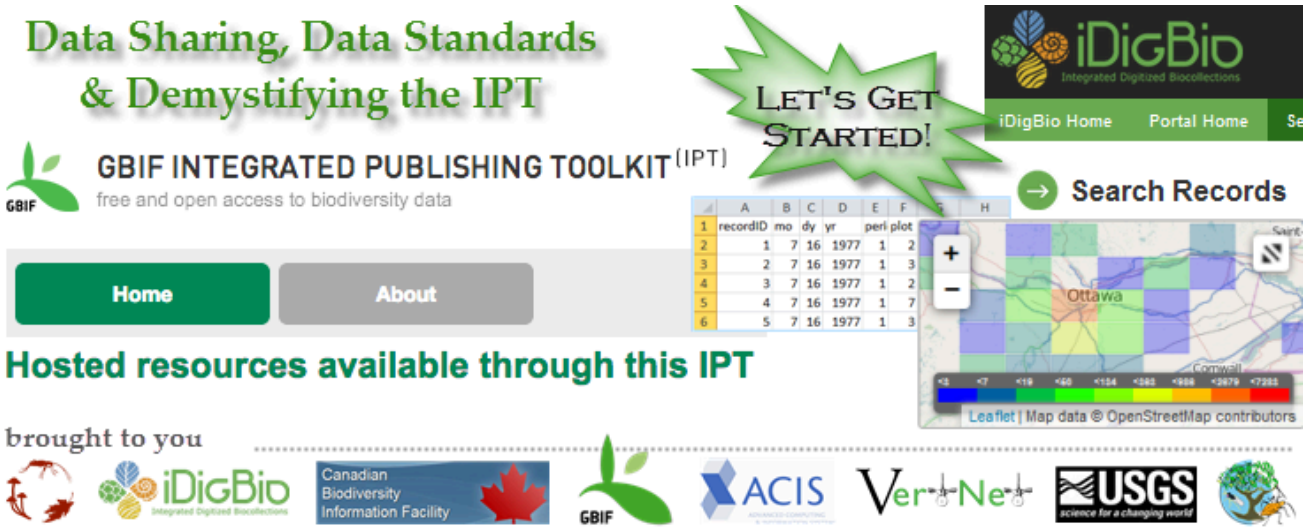

INTRODUCTION

January 13 - 14, 2015
Data Sharing, Data Standards, and Demystifying the GBIF IPT
Two locations: iDigBio at University of Florida, Gainesville and Agriculture and Agri-Food Canada and CBIF in Ottawa, Canada
Workshop Wiki: Data Sharing, Data Standards, and Demystifying the IPT

Welcome to this shared note-taking space. Please add your thoughts, insights, relevant links, and questions here for use during and after the two-day workshop. Note this is also a way to have a conversation with all workshop participants :-) in both locations.

Twitter Hashtag: #IPTWorkshop (https://twitter.com/hashtag/IPTWorkshop)

INDEX

AGENDA (Including questions per session)

Day 1

8:30 - 9:00 local logistics

9:00 - 9:20 1A introduction to the workshop

Volunteers
Day 2 format, opportunities

9:20 - 10:15 1B Theory. publishing basic primary biodiversity data: IPT and other methods.
Alberto Gonzalez-Talavan (GBIF)
1B: Theory: publishing basic primary biodiversity data: IPT and other methods.

DP: Alberto, what core types can we have in the new IPT version?
DP: I learned something :-) we can publish a description of a dataset (metadata), without publishing the dataset.
DPS- new version of IPT has occurrence, checklist, and metadata - same as current production version. No sign that there is support for other core types. I see core types like Event and Material Sample in admin, not sure where a new dataset is created that uses those cores.
DP: I think there is the possibility now to configure the IPT for whatever core type one wants, but i am not certain of how this works. Email from Kyle Braak points me to: the possibility to add new custom cores to the IPT. Instructions can be found in the IPT wiki here: https://code.google.com/p/gbif-providertoolkit/wiki/IPT2Core Kyle writes: The next release v2.2 will still ship with the two default cores: occurrence and taxon.

10:15 - 10:45 break

10:45 - 11:40 1C Theory: What are the metadata? David Shorthouse (Canadensys)
1C: Theory: What are the metadata?

Rights over biodiversity data in Europe. Recent paper: http://dx.doi.org/10.3897/zookeys.414.7717

no spaces in short names for resources
these short names end up in the URL created in the IPT as a link to your dataset, so give some thought to what you'd like them to be.

please fill out as much as you can -- this is the key to your data being discovered, used, re-used, cited, attributed, and counted!

AML: Be aware that metadata has to be compatible with the information in the dataset (ex geographic coverage indicated in the metadata vs GIS coordinates in the dataset, taxonomic coverage, etc). Is there a way/tool to verify this easily? (I think David said there wasn't)

11:40 - 12:00 1D Theory: publishing with the IPT Laura Russell

Please put your questions here:

DM: I am following remote from home - is it possible to log into my ipt instance from off site? Thanks
Deb - yes. You can. Please see instructions in the Private Adobe Connect Chat -- I'm going there now to send you instructions.

to Laura: what is ISO?
what is character encoding?
note to use UTF-8

Text editing programs, recommended for speed and capabilities:

- Notepad++ (http://notepad-plus-plus.org/) , file size limitation
- Sublime Text (http://www.sublimetext.com/) , can open very large files
- Text Wrangler (Mac OS)
- Atom (http://atom.io/)

12:00 - 1:00 lunch (provided)

1:00 - 1:30 1D continued. Demo and hands-on exercise for publishing with the IPT

DPS - We discovered browser version issues with file upload in the IPT in Ottawa. IE9 does not work & that's not unexpected. The issue we had here was with the map not loading b/c it uses a script (people had to "allow" in order to use the map to share geographic coverage.

DM: If possible, could the exercise be placed in the Gainesville drop box as well? Thanks
sure will do

Filters - will let you just publish certain records based on your selection.

Hi All, see the 1D .docx file in the Ottawa / or Gainesville Dataset Drop Boxes

1:30 - 2:30 1E Theory. Complex primary biodiversity data, extensions

DM: When possible, could the second exercise be placed in the Gainesville drop box as well? Many thanks

2:30 - 3:00 break

3:00 - 4:30 1E continued. Demo and hands-on exercise for complex biodiversity data

4:30 - 5:00 1F wrap-up and review for tomorrow.

