Data Aggregators:

A data provider's perspective

Talia Karim

University of Colorado Museum of Natural History

University of Colorado **Boulder**

Data Aggregator?

Data Aggregator compiles information, from databases or datasets, with the intent to prepare combined datasets for data processing (adapted from Wikipedia)

Cleaning algorithms

Data Normalization

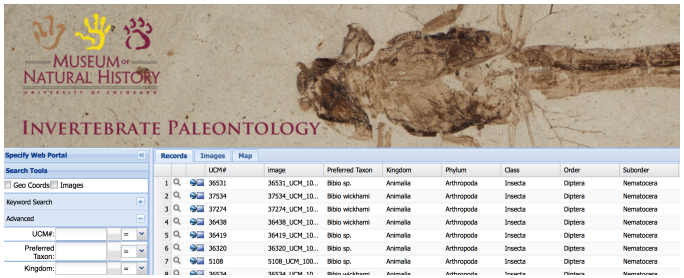
Integration of Different Data Types

Aggregators also make the compiled data available online for research and other uses.

Permissions or Restrictions?

Who Are The Data Aggregators?

- Institution/Collections Based
- Project Based Portals
  - SCAN
- National/Regional Portals
  - iDigBio, ALA, CONABIO, etc.
- Global Portal
  - GBIF



SCAN
Symbiota Collections of Arthropods Network

iDigBio
Integrated Digitized Biocollections

Atlas of Living Australia
ala.org.au

GBIF
Global Biodiversity Information Facility

CONABIO
Comisión Nacional para el Conocimiento y Uso de la Biodiversidad

Types of Datasets

- Resource Metadata
- Checklist Datasets
- Occurrence Datasets
  - Specimen or observation based
- Sampling-Event Datasets

More Information: https://www.gbif.org/dataset-classes

Format Your Data

- Darwin Core Archive (DwC-A)

