



It's About Data

iDigBio Wet Collections Digitization Workshop

March 4 – 6, 2013

KU Biodiversity Institute, University of Kansas – Lawrence

Deborah Paul, Greg Riccardi, Joanna McCaffrey



Overview

- **Identifiers**
- Darwin Core
- Collaborating with TCNs and PENs
- Contributing to iDigBio
- Data Aggregators
- Guidelines for selecting a database
- Digitization Maturity



Identifiers

- What good is identification?
- How are identifiers used by consumers
- Providing IDs
- Annotations and Feedback

What good is identification?

- **Aggregation**
 - If you get info from 2 sources that are about the same object, you can combine the info
- **Resolution** (finding information about object)
 - Types of resolution
 - Determine where to get information
 - Determine how to get information

DOI example

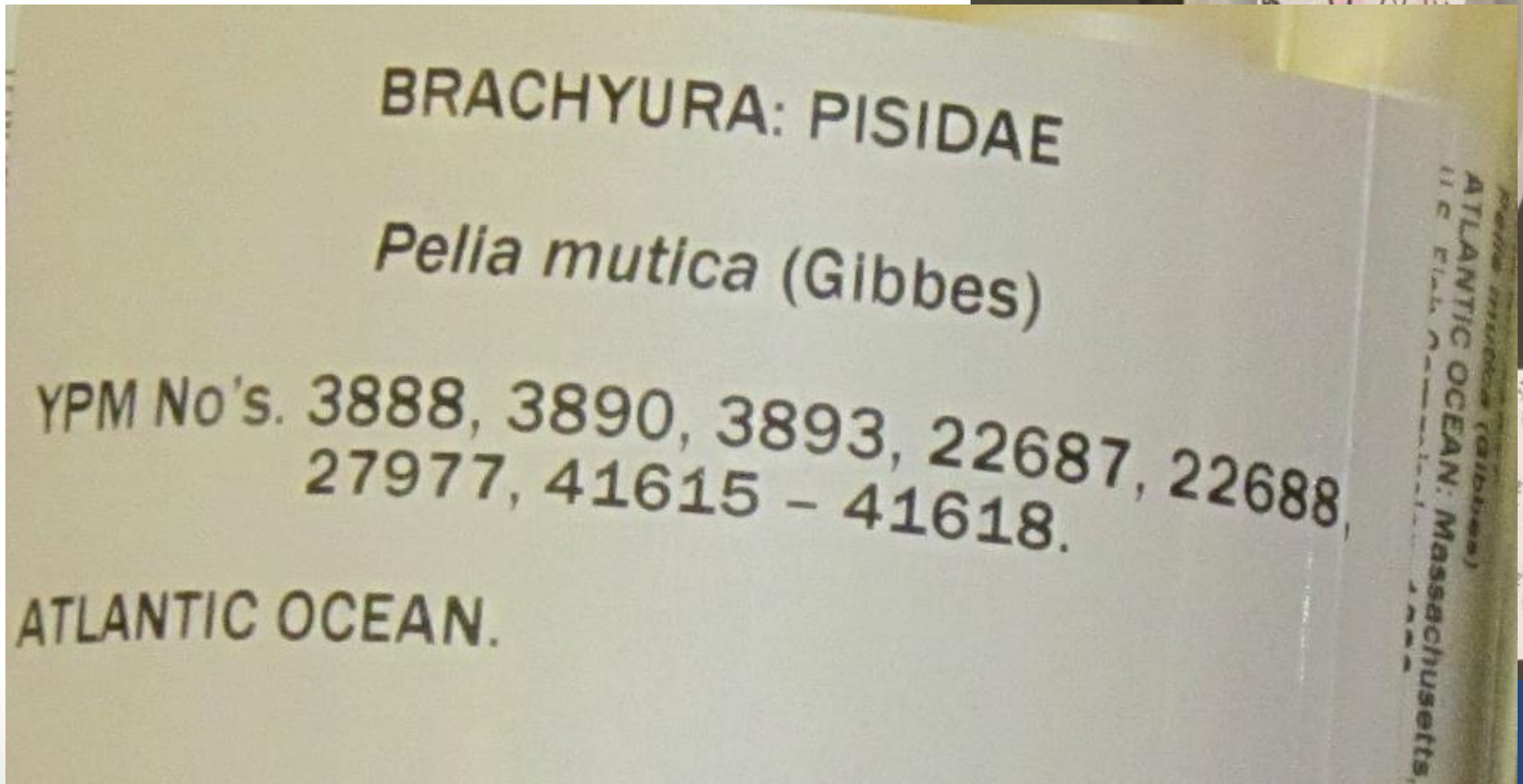
- The DOI is
 - [10.3897/zookeys.209.3135](https://doi.org/10.3897/zookeys.209.3135)
- URI (for aggregating & display) is
 - [doi:10.3897/zookeys.209.3135](https://doi.org/doi:10.3897/zookeys.209.3135)
- A URL for information retrieval (proxy resolution) is
 - <http://dx.doi.org/10.3897/zookeys.209.3135>
- Information fetched from
 - HTML:
 - <http://www.pensoft.net/journals/zookeys/article/3135/abstract/five-task-clusters-that-enable-efficient-and-effective-digitization-of-biological-collections>