[for Gainesville participants] 6 PM meet in Lobby for **Night at the Museum**

Day 2

8:30 - 9:00 check in and set up

9:00 - 10:15 2A Open practical session (participant designed around data sets brought)

participant data sets

10:15 - 10:45 break

10:45 - 12:00 2B Open practical session

producing a DwC-A file for publishing

12:00 - 1:00 lunch (provided)

1:00 - 2:30 2C Admin functions and user management in the IPT

2:30 - 3:00 break

3:00 - 4:00 2D Collaboration and the way forward

4:00 - 4:30 2E Summary of webinar, workshop, evaluation link, pay-it-forward feedback

---

NOTES

--Tues Jan 13 2015 9am --

--1A. Intro / Overview / Adobe Connect / Goals of Workshop--

Workshop Welcome and Introduction:
Data Standards, Data Sharing, and Demystifying the GBIF Integrated Publishing Toolkit (IPT)

Debbie Paul, David Shorthouse, James Macklin, et al
iDigBio - Gainesville, Florida, Canadensys, and CBIF and Agriculture and Agri-Food, Ottawa, Canada

Brought to you by: (in alphabetical order) Reed Beaman (NSF), Cathy Bester (iDigBio),Kyle Braak (GBIF), Matt Collins (ACIS - iDigBio), Shari Ellis (iDigBio), Alberto González-Talaván (GBIF), Chris Lewis (CBIF), Anissa Lybaert (CBIF), Kevin Love (iDigBio), James Macklin (CBIF), Derek Masaki (USGS - BISON), Andrea Matsunaga (ACIS - iDigBio), Joanna McCaffrey (iDigBio), Deborah Paul (FSU - iDigBio), Bénédicte Rivière, Laura Russell (VertNet), Katja Seltmann (TTD-TCN), David Shorthouse (Canadensys), Dan Stoner (ACIS - iDigBio), Alex Thompson (ACIS - iDigBio)

iDigBio – Integrated Digitized Biocollections

10 year effort to digitize and mobilize the scientific information associated with vouchered specimens held in U.S. research collections (Larry Page – PI)

Project Management (PM) and PM skills development by David Jennings

Provide for cyberinfrastructure
needs (Jose Fortes et al.)

Effective digitization
standards and workflows
(Greg Riccardi et al.)

Develop research use cases and collaborations (Pam Soltis et al.)

Education and outreach (Bruce McFadden et al.)

Plan for long-term sustainability of the national digitization effort

Estimated 1 billion specimens
in 1,600 collections in the US over 203 institutions

Thematic Collections Networks (2 of 13)…Lichens/Bryophytes, Tri Trophic Relationships

Need two volunteers - one thing you learned, how you will implement, how you will pay it forward. Two from Gainesville, two from Ottawa - each create one slide and spend 5 minutes for discussion.

Night at the Museum planned for Gainesville participants, 8pm

Wiki URL http://tinyurl.com/iptworkshop

A webinar about the installation of the tool was held last week. The feedback through the survey was very positive.

Workshop goals

- Enhance your skills / knowledge for getting a dataset into a standard format and improving and extending that data
- Show how to use the GBIF IPT software, as one way to share data
- (but not the only way)
- Enhance your Darwin Core and Data Sharing Standards knowledge
- You become ambassadors for data standards and data standards development
- Clarify for you, just what is in a Darwin Core Archive file (DwC-A)
- Including extensions, what are they, why have them?
- Teach you how to create or update Darwin Core Archive files using the IPT
- Enhance your knowledge about where data goes and how it gets there once it leaves your collection
- Show how a taxonomist with an occurrence dataset can publish a dataset as a DwC-A
- Help you mobilize your data

Useful Links:

Wiki
Google Doc
Participant list
Twitter: #IPTWorkshop
Twitter: https://twitter.com/hashtag/iptworkshop/

Data and Metadata. It's about discovery and data re/use. It's about feedback and accountability. It's about credit and attribution. (It's about curation not ownership). Make sure your data's not under a rock.

Ottawa Introduction

Canadian Biodiversity Information Facility - Canadensys

David Shorthouse - Biodiversity Information Manager. Responsibilities, biodiversity data, checklists.

Developer, Christian (Last Name?), Explorer product.

--Tues Jan 13 2015 920am --

--1B Publishing Primary Biodiversity Data--

Alberto González-Talaván1
Data Sharing, Data Standards, and Demystifying the IPT
Gainesville, FL, USA. 13 January 2015

Structure of Presentation

What is biodiversity data?
Rationale for biodiversity data publishing
Data publishing procedure
Data exchange standards
The technical infrastructure
Data publishing software
GBIF Integrated Publishing Toolkit

What is biodiversity data?

Digital text or multimedia data record detailing facts about the instance of occurrence of an organism, i.e. on the what, where, when, how and by whom of the occurrence and the recording.

Specimen Labels - describe location, species. Traditional concept of biodiversity data.

Other Sources: Journals, Checklists, Assessments, Urban Biodiversity, Citizen Science, Genetics, Camera Traps, Satellite Images.

The many sources of biodiversity are similar but also have differences that introduce challenges to those curating and making it accessible.

Rationale for biodiversity data publishing.

"Publishing" refers to making biodiversity datasets publicly accessible and discoverable, in a standardized form, via an access point, typically a web address (a URL).

Reference.
Chapman, A.D., 2005, Uses of Primary Species-Occurrence Data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. 100 pp. ISBN: 87-92020-01-1. http://www-old.gbif.org/orc/?doc_id=1300

Uses of Data:

1. Agriculture, Forestry, Fisheries and Mining
2. Health and Public Safety
3. Bioprospecting
4. Forensics
5. Border Control and Wildlife Trade
6. Education and Public Outreach
7. Ecotourism and Recreational Activities
8. Society and Politics
9. Human Infrastructure Planning

Three Sources for Review - Find a use case related to data that you work with. Share in chat or document.Use the following sources:

Featured data section in GBIF.org
http://www.gbif.org/newsroom/uses

GBIF Public Library in Mendeley
http://goo.gl/btrzDa
(requires Mendeley account)

GBIF Science Reviews
http://www.gbif.org/resources/3094

Rationale for Data Publishing: data quality

Verbatim data, visualization of messy, as provided data - locations indicate issues (missing lat/lon)

Processed data, visualization of cleaned data after spatial issues addressed/corrected.

"We believe that the lack of incentive similar to the impact factor for scholarly publication remains a major impediment to the provision of free and open access to biodiversity data"

GBIF Data Publishing Framework Task Group

"We believe that the lack of incentive similar to the impact factor for scholarly publication remains a major impediment to the provision of free and open access to biodiversity data"

GBIF Data Publishing Framework Task Group

Rationale for Data Publishing: benefits

Data Paper
A scholarly publication of searchable metadata document describing a dataset, or a group of datasets

Promote and publicize the existence of the data
Provide scholarly credit to data publishers through citable journal publications
Describe the data in a structured human-readable form

Data Publishing Procedure:

- Prioritization and planning
- Capture

- Curation
- Export and preparation
- Publishing

Comment in Gainesville - include in data publishing workflow, the work in the field of getting data (getting dirty and muddy).

Data Exchange Standards. Coordinating activity.

www.tdwg.org

ABCD Access to Biological Collection Data
DwC Darwin Core
DwC-A Darwin Core Archive
NCD Natural Collection Descriptions
AC Audubon Core
Biodiversity Data Standards

Example Darwin Core (DwC)

Simple Darwin Core - flat file, few restrictions
http://rs.tdwg.org/dwc/terms/simple/index.htm

For more complex data - Darwin Core Archive (DwC-A)
http://rs.tdwg.org/dwc/terms/guides/text/index.htm

Archive elements: taxon core, extensions, eml.xml, meta.xml

The technical infrastructure: Summary

Workflow:

1. Determine type of data
2. Select a tool: spreadsheet, integrated publishing toolkit, DwC-A assistant
3. Register with GBIF
4. Publish data
5. Discovery

The technical infrastructure: processing

Video - Tim Robertson describing making data available using GBIF

Official launch of the new GBIF.org

http://vimeo.com/77782067 - from 24:15 to 27:00

Data publishing software: some options

1. Integrated Publishing Toolkit (high technical capacity, low data management capacity)
2. Create your own DwC-A
3. Publish with Spreadsheets

4. Data Hosting Center (low technical capacity, high data management capacity)

The GBIF Integrated Publishing Toolkit: Vision

- A single platform allowing the sharing of, Primary biodiversity data, Species name information, Dataset descriptions (metadata)
- The ability to register with GBIF, Technical contact information, E.g. Internet URLs
- Physical contact information, E.g. telephone details
- Institutional affiliations, Attribution
- Connect , Databases

Upload text files, Lower the technical threshold for participation

Flexibility to accommodate data extensions

Support efficient and simple transfer of content

An open source project

Questions:

Question (Gainesville) regarding languages and datasets - missing Russian and Japanese character sets. Data in other languages represents a challenge - the system should allow sharing of information in other character encodings.

