Sample Data Set for Workshop: Data sharing, data standards, and demystifying the IPT.

This document describes the known (intentional) issues in the dataset. Participants don’t see this until after or at the end of Day 1. It is intended for the instructors and assistants. Errors are representative of the majority of the error types seen. This error-filled dataset is in the drop box where the presenter’s powerpoints are located.

**occurrence.txt issues**

1. Headers need standardization (note not in camelCase, spelling errors)
   1. Collecter, Occurence
   2. Country (header is missing, so data would be off by one column in mapping in the IPT?)
   3. see table below
2. Data must have globally unique identifiers, for example UUIDs
   1. make sure to store them in your database if you are creating them on export.
   2. make sure there are no duplicates (there is a dup in the data set).
      1. how to fix?
      2. check: are the records truly duplicates OR is the identifier in error
3. Use standard “controlled” vocabularies where possible / suggested / required
4. Some fields can / will need to be added either
   1. by a literal in an SQL query
   2. provided in the csv or
   3. provided in the IPT
      1. **collectionCode** (=Herbaria, =Fish, =UniqueIdentifierForCollection)
      2. **type** (=PhysicalObject
      3. **basisOfRecord** (=PreservedSpecimen, or appropriate value from controlled vocabulary). \*Note this term is being deprecated.

|  |  |  |
| --- | --- | --- |
| SampleData.zip Headers | map terms on left to dwc terms | fields in order from IPT? |
|  |  |  |
| id | occurrenceID(=**uuid**) |  |
| Kingdom | kingdom |  |
| Phylum | phylum |  |
| Class | class |  |
| Order | order |  |
| Family | family |  |
| Genus | genus |  |
| Common Name | vernacularName |  |
| Specific Epithet | specificEpithet |  |
| Infraspecific Epithet | infraspecificEpithet |  |
| Scientific Name | scientificName |  |
| Rank | taxonRank |  |
| Type of Preservation | basisOfRecord(=PreservedSpecimen) |  |
| Coden | institutionCode |  |
| Occurence Remarks | occurrenceRemarks |  |
| Catalog | catalogNumber |  |
| Determiner | identifiedBy |  |
| Record Number | recordNumber |  |
| Collection Date | eventDate |  |
| Collection Date2 | verbatimEventDate |  |
| Collecter | recordedBy |  |
| Life Stage | lifeStage |  |
| Nomenclatural Notes | taxonomicStatus |  |
| Preparation | preparations |  |
| Country | country |  |
| Country Code | countryCode |  |
| State | stateProvince |  |
| County | county |  |
| Locality | locality |  |
| Habitat Description | habitat |  |
| Longitude | decimalLongitude |  |
| Latitude | decimalLatitude |  |
| Georeference Remarks | georeferenceRemarks |  |
| Datum | datum |  |
| Group | group |  |
| Formation | formation |  |
| Member | member |  |
| Bed | bed |  |
| Dynamic Property | dynamicProperty |  |
| Morphbank URL | associatedMedia |  |
| License | accessRights | Information about who can access the resource or an indication of its security status. Access Rights may include information regarding access or restrictions based on privacy, security, or other policies. |
| Rights | rights | <https://creativecommons.org/publicdomain/zero/1.0/>  or is this the plain text of  dc:rights (instead of dcterms:rights which is expected to be a URI) |
|  | type(=**PhysicalObject)** |  |
|  |  |  |
|  |  |  |
|  |  |  |

**NOTE: SQL can alleviate a lot of these mapping issues, and make things much easier like creating JSON output for the dynamicProperties field.**

