specimen Catalog Number / bar code	bbbrc000123
Darwin Core Triplet (DwC)	InstitutionCode:CollectionCode:bbbrc000123
DwC OccurrenceID	urn:catalog:institutionCode:collectionCode:bbbrc000123
Specimen GUID of type Isid	urn:lsid:biocol.org:bbbrc:bbbrc000123
Specimen Opaque Identifier(UUID)	424854d7-baec-42cf-a142-805b64117b9f
Specimen GUID of type URI	http://ids.flmnh.ufl.edu/herb/bbbrc000123

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

One Pathway to sharing?

- **Discoverability** for use / re-use / re-purpose / purpose discovery
 - Identifiers are key
 - Metadata is key

- Data in more than one place
 - + Aids discoverability
 - - Can be a issue to track
 - Identifiers help
- Dataset identifiers too

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

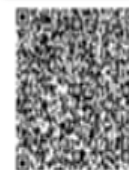
KUNHM-ENT

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

- 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>

Import & Export Clean Data

- Workbench-type strategies
- No matter the chosen database
 - clean the data first e.g.
 - Kepler Kurator
 - Google Refine
 - SQL, Reports,

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 and replaces an old dataset with a newer one.

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/>

Import to iDigBio

- iDigBio
 - CSV files, DwC-A file + extensions, and...
 - all data without limitations from a given standard
 - “if a field is valuable – it will someday be in a standard” (Schuh 2012)
- standards

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

- Visualization - 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

Feedback with Identifiers

- Annotations

- Target of annotation

Related Annotations					
Taxonomic Name	Taxon Author	Prefix	Suffix		
Opuntia humifusa	(Raf.) Raf.	none	none	1	0

- Repatriation

- Filtered PUSH
- Scatter Gather Reconcile (SGR)

- Linked Data, aka the Semantic Web

- BiSciCol

- updating the database

- be(a)ware
- store and share other IDs

Biodiversity Information Standards

Biodiversity
Information
Standards
T D W G

- formerly (still) 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
- Your **input** and **participation** make standards **robust and useful**

The data is born (digital)?

- researcher collects data
- organizes it for their purpose (publishing)?
 - or not
- non-standard metadata
- non-standard file formats, file-naming, packaging
- user file system
 - unique
 - sometimes enigmatic?

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. lörtyneiden lep. tiedot, pöytä SU12

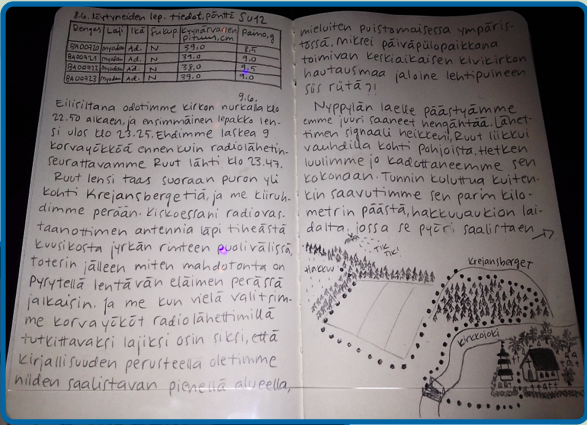
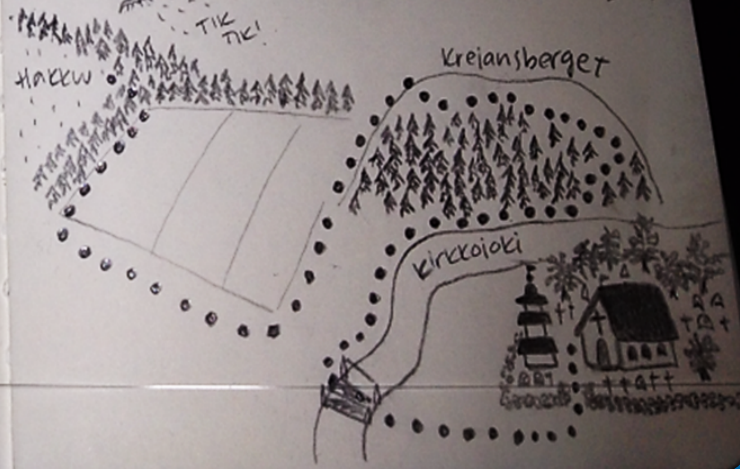
Benges	Laji	Ikä	Sukupu.	Kyynärvarren Pituus, cm	Paino, g
BA00710	Myöskä	Ad.	N	39,0	8,5
BA00711	Myöskä	Ad.	N	39,0	9,0
BA00712	Myöskä	Ad.	N	38,0	9,5
BA00713	Myöskä	Ad.	N	39,0	9,0

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 yli 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 trossä, mikrei päiväpuloa toimivan kerkiaikaisen 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

Mahalo nui loa to you from



East--West Center, Bishop Museum,
University of Hawaii, and the Pacific Science Association

We look forward to your continued input at iDigBio.
We need your voices, your ideas, your participation.

Here's to museum specimen and related data online accessible to all.