Comment (Gainesville) - search produced result where the title of paper is translated into English, body of paper remains in Spanish.

Question (Ottawa) - citation and usage. In library world, AltMetrics and ImpactStory. Is GBIF considering similar applications for use metrics. In theory, GBIF implementation of DOIs/UIDs will provide ability to track. IPT Version 2.2 already supports DOIs, GBIF Portal will soon.

Comment (AdobeConnect) - OBIS recommends citations produced in English first.

Question (Gainesville)- for the citation section of the instance metadata, can anyone just create a DOI that will permanently link to the relevant data?

Question (Gainesville)- what are (where can I find) best practices/standards for fields that are missing data? When to use null vs zero? Leaving fields blank may cause issues when a dataset is being actively added to, but placeholders are not recommended for publication.

Note: DataCite DOIs in Canada: NRC/CISTI ("Datacite Canada") is the Canadian DataCite registrar:

- https://www.cisti-icist.nrc-cnrc.gc.ca/eng/services/cisti/datacite-canada/index.html
- FAQ: https://www.cisti-icist.nrc-cnrc.gc.ca/eng/services/cisti/datacite-canada/info-prospective-clients.html
- "***Question: Who can use DataCite Canada services? Answer:*** *Organizations that manage research data are eligible for an account with DataCite Canada. This includes universities, libraries, government departments and other research data centres. To receive a DataCite Canada account, data centres must have the authority and responsibility to store the data as well as to maintain access to the data to which the DOIs will be assigned.*" -- https://www.cisti-icist.nrc-cnrc.gc.ca/eng/services/cisti/datacite-canada/info-prospective-clients.html#q6

Comment (AdobeConnect) - OBIS recommends citations produced in English first.

--Tues Jan 13 2015 1015am --

--Break (with questions) --

Comment (AdobeConnect) from Mary Kennedy: Quick question about language - is there a site where all the DwC terms are translated? i found a page where basis of record is described in many languages but it would be more useful if there was a page with all terms and definitions in say French or Spanish. The page that i found with terms in german, spanish, french, japanese and chinese is terms.tdwg.org/wiki/dwc:basisOfrecord (perhaps this question can be addressed later in the workshop when DwC terms are discussed)

Follow up from Alberto: The translations are community provided so probably no one has stepped forward to translate the terms in to those languages. Also, the translations can be piecemeal based on the time and interests of the person volunteering to do the translation.

--Tues Jan 13 2015 1045am --

--1C Theory: What Are The Metadata --

David Shorthouse

What are the metadata?

« A set of data that describes and gives information about other data »

Metadata can take many forms.

One person’s metadata is another’s data.

Traditionally, metadata can be unstructured (such as a collection of keywords) or structured documents (such as XML with schema), which may include controlled vocabularies.

Why do we need metadata?

We’re forgetful
We misplace things
We’re poor communicators
Our science is dynamic

We use machines to help us.

What aren't metadata for us?

Darwin Core standard, http://rs.tdwg.org/dwc/terms/
WHO, WHAT, WHERE, WHEN
recordedBy, scientificName, stateProvince, eventDate

Not Metadata - Darwin Core Standard
A stable glossary of terms based on taxa & their occurrence in nature as documented by observations, specimens, and samples

180 terms split up among 11 classes
Some terms borrow from Dublin Core
Managed by the Darwin Core Task Group within Biodiversity Information Standards (TDWG)

What Are Our Metadata?
Data that describe a biodiversity dataset, provide the context
We want to facilitate:
Search & retrieval
Reuse: licensing, community norms
Provenance & attribution
Communication: contacts, responsible parties
Expressions of fitness-for-use

Metadata Search and Retrieval
Primary reason for structured metadata is to faciliate appropriate exposure. That means our data are capable of being acquired by aggregators like GBIF, iDigBio, and Canadensys & that our data can be placed alongside all others' data without losing their provenance.

Metadata search and Retrieval
Some of the basic metadata elements include title, description (abstract), a broad description of the geographic coverage (bounding box), higher taxonomy (if relevant and appropriate), temporal coverage, and keywords from thesauri and using Flickr-style, freely produced keywords.

Metadata – Reuse: Attribution & Licensing
CC Zero
CC By
CC By NC

Background – Free and Open Access
Bouchout declaration
http://bouchoutdeclaration.org

Background – Community Norms

http://www.canadensys.net/about/norms

Background – Community Norms
http://www.vertnet.org/resources/norms.html

Canadensys – uses metadata generated in IPT and transforms it for submission to DataCite.
Soon to be released, new version of IPT will have the capacity to assign DOIs to datasets in hosted IPTs.
GBIF registration assigns a UUID to a dataset, included in the metadata & also in the GBIF's URL for the resource when harvested

Darwin core archive
Star schema
http://rs.tdwg.org/dwc/terms/guides/text/index.htm

EML

Developed for ecologists by ecologists

Lead by NCEAS, DataONE
XML schema for the structural expression of metadata
Basic subset of EML used by GBIF, « eml-gbif-profile », http://www.gbif.org/resources/2559
IPT automagically creates eml.xml from web form data when dataset is published

Where Does the Metadata Go?
GBIF data portal and registry, iDigBio, Canadensys, other aggregators
GlobalNames
Core of a data paper: ZooKeys, PhytoKeys, MycoKeys

Questions

Question: Ottawa - How extensible is the GBIF file for metadata? You can extend the EML XML yourself. The registered schema is fixed - though amendments can be made through proper channels.

Comment: remote Participant - Other versions of metadata can be added to the IPT 'additional metadata' tab as an alternative identifier

Questions:

1. How extensible is the EML file? Can we share more metadata than those included in the subset?
   You can add additional elements, but not within the IPT. You need to edit it yourself. There are software that can load EML, and can be used to analyse their contents and probably extend it.

Demo Exercise

Each attendee has access to test IPT instance

Create a sample resource.

Create an occurrence dataset.

Make a new resource.

Shortname is a quasi identifier, it will appear in the URL of the resource, make it short/concise.

Shortname: ttrs-ferns-specimens
Type: Occurrence

(Note) You can use the same shortname as the demo example - everyone has their own IPT instance.

(Note) You cannot have spaces in shortname, use underscores or hyphens.

Should see "Overview:ttrs-ferns-specimens"

Select edit under Metadata

Fill out Basic Metadata:

Change title to a more human readable entry. "Tall Timbers Research Station Ferns"

Description: "Description of tall timbers dataset"

Can specify metadata language, resource language. Make both English.

