iDigBio Nuts and Bolts

Data Integration and Ingestion

Data Integration

“Data integration involves combining data residing in different sources and providing users with a unified view of these data.” - Wikipedia, Data Integration

What we’re trying to do: Mobilize as much of the data from museum collections as is possible, in a way that most consumers of this information can understand it, and without losing linkages back to the information providers so that we can track curatorial control of the information.

Core Standards

- Specimen Data - Darwin Core
    - Currently expecting a flat darwin core representation.
    - Recommended Minimum: Record ID, Scientific Name, Occurence ID, Event Date, Collector Name, Locality Data
- Media Data - Audubon Core
    - Currently expecting flat audubon core with some provision for providing multiple Access URIs.
    - Recommend Minimum: Access URI, rights, provider, scientific name, title, description, tags
- Guidance on other terms to use might be available in the MISC Working Group documents

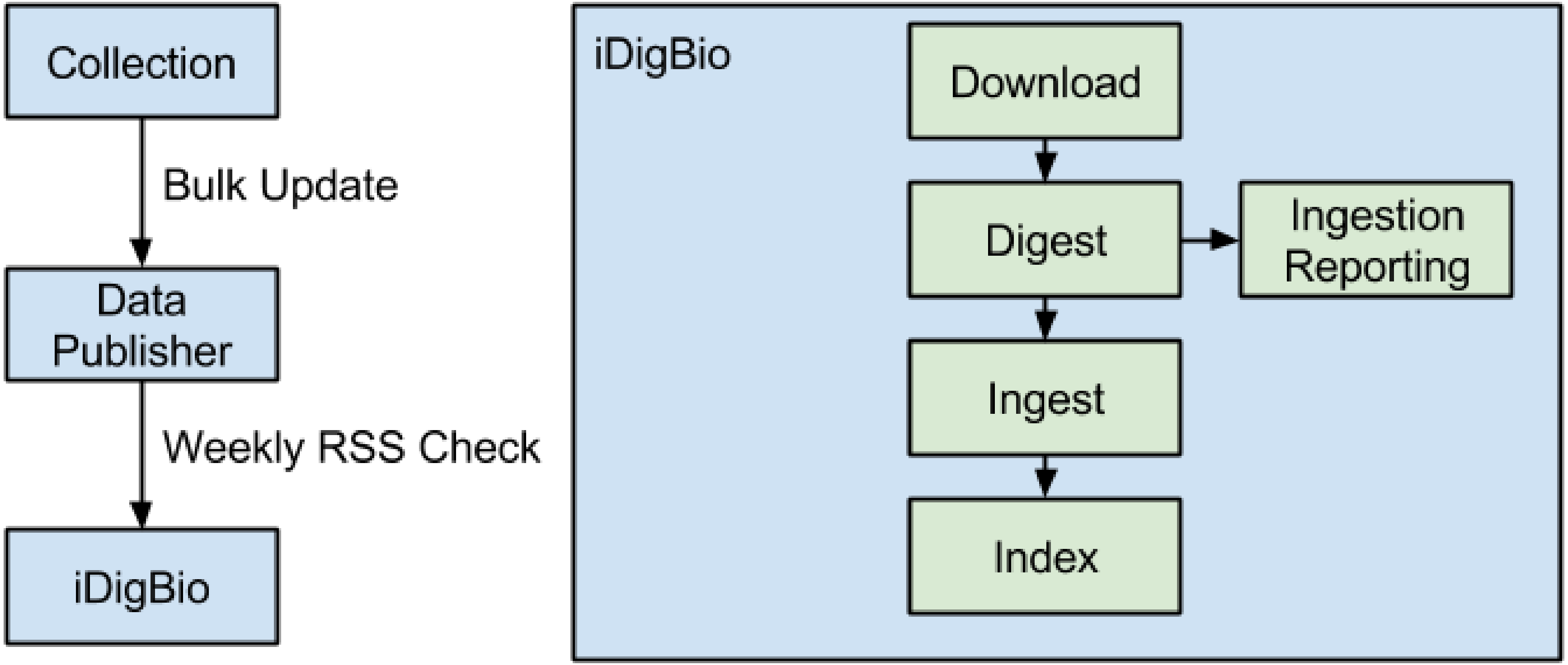
Things the standards don’t cover.

- Other people have covered many of these topics very well: See Apple Core for herbarium targeted advice that is fairly generally applicable.
- Data Formats
  - ISO 8601 Dates
  - WGS84 Decimal Lat/Long
- Controlled Vocabularies
  - ISO Country Codes and Subdivision Codes
- Identifier Formats
  - URN, ARK, DOI, URI, URL, LSID

Data Ingestion Formats

- Darwin Core Archive
  - Delimited files + Metadata in a Zip archive
- CSV
  - With darwin core/audubon core based field-names
- RSS
  - For publishing multiple recordsets
- HTTP
  - For single files