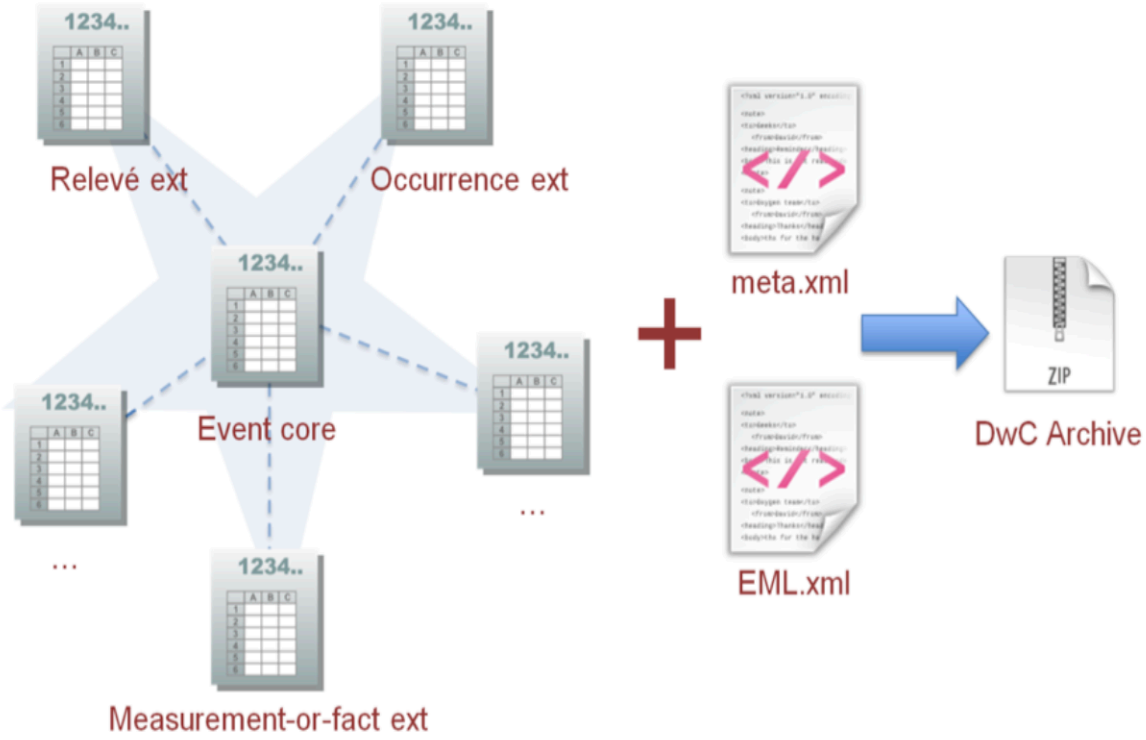
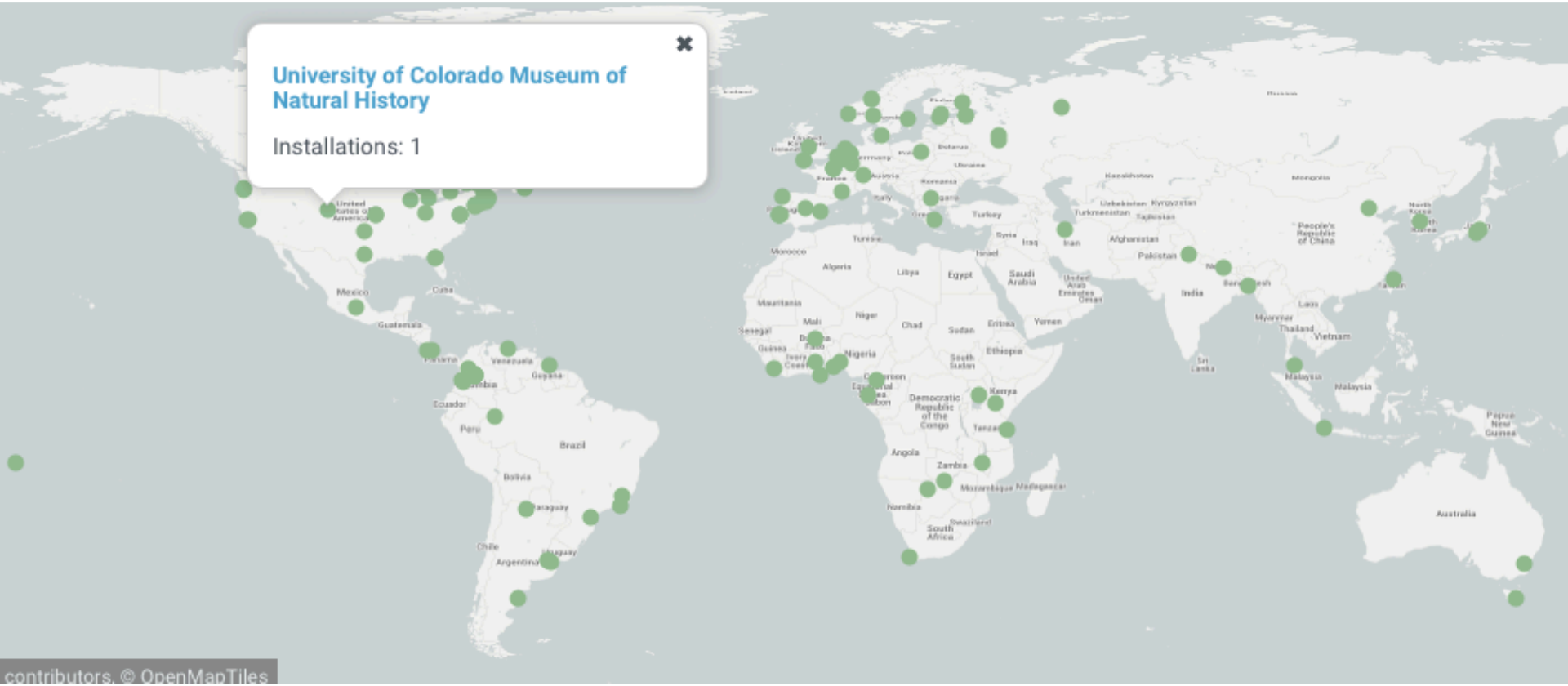


Image and More Information: https://github.com/gbif/ipt/wiki/DwCAHowToGuide

IPT: The Integrated Publishing Toolkit

A free open source software tool used to publish and share biodiversity datasets through the GBIF network.