Type: Occurrence; Subtype:Specimen;

Fill out Resource Contact and submit.
NOTE: if you use upper case letters in your email address you will get an error when you press SAVE
Plus, remember to press SAVE at the bottom of each page before you jump to the next tab or else you may loose all your edits.

If required fields are not entered, error will indicate .

Fill out other sections as best you can. Basic Metadata; Geographic Coverage; Taxonomic Coverage; Temporal Coverage…

Questions

Question: Versioning. Newer version of IPT will enable more efficient at managing versions.

Question: How to associate media with data. Extensions are available for media associated with records.

Question: How to handle multiple geospatial coverages, eg. area in Australia and area in Canada. Probably best to use global coverage in metadata and specify location in dataset.

Question: How to manage metadata quality. Would be good to have feedback / comments provided by community or users.
in OBIS we have created a manual where we suggest the addition of a few canned phrases that can or should be added to different sections in the metadata.

If you need help to figure out what goes in each field, you can check the metadata profile but also in the inline help that is available in the IPT. Look for the little blue circles with an ‘i’ (info) on them.
http://www.gbif.org/orc/?doc_id=2820

--Tues Jan 13 2015 1140am--

--1D Publishing with the IPT --

Note: IPT shortname, once submitted, is permanent and cannot be easily changed.

The two sections that we will be talking about is:

- Source data; and
- Mapping

Source data

In the source data, we will have two options:

1. A file upload: compressed or uncompressed. There’s a limit of 100 Mb for uncompressed files. There are different formats accepted.
2. A database connection. Again different databases are supported: MS SQL Server, MySQL, Oracle, PostgreSQL, Sybase, ODBC...

You need to know where your data is. If you are using a managment software, it is inside that software. Other people may have custom databases of simple datasheets.

There is a number of required fields that are formally required: you cannot publish data if they are missing. Others are not formally required but effectively required to publish: identifier (occurrenceID), Nature of the record (basisOfRecord), Institution Acronym…

There are hundreds of fields in the DwC, grouped in categories. Different initiatives have a different list of required fields. Everyone will encourage to share as much as you can. A common average is sharing around 50 fields. A more adequate number would be around 80.

If you are restricting access to certain fields or to certain records, you should describe it in your metadata, so those who read the comments know why you are restricting access to, and then they can contact the origin one to one.

There is no better moment to share data than now. Don't wait till you have made certain checks. Publish them, improve them, republish…

Try to follow the recommendations provided with the DwC guide. Use the recommended vocabularies.

Aggregators can use their experience to help you to identify and solve data quality issues.

Character encoding can be a source of problems. UTF-8 is the recommended. There are documentation on the internet justifying this decision. It has to be consistent when we assign it to the source and when reporting to the aggregator. Otherwise there will be errors shown in the search portals or when people access the data. It is also essential that whoever opens the file also assigns to it the correct encoding.
IPT will do its best to recognize the encoding, but it is important to check it.

Mapping

There are two cores: Taxon Core and Occurrence Core. There will be others coming.

Questions

QUESTION: Do people actually have time/money to address the error reports from VertNet. I think I have something like 8000 "errors." - Rob Faucett

From Deb: Rob, this is a great question. What kinds of errors? Is there anything we in the community can do to help you fix these? Is a matter of time? Or is there some programming or scripting that could help?

QUESTION: If you have multiple excel files, can you upload all of them in the IPT ? Yes, you can upload multiple files to map

QUESTION: If you have a database connection, what is it using at the source? is it static?
is it a

Extracting Source Data – Format

What formats are usable within the IPT?

- File upload (uncompressed up to 100MB)
- Delimited text files
- Excel
- Database connection

MS SQL Server, MySQL, ODBC (Sun Java 5)

Oracle, PostgreSQL, Sybase

Locations of Source Data:

Collections Management Software
Desktop Databases
Spreadsheet Applications

Extracting Source Data - Fields

"Required" fields (for an occurrence file)
Identifier (occurrenceID)*
Nature of the data record (basisOfRecord)*
Institution acronym (institutionCode)
Collection short name (collectionCode)
Catalog number (catalogNumber)
Taxon name (scientificName)
*required by the IPT

Extracting Data - Fields

Every aggregator is going to have their own "required" list
Share as much as you can
Simple DwC has 164 terms
Most institutions are sharing between ¼ to ½ the terms
Encumber data when necessary
Protected species
Protected localities

Extracting Source Data - Data Quality

Should data be "perfect" before publishing?

"Errors using inadequate data are much less than those using no data at all."
-- William Kenneth Richmond (1969), The Education Industry

Extracting Source Data - Data Quality

- Follow recommendations provided in the Darwin Core Guide and use recommended vocabularies. (http://rs.tdwg.org/dwc/terms/index.htm)
- Be consistent
- Dates formatted as ISO 8601:2004(E)
- Don't use placeholders
- Know when to use null vs zeros
- Ask aggregators for feedback

Extracting Source Data - Character Encoding

UTF-8 is recommended
http://www.w3.org/International/getting-started/characters

IPT and Character Encoding

IPT will try to recognize the encoding on upload or database connection. If you know it to be different you have the opportunity to set the proper encoding.
IPT will use UTF-8 encoding when the DwC Archive is published.

Use Better Text Editors

Notepad++ (Windows)

Sublime (Windows, Mac)
TextWrangler (Mac)

IPT Mapping - Mapping Cores(2 Cores)

Taxon Core

The category of information pertaining to taxonomic names, taxon name usages, or taxon concepts. Updated Nov 2011 with newly ratified terms.

Occurrence Core

The category of information pertaining to evidence of an occurrence in nature, in a collection, or in a dataset (specimen, observation, etc.). Updated 2 Apr 2014 with materialSampleID.

Questions

Question: When you have several excel files, can you upload the multiple files and merge the data. It is possible with text files - unknown if possible with excel. Suggestion to test first.

Question: Database connections, does IPT use view tables or is it dynamic. You can connect to source table, can connect to view.

Question: Placeholders. Collection dates are important - sometimes when missing, data operator will use placeholder like" not on record label" , to indicate field may be completed in the future following further investigation. Suggestion to put extended information in comments field.
Question: what about using verbatimEventDate field to indicate that 'date ?
Comment from AC: You can perform any query in IPT to your database, specifying an SQL query. Thus, it allows you to dynamically specify transformations or consult a predefined view. Use of specific functions, joins, and inner selects depend on the database technology being used.

--Tues Jan 13 2015 12pm--

--Lunch --

--Tues Jan 13 2015 1pm--

--1D: Theory Publishing with the IPT cont'd demo / exercise--

Adding a data source and mapping data

File upload. For text files, in source data use "Choose File"
Data upload preview useful for looking at source for errors/misalignment

For database connections, use connect to database option - analyze data option is useful. You need connection information and permissions to connect to desired database resource.

IPT Darwin Core Mappings - use add button, select source data.

IPT Automapping - will attempt to recognize and map fields from source data. Will provide alerts to detected missing fields. Will detect improperly formatted date fields. A shortcut to datasource is provided. All fields that do not automap will be listed at bottom of page -for example, if terms do not match.

IPT DC Mapping - Dropdowns -some terms like type have recommended vocabularies that you can use if the data is not present in your source.

