Aggregation &
The Mechanics Of Data Sharing

Patricia Holroyd
Museum of Paleontology
University of California

UCMP

Pop Quiz!

- **Downloaded online digital data for research?**
- **Shared research data online?**
- **Share your collections online? On a site not your own?**
- **Updated data sets online?**

Morphobank media number
M332999

Taxonomic name
†*Aceroryctes dulcis*

Specimen
†*Aceroryctes dulcis* (UCMP/UCMP:131849)

Specimen notes
holotype, from locality UCMP V71237

Media loaded by
Patricia Holroyd

Copyright holder
Patricia Holroyd

» Zoom » Download

Data aggregation

compiling of information from multiple sources with intent to prepare combined datasets for data processing or analysis

Can be “automated” or “manual”

Goals

Standardize data

Facilitate comparisons

Expose resources

Improve data through feedback

Document research

Make studies repeatable

Find new patterns and relationships

“Automated” vs. “manual”

- Shared via automated to semi-automated processes
- Primarily specimen occurrence data & images
- Collections data ≠ research data
- Use international metadata standards
- Examples: iDigBio, GBIF

- Actively uploaded and shared
- Primarily research driven
- Often static archives associated with publications
- Many kinds of data and standards
- Examples: Dryad, Morphobank, Morphosource

“Manual” aggregations of fossil data

The Paleobiology Database
revealing the history of life

NOAA
National Oceanic and Atmospheric Administration
U.S. Department of Commerce

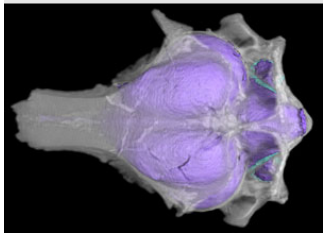
figshare

MORPHOBANK
HOMOLOGY OF PHENOTYPES OVER THE WEB

NEOTOMA PALEOECOLOGY DATABASE

MORPHO SOURCE

DIGITAL MORPHOLOGY
A National Science Foundation Digital Library at The University of Texas at Austin



DRYAD

morphbank
http://morphbank.net

Combination of static archives
& research databases updated with expert input

“Automated” data aggregators

9,658,971
fossil specimens
& occurrences

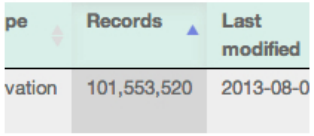
4,075,098
fossil specimens

Largest online data set… but are only a small number of collections

Automated aggregation: IPT

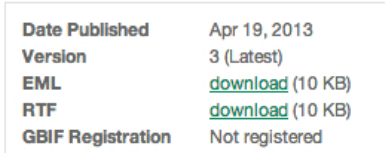
Integration with GBIF, DataCite, EZID

Can automatically register datasets with GBIF making them globally discoverable through the GBIF website. Can also automatically connect with either DataCite or EZID to assign DOIs to datasets.



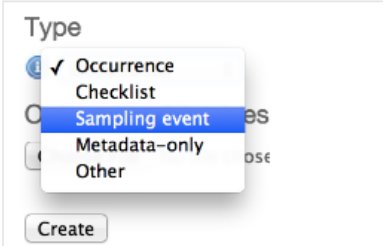
Support for large datasets

Can process ~500,000 records/minute during publication. Disk space the only limiting factor. For example, a published dataset with 50 million records in DwC-A format is 3.6 GB.



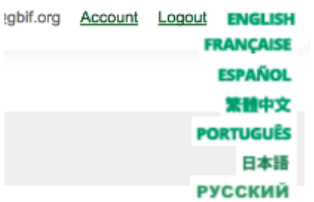
Standards-compliant publishing

Publishes a dataset in Darwin Core Archive (DwC-A) format, a compressed set of files based on the Darwin Core terms, and the GBIF metadata profile built using the Ecological Metadata Language.



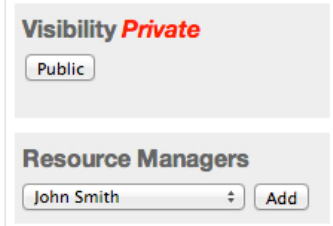
Publication of four types of biodiversity data:

I. Primary occurrence data (specimens, observations)
II. Species checklists and taxonomies
III. Sample-based data (data about sampling events)
IV. General metadata about data sources



Internationalization

User interface available in seven different languages: English, French, Spanish, Traditional Chinese, Brazilian Portuguese, Japanese and Russian.