**Laundry List of Issues added to the occurrence dataset. (too many to fix in our sessions).**

1. 1 duplicate identifier in **id** and **occurrenceID** field
   1. GR (may need to suggest no prefix urn:uuid)
2. **eventDate**, needs to be in ISO std format
   * 1. 10-May-10 = 2010-05-10 or is it 1910-05-10 or 1810-05-10 …
     2. 12-Jun-00 = 2000-06-12 or 1900-06-12 or 1800-06-12 …
     3. 10/13/2000 = 2000-10-13
     4. January 12 - 15th, 2015 = 2015-01-13/15
     5. Jun 1906 = 1906-06
3. no **collectionCode** (**can add as literal, or “on export”**)
   1. prefer at something unique within your collection (herps, fishes, seeds, herbarium,…)
   2. we (iDigBio) prefer that you populate this field
      1. supports those who still use / want to use triplets
      2. and provides easier way for source to recognize where records come from (Katja’s TCN data issues as example)
      3. example Harvard University dataset could not be distinguished w/o collectionCode
4. non-standard **stateProvince**
   1. Miss = Mississippi
   2. AL = Alabama
   3. alabama = Alabama
5. non-standard **country** (USA, U.S.A. etc).
   1. USA, US, united states, UNITED STATES OF AMERICA, U. S. A.
6. **basisOfRecord**=Specimen
   1. change to value=**PreservedSpecimen**
   2. Specimen (in 7 records)
7. In participant data sets latitudes and longitudes may need converting to degrees decimal, and / or populating verbatimCoordinates field and converting…
8. latitude
   1. put 0 – replace with “empty string”
9. longitude
   1. put 0 – replace with “empty string”
   2. NULL – replace with “empty string”
   3. redacted – replace with “empty string”
      1. note: this redacted information belongs in dwc:informationWithheld
10. License
11. Rights
12. Life Stage
13. Determiner

**Fields Added**

1. add field **Country Code** (ISO standard needed)
   1. need to populate
2. occurrenceRemarks, georeferencingRemarks
3. geodeticDatum
4. dynamicProperties (must be in JSON format)
   1. heightInMeters=1.5, natureOfID=expert identification
   2. {"heightInMeters":1.5, "natureOfID":"expert identification"},
5. Paleo fields added: group, formation, member, bed
   1. use standard, controlled vocabulary terms
6. habitat

Leave rights and License (they belong in the AC file) in the occurrence.txt file. David says to leave them in. Then we will need to clarify <http://www.vertnet.org/resources/datalicensingguide.html>. Clarify rights / attribution for a data set (expectations) vs. the copyright options for media.

Now, we need to do the same with the **metadata file**, and **ac** file (although we agreed not to make this one complicated).

For metadata, we need a template, see word document. (BISON uses this method). Derek sent this template (based on the IPT fields).

**Audubon Media**  (see multimedia.txt)

For ac, see acMapping.txt file in sample data folder. Simple multimedia is all dcterms and this sample data was mapped from Simple multimedia to Audubon Media.

multimedia.txt file

Not many issues in this file. Should map easily. The *modified* field data values are in error, and is unclear. Data should reflect <http://en.wikipedia.org/wiki/ISO_8601> format and refer to the last time the media resource itself was modified (not the metadata for the resource).

|  |  |
| --- | --- |
| **Date and time (current at page generation) expressed according to ISO 8601:** | |
| Date: | **2015-01-06** |
| Combined date and time in [UTC](http://en.wikipedia.org/wiki/UTC): | **2015-01-06T03:51:57+00:00** **2015-01-06T03:51:57Z** |
| Week: | **2015-W02** |
| Date with week number: | **2015-W02-2** |
| Ordinal date: | **2015-006** |

Purpose to highlight a new soon-to-be released standard and provide an opportunity to talk about alternative “core” files (as it may be possible that this may be a new Core type and a potential one-to-many situation too.

Determination History

determinations.txt

No intentional errors in this file. It is here to represent how this type of information can be shared (1 to many example).

**Notes for issues seen in user data**

determinations

can identifiedBy and dateIdentified be suppositions?

assumed to be collector, assumed to be collection date

Other issues.

1. David S. to provide a second dataset (for use on day 2 for those that do not bring their own data)…
2. See <https://www.idigbio.org/wiki/index.php/Data_Ingestion_Guidance> for more common data issues.
3. From Derek. Workflow observation and question.
   1. public | not published – can still edit metadata w/o repeating the process
   2. published | no edits
4. resources / file size limits
   1. Question: what if your CSV is a million records? A scalability question. Are we going to demo using SQL to connect to, and query to create a dataset? (in this use case). 100 MB overhead limit even with SQL. (1.5 million – BISON). Tim suggested zipping, break-up individual files. (VertNet goes direct to GBIF).