Darwin Core Mappings - Preview Data. Show mapped fields with data of first 5 records.

Fill In Values - if no mapping provided, a value can be supplied. Supply here only if it is the same for every record in the data source.

Translations. Example state abbreviations can be translated , eg. AL >> Alabama. Ok for small number of values, for larger number better to change at the data source itself.

Filters - at automapping step a filter can be used to hold back records based on specific field value.

--Exercise: Publishing with the IPT--

Adding a data source and mapping data

Data Sharing, Data Standards, and Demystifying the IPT Workshop – Day 1

1. Continue building on the IPT resource created in the Meta Data exercise.
2. Upload the Sample Occurrence dataset from the participant DropBox folder as Source Data in the IPT.
3. Preview the data from the data source and become familiar with the supplied fields and the data within the fields.
4. Save the data source.
5. Create an Occurrence Mapping using the data source uploaded in step 2.
6. Finish mapping the unmapped fields.
7. Use a standard vocabulary as a mapping.
8. Use a fill-in value as a mapping.
9. Set up a value translation.
10. Note which terms you mapped for steps 7, 8, 9. (We will examine these on Day 2.)
11. Save the mappings, making sure that required fields are set and that no alerts appear at the top of the mappings.

Question: If there are multiple fields that need to be concatenated in a data source, can that be managed in the IPT. At this point, no.

Comment: Sometimes the source datafile has column headers that do not match the DwC terms and one has to manually map these when uploaded to the IPT. I believe tho that if the header is close then the IPT will recognize the match. Example Order1=order; year1=year. sometimes terms like order and year cause issues in source databases so try adding 1 as a suffix.

--Tues Jan 13 2015 145pm--

--1D: Complex Primary Biodiversity Data--

Standards and sharing complex primary biodiversity data;and what is an extension anyway?

Standards are wonderful, everybody loves them. That's why there are so many.

The data landscape (silos)

Evolving standards
What are extensions, and why does Darwin Core need extensions?
One-to-many relationships

Identifiers are the key
Audubon Media Description
Darwin Core Identification History
Global Genome Biodiversity Network (GGBN)

We have lots of different types of data out there in the wild that we'd like to share. Importantly, we'd like to preserve and share the relationships between the data types. Darwin Core provides the basic foundation for sharing **core** information about natural history museum specimens. But, in order to share the many derivative types of information that may be created at the time of collection, to many years after collection of that specimen we need

1.A way to relate the supplementary data to the **core** data, and

2.We need the data standards to map our data to, for sharing.

3.(which includes the need for robust identifiers!).

Evolving Standards

http://xkcd.com/927/

Keep in mind that standards evolve.

Ideally, as with Darwin Core, standards are community developed. If you discover you have data that doesn't fit, the community needs your input – to develop the necessary standards to support that data.

It's similar to finding out that your database needs a new field so you don't stick everything in the notes.

Complex Primary Biodiversity Data

DwC does not provide fields for every possible type of data.

But you have lots of other types of data, right?

Introducing the extension

–http://tools.gbif.org/dwca-validator/extensions.do

–There are many! And (no doubt) more to come.

22 registered

23 under development

Examples

– Audubon Media Description (aka Audubon Core)
– Darwin Core Identification History
– Global Genome Biodiversity Network (GGBN) extensions

One Specimen So Many Kinds of Data - Darwin Core Extensions

ttp://www.morphbank.net/?id=551005

http://herbarium.bio.fsu.edu/view-specimen.php?RecordID=22924

Mast, A. R., A. Stuy, G. Nelson, A. Bugher, N. Weddington, J. Vega, K. Weismantel, D. S. Feller, and D. Paul. 2004 onward (continuously updated). Database of Florida State University's Robert K. Godfrey Herbarium. Website http://herbarium.bio.fsu.edu/ [accessed 06 Decembe 2013].

Defined and Registered with GBIF registry

Allow extension while retaining compatibility

Extensions are optional data files linked to core

A row in an extension file always references the core id corresponding to a taxon or taxon occurrence

Laura provided background for the next step: How to create a resource in the IPT in the section: 1D. Theory: publishing with the IPT data sources where she discussed issues of data quality, character encoding, and mapping, including a discussion of the different core types. Participants again logged in to follow along and map the data found in the sampleoccurrence.txt file. This is the core file in this example. The coreid, becomes the dwc:occurrenceID and is the identifier that links other files (image data, determination data, etc) back to the core file. The sampleoccurrence.txt file contains plant data, some of it is real, some of it customized for this example.

After this, the next step is to add the determination (aka verification, identification) history data, and lastly the multimedia data. To do this, each extension data set will need to be checked for data issues (data quality, standardization, enhancement), then the extension fields will need to be added in the IPT to the resource, the relevant text file uploaded to the IPT, followed by mapping. For these examples, there are very few errors in the samplemultimedia.txt and sampledeterminations.txt files.

Note that the meta.xml file is created for you by the IPT tool. There are other online tools that can help you do this in addition to the IPT (the dwc-a assistant). The meta.xml file contains the name of each file inside the zipped DwC-A file, and shows all the fields (elements, properties, column headers) inside each file, noting in which column a particular field appears.

Audubon Media Description - extension provides for description of associated media

Sharing media

What's in the image, recording, video?

Who took the photo, made the recording, created the SEM, CT scan?

Is the media under copyright? Or is it public domain?

Where can more / different formats of the media resource be found?

A given specimen may be photographed, once, or potentially many times and with different methods and processing techniques. We need to be able to share all the relevant media desired and information about these media files were created. In this situation then, we have a one-to-many use case., that is, one specimen with many media files.

For media, there are several relevant extensions to Darwin Core, one is Simple Multimedia, another, used here is the Audubon Media Description (aka Audubon Core).

The AC has 10 groups of terms (see next slide), for providing richly detailed information about any given media object. If most of the information about what is in the media object can be found in the occurrence file, then very few fields can be used. Alternatively, if only this media file is to be shared, for media-only sharing, then this standard provides a broad and deep set of terms to make the accompanying media metadata quite detailed.

If we have this data, it aids discoverability, and provides methods for discovering if a given resource meets one's needs. This aids researchers when evaluating media data, and fitness-for its potential use.

Global Genome Biodiversity Network (GGBN) Extensions: require a Material Sample Core

The category of information pertaining to the physical results of a sampling (or subsampling) event. In biological collections, the material sample is typically collected, and either preserved or destructively processed. Created 2 Apr 2014 with all Simple Darwin Core ratified terms.

For GGBN info see http://ggbn.org/

Question from AC: Question: what fields should one populate if tracking one animal. will have several records for the same taxon but different location and date. IndividualID doesn't seem to be in DwC anymore. not sure when it would be appropriate to ask this question. Perhaps this raises the question what happens when DwC terms are updated and old terms dropped or replaced by new terms (example was the term rights replaced with licence in December?

Followup from Andrea: IndividualID has been replaced with organismID. When changes in standards take place, you need to remap your data to use the updated term. Sometimes there is a need to also update your data in addition to remap the term.

Question from AC: Question from Ottawa: Are there extensions for hosts or substrates?

Followup from Alberto: There is a list of registered extensions at http://tools.gbif.org/dwca-validator/extensions.do

Followup from Andrea: To my knowledge there are no extensions for host or substrate. That said, iDigBio receives a set of host terms that has been defined by the TTD TCN. There is currently an effort to make those terms into semantics standards. For substrates, most users seem to find the substrate term in DwC to be sufficient.