Data Security

Controls access to datasets using three levels of dataset visibility: private, public and registered. Controls which users can modify datasets, with four types of user roles.

Java Script Software that provides translation of your data to international standards

- http://www.gbif.org/ipt

Sharing collections data via IPT

Biodiversity Information Standards

TDWG

You don't change your database, only map data to international standards

Dublin Core – any data

Examples:

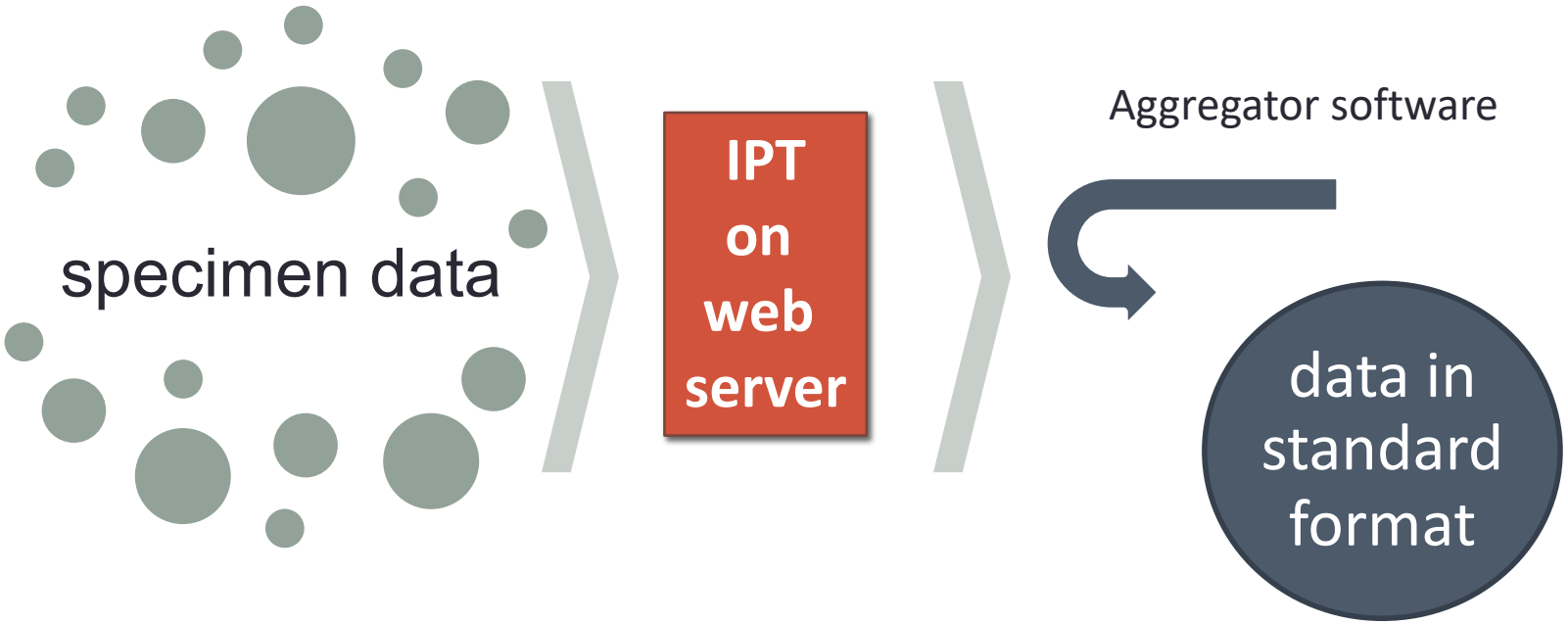
UCMP SpecNo = dwc:catalogNumber

UCMP element = dwc:Preparations

Darwin Core – biodiversity data (including paleo)

Audubon Core – multimedia (images, video, audio, other data types)

Sharing collections data via IPT



may be in various formats

(.csv, Access, Specify, KeEmu, etc.)

iDigBio

GBIF

Sharing collections data via IPT

- Set up your own IPT
  - Gives you greatest control over what data are served
  - Ask your system and database administrators if this is viable
- Use a shared IPT (e.g., Vertnet)
  - Get expert assistance in translating your data to standards
  - Currently serving more than 20 million biodiversity records
- http://vertnet.org/

What data could and should we be better sharing and aggregating?

- Trait data – measurements,
- Isotopic and geochemical measurements
- Photos for morphometrics and image stacking
- CT and other types of derivative imaging
- Observations – geologic and field censuses
- Detailed Geologic context

How do we get more and better data online?