https://www.gbif.org/ipt

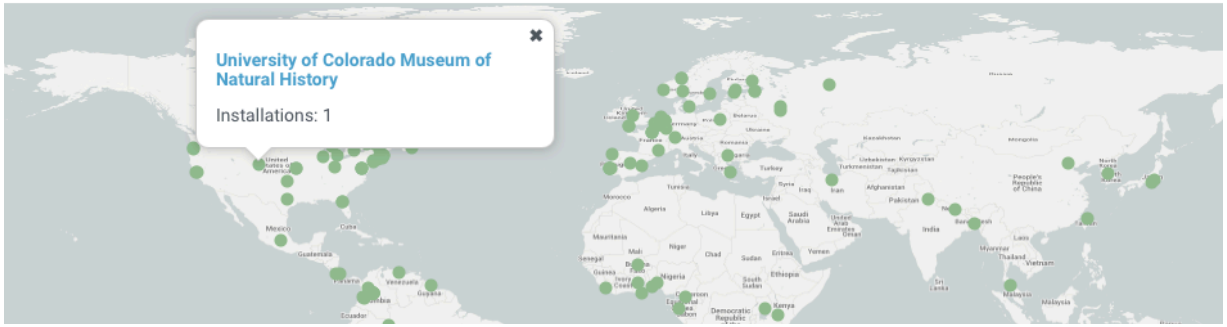
How Do Aggregators Find Your Data?

- Integrated Publishing Toolkit (IPT)
- RSS Feed
- Email/Share a spreadsheet

IPT: The Integrated Publishing Toolkit

A free open source software tool used to publish and share biodiversity datasets through the GBIF network.

https://www.gbif.org/ipt



@Translators

The IPT user interface and wiki both need internationalisation, but it's a community effort and everyone is welcome to join. Full instructions aimed at translators can be found here.

Thanks to an enormous community effort, and by leveraging the power of the Crowdin localisation tool, the user interface has already been translated into seven different languages: English, French, Spanish, Traditional Chinese, Brazilian Portuguese, Japanese, and Russian.

https://www.gbif.org/translators

Post-Data Publishing Lifecycle

- Data Ingestion and Integration by Aggregator:
  - Cleaning algorithms
    - Not all the same
  - Differences in fields reported
  - Indexing of data
  - Non-Standard vocabulary or formats (dates)
    - Makes your data uninterpretable by aggregator and downstream users

- Data Attribution
  - How do people cite datasets in their research?
    - Assign DOI to dataset
    - Include “datasets used” in Data Citation/Reference section of manuscript
  - How do we track data usage outside of traditional publications?

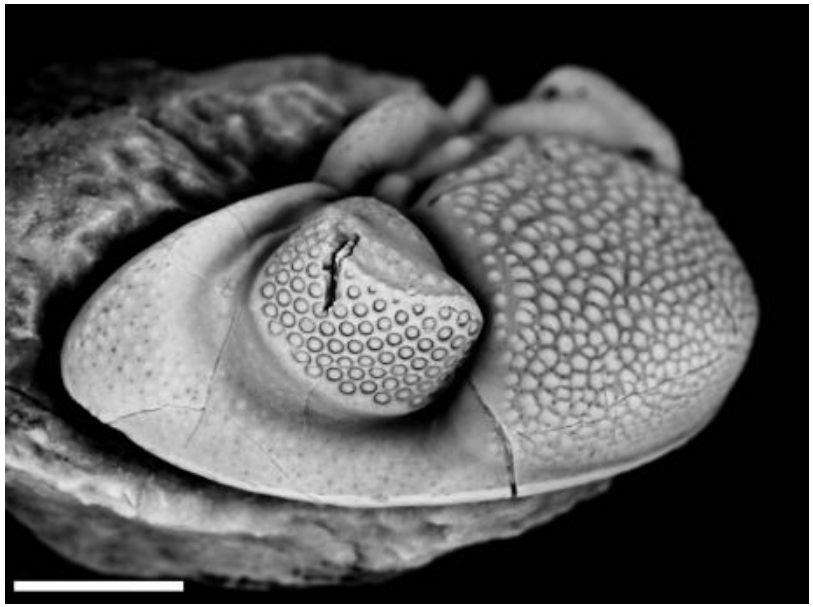
- Annotations
  - How do these get routed back to the data provider to improve data quality?

Future- Seamless Global Data Lifecycle

- What does this look like?
- Data publishing streamlined
- Persistent and resolvable identifiers
- Easy integration of other data types
  - Climate, genomic, trait data, etc.
- Tool development for data use downstream
- Better ways to track data use
- Broadening use of data
  - Food security, ecosystem services, government planning, pubic health

Image: https://whatsthebigdata.files.wordpress.com/2015/10/datascience_unicorn.png

You Might Also Like:



- WT44 DWG Paleo IG Workshop: Coordinating best practices for fossil specimen data mobilization
  - Room: Jan Willem Schaapfoyer
  - Time: Wednesday (23rd) 13:30-15:00

- Reducing Dependence on Digital Biodiversity Data Silos Through Global Alignment and Collaboration
  - Room: Jan Willem Schaapfoyer
  - Time: Thursday (24th) 11:24