--Tues Jan 13 2015 2pm--

--1D: Complex Primary Biodiversity Data - Demo--

Adding data sources and mapping for the multimedia extension.

In IPT choose multimedia file.

Review just like the core file previously update.

Now there will be two files, occurrence and media extension.

To use extensions, you need to enable in the IPT. By default only core is loaded.

Demo example, use Audubon Media Extension file. Also add Identification History.

Once extensions are added to IPT, fields for chosen extensions are available in the dropdown list.

--Tues Jan 13 2015 220pm--
--Break--

--Tues Jan 13 2015 250pm--

--1D: Complex Primary Biodiversity Data - Exercise--

Exercise: Complex Primary Biodiversity Data

Adding data sources and mapping for the multimedia (Audubon Core) and identification histories extensions

Data Sharing, Data Standards, and Demystifying the IPT Workshop – Day 1

1. Continue building on the IPT resource created in Data Source/Mappings exercise.
2. Upload the Sample Multimedia dataset from the participant DropBox folder as Source Data in the IPT.
3. Preview the data from the data source and become familiar with the supplied fields and the data within the fields for the new data source.
4. Save the data source.
5. Upload the Sample Determinations History dataset from the participant DropBox folder as Source Data in the IPT.
6. Preview the data from the data source and become familiar with the supplied fields and the data within the fields for the new data source.
7. Save the data source.
8. Using Admin functions, add the Audubon Core and Identification History extensions.
9. Create an Audubon Core Mapping using the data source uploaded in step 2.
10. Finish mapping the unmapped fields.
11. Save the mappings, making sure that required fields are set and that no alerts appear at the top of the mappings.
12. Create an Identification Histories Mapping using the data source uploaded in step 5.
13. Finish mapping the unmapped fields.
14. Save the mappings, making sure that required fields are set and that no alerts appear at the top of the mappings.

Possible topics for discussion or questions ask people in the room:

Simplify IPT data process; Unique IDs; Extensions; DOIs for datasets; Rights, licenses, community norms; Specify and Symbiota

Unconference - need to vote on topics for Wednesday session. Topics - Filemaker Pro; Break out groups (paleo,botany, marine); Live SQL demo; Spreadsheets; Georeferencing; Data Cleaning

--Tues Jan 13 2015 340pm--

--Demo Publishing IPT Data--

3 source files, mapped.

Publish dataset. Under “Publish Release” click publish. Check fails with errors - duplicate ID. Which file has the duplicate?

Can keep publishing after each modification to your IPT resource, as a data check step (eg. after adding metadata, after each extension being added). This can make it easier to track where errors are.

Excel - open text file with encoding specified. Open file, select text type, select “65001 Unicode (UTF-8)”, highlight all columns and text type UTF-8 will be enforced.

Excel find duplicate values - select columns, conditional formatting>>highlight duplicates. Duplicates will show in red.

Correct errors and publish again.

Make resource public.

Test download - click on link and extract.

--Tues Jan 13 2015 415pm--

--Open Discussion--

Steps for finding Duplicates in Excel on Mac: (this is one method, there might be others)

- Data
- Filter
- Advanced Filter
- Check option to show “Unique records only”
    - This method will collapse the view of duplicates within the range of data selected
    - To show all the records again, Data -> Filter -> Show All

Question (Gainesville)-Where can I find the openRefine manual that Alberto mentioned?
http://www.gbif.org/resources/12358

Matt proposed a lunchtime demo on Wednesday. Using OpenRefine for data cleaning.

Comment: Data cleaning - tools, Python, R, Open Office, spreadsheets, Curator (Ottowa suggestion)

Comment: Mary Kennedy (OBIS) running distinct names through WORMS (names registry). World Register of Marine Species

Mary Kennedy: any thoughts about discussing QC of the data before uploading it to the IPT? for marine taxa we highly recommend running a list of distinct names thru the WoRMS taxon match tool - this helps clean up spelling variations and helps map to classification hierarchy.

Mary Kennedy: plus we highly recommend extracting a list of distinct coordinates and plotting on a map to confirm that they are all in the correct area. etc.

OBIS has developped a manual that suggests fields that should be checked and suggests tools that could be used so if anyone is working with marine taxa they should have a look at a few of the tools used by lifewatch.be

Workflow for data cleaning - various stored procedures in database.

Alberto - working group for data quality, identifying data quality checks.
http://community.gbif.org/pg/groups/21292/gbiftdwg-biodiversity-data-quality-interest-group/

TDWG now Biodiversity Information Standards, guidelines on data quality, data cleaning procedures.

Question: provide name but don't populate classification fields…
When I work on a marine dataset i map all my names to WoRMS codes and then i map these codes to a table containing a matrix from WoRMS with all their classification - this results in a record that contains'my name' plus the WoiRMS classification. I don't have to worry about populating these fields - it is semi-automated!

GBIF processing adds valid name and higher taxonomy, that is available when data is downloaded from GBIF after upload from IPT.

--Questions and Minute Card Entries From Day 1 Jan13 2015--

Question: Mapping

1. When do we map to "verbatim-xxx" rather than "xxx"?

   You map to "verbatim-xxx" when your data does not quite fulfil the requirements of the "xxx". Three noteworthy examples are: (a) in eventDate you should not have "Fall 2010" or "Dec-13-1990" since the best practice is to use ISO format dates "2010" and "1990-12-13" respectively; therefore you should use verbatimEventDate if you do not have dates in ISO format; (b) decimalLatitude as the name suggests should only contain data in decimal; for latitudes in degrees/seconds/minutes, you should use verbatimLatitude; (c) if you "clean/improve" your locality, for example to avoid abbreviations, you may want to keep the original description as it appears in the label for future comparison; in this case you would store the "clean" version in locality, and keep the original description in "verbatimLocality".

2. What if the field does not have data in it? Do we map it?

   It would be preferable to not map a field that contains no data.

What I learned

1. I knew nothing about how the IPT software worked so all that is what I learned, mapping, add data sets, core types, etc.
2. Learned difference between metadata and collection data, use of unique id#s.
3. Mapping source data.
4. Learned role of Audubon Core.
5. IPT metadata can serve as basis for a data paper.
6. Adding source data and metadata in the IPT.
7. Connect DB (PostgreSQL) to IPT with interactive mapping and translation.
8. IPT is a useful tool for mapping data into a particular format but it is also flexible and allows for inclusion to accommodate for different media types.
9. Basic needed data and info structure for uploading to the GBIF IPT.
10. Value of iDigBio.
11. How to import data file and map fields.

12. How to map data, sort of.
13. Datasets as well as collections can both be entered into IPT and it's best to get something in as it can always be updated.
14. Lots of great resources for tracking how people are using data from GBIF and published data - useful in showing value of collections.
15. I learned that the Audubon Core is a set of standards for multimedia and DwC is for metadata.
16. How to handle extensions - love it!
17. Hands on experience with IPT sample data.
18. How to start with IPT mapping.
19. Mapping fields in IPT.