Data Ingestion Process

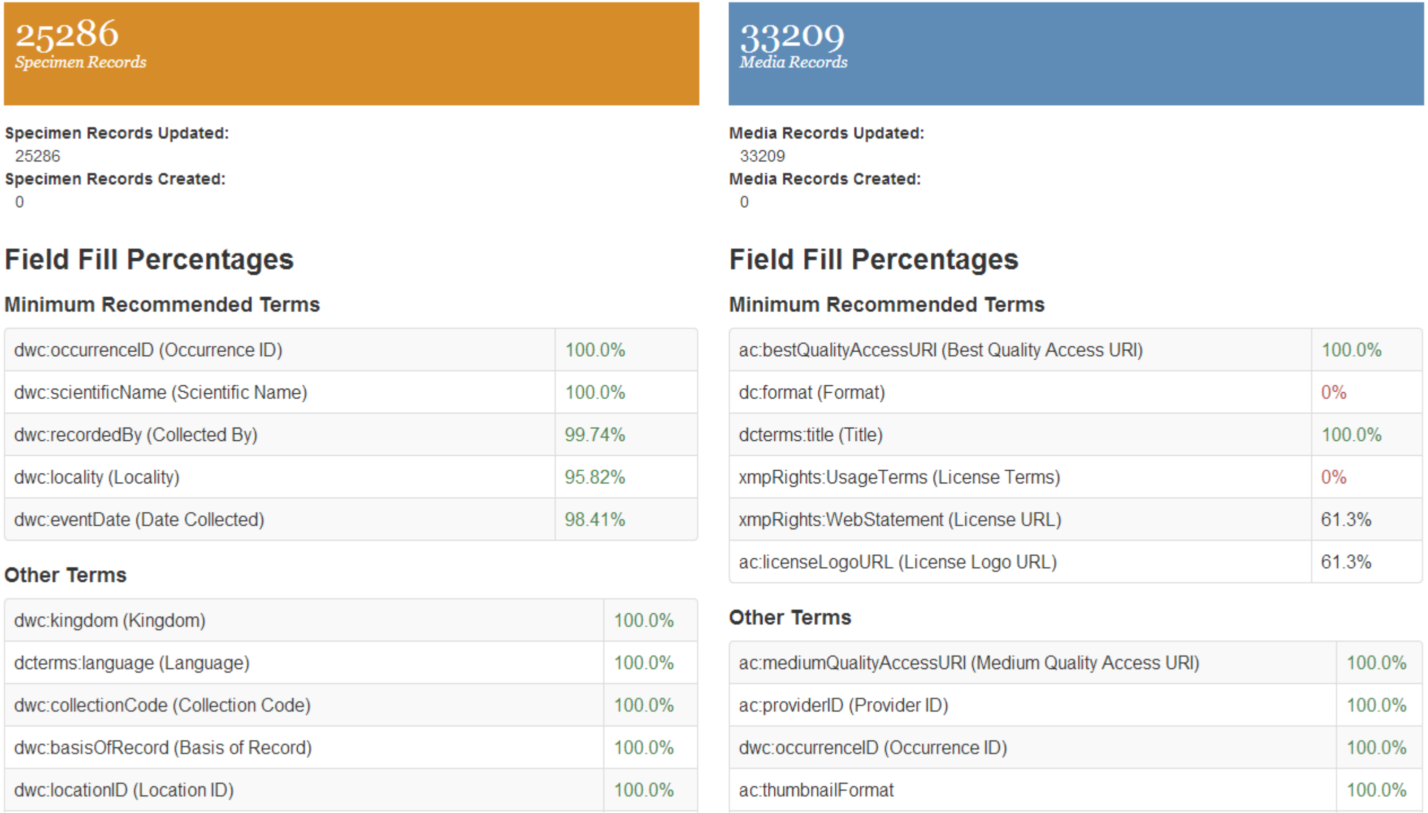


Other Options for Data Publishing

iDigBio will act as a data publisher, accepting files from collections that can’t or don’t wish to work with another publisher or run one themselves.

We’ll be publishing to ourselves with a very simple PHP script that should be able to run on nearly any web server. If you’d like a copy to run your own publisher that way, contact me

Ingestion Reporting



25286
Specimen Records

Specimen Records Updated:
25286
Specimen Records Created:
0

Field Fill Percentages

Minimum Recommended Terms

dwc:occurrenceID (Occurrence ID)	100.0%
dwc:scientificName (Scientific Name)	100.0%
dwc:recordedBy (Collected By)	99.74%
dwc:locality (Locality)	95.82%
dwc:eventDate (Date Collected)	98.41%

Other Terms

dwc:kingdom (Kingdom)	100.0%
dcterms:language (Language)	100.0%
dwc:collectionCode (Collection Code)	100.0%
dwc:basisOfRecord (Basis of Record)	100.0%
dwc:locationID (Location ID)	100.0%

33209
Media Records

Media Records Updated:
33209
Media Records Created:
0

Field Fill Percentages

Minimum Recommended Terms

ac:bestQualityAccessURI (Best Quality Access URI)	100.0%
dc:format (Format)	0%
dcterms:title (Title)	100.0%
xmpRights:UsageTerms (License Terms)	0%
xmpRights:WebStatement (License URL)	61.3%
ac:licenseLogoURL (License Logo URL)	61.3%

Other Terms

ac:mediumQualityAccessURI (Medium Quality Access URI)	100.0%
ac:providerID (Provider ID)	100.0%
dwc:occurrenceID (Occurrence ID)	100.0%
ac:thumbnailFormat	100.0%

Future Work

- Publishing to GBIF
- Automated Data Citation file in downloads
- Rights file in downloads
- Usage Reporting
    - Will initially be focused on easy wins
    - Usage counts per institution, dataset, and publisher
        - Including Portal & API Views, as well as downloads