Zoobank Uses Identifiers

- Consider web page
 - <http://zoobank.org/NomenclaturalActs/4DFD6D95-C287-4AFF-8473-E073D8960EC6>
- Look for identifiers
 - urn:lsid:zoobank.org:act:4DFD6D95-C287-4AFF-8473-E073D8960EC6
 - HOLOTYPE: Deposited as No. 7113, Hancock Parasitology Collection, University of Southern California
- Search in google for these

What about Specimen identifiers?



- Identifier on the specimen?
 - readable text



Feedback with IDs

- Annotations
 - Target of annotation
 - <http://www.morphbank.net/818505>

Related Annotations

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

- filtered PUSH
- SGR
- BiSciCol
- **linked data**, aka **the semantic web**
- updating the database
 - be(a)ware
 - store and share other IDs

Specify 6



EMu Museum Management System

FilteredPUSH



Identifiers

GenBank



Kepler Kura

DISCOVER LIFE

Symbiota



VerNe



ZooKeys

Cladistics

Silver COLLECTION



PhytoKeys



vizzuality

ZOONIVERSE REAL SCIENCE ONLINE



Overview

- Identifiers
- **Darwin Core**
- Collaborating with TCNs and PENs
- Contributing to iDigBio
- Data Aggregators
- Guidelines for selecting a database
- Digitization Maturity

Darwin Core Standard

<http://rs.tdwg.org/dwc/terms/>

- 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] .

Darwin Core Standard

<http://rs.tdwg.org/dwc/terms/>

- Darwin Core (often abbreviated to DwC) is

a vocabulary of terms to facilitate the discovery, retrieval, and integration of information about organisms,

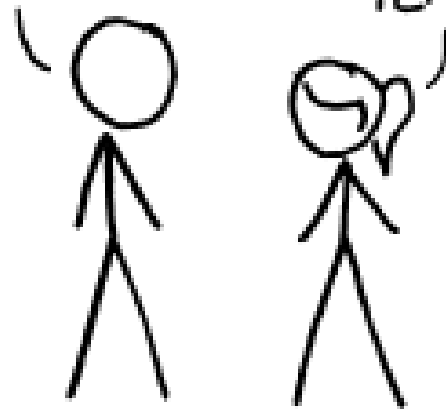
- Does Darwin Core cover every field possible? – No
- Don't panic! There are extensions and other options.

HOW STANDARDS PROLIFERATE:

(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)

SITUATION:
THERE ARE
14 COMPETING
STANDARDS.

14?! RIDICULOUS!
WE NEED TO DEVELOP
ONE UNIVERSAL STANDARD
THAT COVERS EVERYONE'S
USE CASES.



SOON:

SITUATION:
THERE ARE
15 COMPETING
STANDARDS.