What I am still not clear about

1. I am not sure how all this will be shared with GBIF or other portals, is it just an upload or something else?
    a. When you publish a dataset publicly, if your IPT has been registered with GBIF, GBIF will automatically fetch data from your IPT. To share the data with iDigBio, send an e-mail to data@idigbio.org for the data mobilization team with the link to your IPT RSS feed to learn about the existence of your IPT and iDigBio will periodically fetch data from your IPT.
2. How much overlap is planned with Symbiota? Seems many IPT features can be executed by Symbiota
    a. Even though there are clear overlaps from the user/data provider point of view between IPT and Symbiota, the goals of each tool are distinct. IPT is a more general tool for publishing various types of data, and very tied to what are defined by standards such as DwC. Symbiota is a tool focused on publishing specimen occurrences, their determinations and media, with the goal of simplifying the mapping process (push of a button publication of data).
    b. From Deb: Symbiota is a website, and a cloud-based database, that happens to package up data for you, to send to GBIF, if that's what you want to happen. They have their own custom DwC-A file. IPT is software you install, to help you map data and package data up into a DwC-A. It has the added feature of being software that enables you to make a folder on the web that is "Findable" by automated services. The "address" of the folder and the datasets are "registered" with GBIF so that they (and anyone else) can program another computer to go out and find your data and get it.
3. What's the difference between rights and licenses (at the dataset and record level) and where do community norms fit in?
    a. See here: https://www.idigbio.org/wiki/index.php/Data_Ingestion_Guidance#Complete_attribution_and_licensing
    b. The short story is
        i. Data should be public domain, CC0
        ii. Media can be protected, e.g., **CC BY-NC**
    c. Don't make it hard for the the user of your data to use it. For example, if they have selected 1000 records for a study and they all have some sort of different copyright, it will be impossible for them to clear the rights to use the data.
4. Best practices for creating occurrenceID.
    a. See below for UUID, or see here: https://www.idigbio.org/wiki/index.php/Data_Ingestion_Guidance#Specimen_metadata
    b. The short story is
        i. Use a UUID, guaranteed to be unique by probability
        ii. Use a locally unique ID, with a prefix that is known to be globally unique.
5. How does one create UUIDs?
    a. https://www.idigbio.org/wiki/images/0/03/GUIDgeneration.pdf instructions for creating a UUID in an Excel spreadsheet.
    b. SQL has a native datatype for UUID, create a new field in your DB of that type
    c. EMu and Specify both have native UUID identifier fields
6. When to use different extensions and vocabularies.

    a. First think at a high-level about the type of data that you are trying to publish: a list of specimen occurrences, a list of observation occurrences, a list of taxonomic names, a list of media files, etc. Then what are the additional information to the main list that is related to the main list in a one-to-many fashion. The main list is the core, while the additional information are the extensions you need.
7. How paleontology data works/is entered/ is mapped within IPT.
    a. There is a set of DwC terms that are often used by paleontology collections. These terms are grouped under GeologicalContext. You can simply have additional columns in your source data with information that can be mapped to these terms. http://rs.tdwg.org/dwc/terms/
8. I am still not clear on how all of these linked datasets can be discoverable through means other than GBIF, for example if I have genetic data in GBIF, how does the same data in GenBank link back?
    a. Map your comma-separated GenBank URLs to this field associatedSequences: http://rs.tdwg.org/dwc/terms/#associatedSequences
    b. From Deb. This is really a very good question, and a bit complex. Right now, there is no automated way to create a link-back system. If, when you supply your sequences to GenBank, you already have URLs for your specimens (the data is online in a database), you can provide this data to GenBank.
        i. Usually, when you upload your sequences to GenBank, you may not have your data in a public database yet. So, you'd have to go back to GenBank, on your own, and update your files. See the LinkOut feature.
        ii. If you want GenBank to link to your GBIF records, you have to provide GenBank with the GBIF link (URL) to the record there.
        iii. So it's "sort-of" simple to provide your GenBank accession numbers / links with a collections database aggregator. It's a bit more work to get links into GenBank out to these places. The community has a way to go to make all of these databases "automagically" interoperable. But, all are aware, and doing what they can.
9. How to keep track of the DwC terminology and all of the acronyms.
    a. Read the definitions here: http://rs.tdwg.org/dwc/terms/, mark the ones that make sense to you, make a goal of 10, Look at your spreadsheet, make an extra first row for mapped DwC terms, the 10 you aimed for will be easy. Now look at the remaining fields to mapped. Go one column at a time. If you run out of steam, then export just those fields, leave the rest for later. Ask a colleague about their mapping. Build your expertise as you go. You can leave the unmapped fields in your export file and map them later when you have time or more knowledge.
    b. From Deb: try looking at the higher level organization of the standards. For example, in DwC, the terms (elements, properties) are grouped (sometimes called classes) so that all terms are related to a given concept.
10. How to simplify mapping process.
    a. See above (9).
    b. In addition, make use of the automatic mapping of IPT, by using the DwC terms in your header (or as the names of your database fields). This will make the step in IPT faster. Even though the effort to map remains, it is more friendly to look at the data at the time of mapping in the spreadsheet or in the database interface.
11. How to get initial datafile set up properly to match DwC and be able to import and map easily
    a. Rename fields in the database while generating the SQL statement or the view, edit the spreadsheet with an editor.
12. What maps to where (need to practice and read lots of (i) buttons!)
    a. You can also read the definitions at http://rs.tdwg.org/dwc/terms/. (does not require clicking 😄)
13. occurrenceID vs. recordID
    a. Both occurrenceID and recordID should be Globally Unique IDentifiers (GUID)
    b. Look above (4) for different methodologies to create a GUID: it can be a UUID, a URI, a DwC triple.
    c. occurrenceID is ideally assigned by the data source.
        i. From Deb: ideally, assigned by the collector.
    d. a recordID is assigned by anyone that modifies the data from the source and passes on to others.
        i. it is primarily used for aggregators to track records, not for the original provider.

e. Thus, when an aggregator receives data from multiple sources, it is possible to easily distinguish “true duplicates” since occurrenceID will match, and the recordID can indicate that different records come from different flows of data.

Questions that require further details in the question:

14. How to make this work at my institution.
15. Mapping data to core
16. Data preparation
17. Some parts of the mapping of unmapped fields.
18. Still don’t understand mapping MY data.
19. Details of data entry.
20. Not much, want to learn more about using DOIs with citations.

--DAY 2--

--Wed Jan 14 2015 900am--

--2A Open Practical Session--

Actions:

1. Finish yesterday’s exercise
2. Work with your own dataset
3. Data quality exercise on your own datasets
4. SQL Demos (22 votes)
5. New IPT version (20 votes)
6. Open Refine

Class vote - group exercise, review dataset together, IPT pre-process and upload.

Question: Will the standard (Darwin Core) stop growing? Concern that the standard is constantly changing, making it difficult for users/data providers to keep up. Answer: Darwin Core is actually pretty stable. Core fields do not change, eg. species name, date, observer, location.

Deb is asking group if you want to work on your own data right now. Most want to work on their data and go through IPT process.

Work on datasets until 920am.

Question (Gainesville)- Has there ever been a guide of best practices as far as data formatting? For instance, in the collector field (recordedBy) should the name be recorded as First Middle Last or Last, First Middle? I suppose these would end up being local decisions, but are there costs/benefits associated with varying formats?