Overview

- Identifiers
- Darwin Core
- **Collaborating with TCNs and PENs**
- **Contributing to iDigBio**
- **Data Aggregators**
- Guidelines for selecting a database
- Digitization Maturity

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
 - VertNet
 - Morphbank
 - iDigBio
 - GBIF
- Help is everywhere!

Sharing data with iDigBio

- Custom export
- CSV files
- DwC-A files
 - DwC-Extensions
 - MeasurementOrFact
 - ResourceRelationship
 - AudubonCore
- Specimen Identifiers
- Record Identifiers

Data in iDigBio

- goal to allow all possible data w/o limitations from a given standard
- “if a field is valuable – it will someday be in a standard” (Schuh 2012)

- standards

Overview

- Identifiers
- Darwin Core
- Collaborating with TCNs and PENs
- Contributing to iDigBio
- Data Aggregators
- **Guidelines for selecting a database**
- Digitization Maturity

Choosing a database / collections management system

- Establish *institutional motivation* to digitize specimens
 - earnest desire could be generated by a focus group with institutional stakeholders, funding generators, users of data, people who input data, data system supporters (curators), IT,

Considerations for selecting a collections management system

- Document and agree on a **priority feature set** that is **necessary** versus **desired**:
 - system is extensible, customizable,
 - responsive vendor,
 - supports reports, auditing,
 - generates labels,
 - supports loans (partial returns, cataloged and uncataloged specimens),
 - supports pest management,
 - supports multimedia attachments (PDF loan forms, image, sound files, etc.),
 - supports web access and privacy,

Considerations for selecting a collections management system

- all the **input/output scenarios** you might envision:
 - Import/export abilities, can it support DarwinCore field mappings
 - **plan B scenario** if software or the internal project becomes unfunded,
- affordable user **license costs**: per seat, pool,
- has basic, and easily **customizable help**,
- **Mac** versus **PC**, perhaps an issue in your user population,
- has a robust **security model** (passwords, users, groups, permissions, input and query defaults, controlled vocabularies),
- supports accessibility, different **character sets**,

Considerations for selecting a collections management system

- Proprietary, open source, hybrid, cloud-based
 - Who decides what features to develop?
 - Who does maintenance?

Considerations for selecting a collections management system

- Interest in having what your **peers** have: economies of training, user community,
- Beware of **demo-ware**,

Considerations for selecting a collections management system

- **Shop** vendors and score them on their ability to meet necessary features above, with extra points for desired ones,

Considerations for selecting a collections management system

- Get a full demo copy and enter data with a **realistic test case dataset**, score on ease of learning the system,
 - novice and expert user

Considerations for selecting a collections management system

- When choosing preferred system, **consider costs** derived from these sources:
 - upfront software costs,
 - software maintenance,
 - long term costs (server space, server replacement, backup),
 - where it is hosted,
 - IT support of system without being the bottleneck,
 - hidden costs of conversion, cleansing, improvements,
 - institutional ***biodiversity informatics staff*** support to continue development of data, ('data curator').
- <https://www.idigbio.org/content/biological-collections-databases>

Overview

- Identifiers
- Darwin Core
- Collaborating with TCNs and PENs
- Contributing to iDigBio
- Data Aggregators
- Guidelines for selecting a database
- **Digitization Maturity**

Digitisation Maturity?

<http://tinyurl.com/digitizationala>



Figure 4 Digitisation maturity model

Choose your level

<http://tinyurl.com/digitizationala>

- **0 No idea**
 - No-one has any idea what's happening across the organisation
 - Digitisation left to individuals
- **1 Making do**
 - Some of us are doing a good job
 - Digitisation left to individuals who have their standard processes
- **2 Coming along nicely**
 - We are getting our act together and starting to share the same idea
 - Digitisation left to each part of the organisation
- **3 Organized**
 - We all know what to do, how it all fits together, and share same idea
 - Management takes responsibility for digitisation
- **4 Under control**
 - We all know how well we are doing
 - Managers take responsibility for improving digitisation
- **5 On the lookout**
 - We are building the (next) idea together
 - Continuous innovation by everyone



Thank you...