Question: This might have already been discussed - I couldn’t find Darwin Core terms for Township-Range-Section-Quarter. They weren’t listed under Location. Also wasn’t sure how to map Site Number and Site Name. Wondering if there were any suggestions. Thanks - Debra Miller

Note: Classroom IPT installation is being restarted. Users should reconnect. Insufficient memory.

Deb- will you please share the powerpoint you mentioned that goes through the process of calling the geolocate service through refine? Thanks

--Wed Jan 14 2015 930am--

--Demo--

Open Refine demonstration - data processing tool. Graphical interface, powerful tool for data cleaning. Facet feature helps to identify and correct duplicate records/misspelled names.
http://openrefine.org/

Refine tutorial: http://schoolofdata.org/handbook/recipes/cleaning-data-with-refine/

Refine tool was originally part of Freebase project and known as Gridworks. Freebase was acquired by Google. Data tool was re-named Google Refine. Subsequently spun off as an open source project with an active developer community - renamed Open Refine.

Other Data Tools/Sites:

Biovel supports research on biodiversity by offering computerized tools ("workflows") to process large amounts of data from cross-disciplinary source. Information resource for biodiversity data processing.
http://www.biovel.eu/

Kepler Kurator - data curation package for Kepler workflows. Kepler project (https://kepler-project.org/). Kurator information site https://sites.google.com/site/daksucd/projects/kepler-g-pack/kuration-package

Taverna data and text mining tool http://www.taverna.org.uk/introduction/taverna-in-use/data-and-text-mining/

> We use Taverna as part of our workflow/microservices that takes a scientific name entered and then on ingest populates the tree of life data - dmoses@upei.ca .

--Wed Jan 14 2015 10am--

--Audubon Core at iDigBio--

Greg Riccardi - director of Morphbank

Symbiota creates data output that have fields that do not align with Darwin Core. These fields cannot be published using IPT workflow.

Morphbank data - many datasets in one database. Morphbank database is based on DC but not identical. Additional tables in Morphbank (localities table). Higher taxonomy not in specimen table.

Mapping process designed for reuse. 4 stages to create IPT export for a dataset. Write SQL statement to select specimens in the dataset. Generate a SQL file for a particular dataset. use SQL file to create and populate a set of tables that match IPT. Map those tables with PT.

Example - Auburn University herbarium. Tables AuburnID, AuburnOcc, AuburnAC, AuburnRR.

IPT datasource. Select * from AuburnOcc.

--Wed Jan 14 2015 1020am--
--Break--

--Wed Jan 14 2015 1045am--

--Open Session--

Ottawa: What's going on? Two sessions. 1) Questions/Issues; 2) Focus on datasets.

Ottawa morning session discussion included:
Darwin core overview. Publishing, DOIs, ability to publish checklists. Aggregators, value of services (Canadensys). Open refine use for data cleaning.

Ottawa: David showed a link out in Canadensys to GenBank - demonstrating how to connect external resources.

--Wed Jan 14 2015 1125am--

--Breakout Groups--

Work your own datasets

Review submitted datasets - workshop staff have identified some issues found in submitted datasets and corrections.

--Wed Jan 14 2015 1200pm--
--Lunch--

--Wed Jan 14 2015 100pm--

--Aggregators--

Canadensys overview http://www.canadensys.net/ Canadensys makes biodiversity information freely and openly available to everyone. We are a network of researchers, collectors, curators, information technologists, students, and educators that shares data on the occurrence and identity of plant, animal, and fungal species in Canada.

Explore Canadensys Data:
http://data.canadensys.net/explorer/en/search

Canadensys Tools:
http://data.canadensys.net/tools/coordinates

Brazil Biodiversity Information Facility:
http://gbif.sibbr.gov.br/explorador/en/search

Vertnet http://www.vertnet.org/ VertNet is a NSF-funded collaborative project that makes biodiversity data free and available on the web. VertNet is a tool designed to help people discover, capture, and publish biodiversity data. It is also the core of a collaboration between hundreds of biocollections that contribute biodiversity data and work together to improve it.

VertNet offers IPT hosting, assistance with IPT configuration, data management.

Explore VertNet Data:
http://portal.vertnet.org/search

VertNet Norms - This document describes the VertNet norms for data publication and use:
http://www.vertnet.org/resources/norms.html

Question: Are all Vertnet records available in GBIF? All records from providers registered with GBIF get into GBIF.

Non-Vertebrate records from VertNet providers are not made available through VertNet, the records are uploaded to GBIF.

--Wed Jan 14 2015 140pm--

--Publishing Data to iDigBio--

Getting data into iDigBio. Set up IPT and communicate with iDigBio front end team.

Initial scope of iDigBio, all US non-Federal, specimen. Now scope is all specimen data, priority on US non-Federal, other specimen data accepted provided level of effort to incorporate is low.

Question: Registration, does each collection need to register separately? Collections can be added to an IPT set up by an institution or aggregator.

Question: Would iDigBio redirect to another provider? iDigBio would ingest and reference external provider.
Certain IPTs specialize in different types of data so for example if someone were to contact OBIS Canada but had a herbarium dataset i would assist them but suggest that they post their resource on Canadensys. I would like to recommend uploading to an OBIS node if the resource contains marine species as there are recommended best practices for marine (including freshwater - Great Lakes, etc) records that are different from terrestrial datasets. we all need to collaborate and work together. we can all access resources from other IPTs but certain groups would prefer if certain DwC fields are populated.

iDigBio search (beta) http://beta-search.idigbio.org/v2/meta/fields/records

--Wed Jan 14 2015 150pm--

--GBIF Overview--

GBIF Vision and Mission

A world in which biodiversity information is freely and universally available for science, society, and a sustainable future.

To be the foremost global resource for biodiversity information and engender smart solutions for environmental and human well being.

GBIF provides monthly updates. Metrics - records contributed, downloads, visits, citations.

Question: How does GBIF know when it is cited in publications? GBIF has team that searches and records citations when and where GBIF data is referenced.

Question: Can you search GBIF for just specimen records? Yes, use basis_of_record to filter.

Question: Where does GBIF get funding? Participant country/agency contributions, dues. Collaborative efforts can also bring external funds in.

Comment: Not a competition. Make data available in as many ways possible.

Question: Until Oct 2013 GBIF had a good logging function to track use of data (downloads/specimen data served) by data provider. These statistics at the dataset level have been lost. Stats downloaded could show for each provider or data resource at least how many queries and how many records are downloaded e.g. monthly. Now participants can request and have a report manually generated. No indication when or if the portal function for downloadable provider statistics will return.

Break out group reports. One group worked with Paul Mayer's fossil dataset. Discussion ovetr basis of record. Tricks for excel in processing data. ISO country codes. Use of literals. Fields that only contained a dash. Verbatim elevation data. Attribution, copyright. Recognizing attribution for every dataset in an aggregation is complicated.

Filemaker break out. Function in Filemaker for unique id. Vertnet has hosting capability. There is a geology extension for Darwin Core.

Bad data discovery.

--Wed Jan 14 2015 230pm--
--Break--

--Wed Jan 14 2015 300pm--

--Lightning Talks-

--Gainesville--

Something new learned:

Audubon Core, Dublin Core, EML
OccurrenceID vs RecordID

Still confused:

How an occurrenceID is generated, added to each record in dataset, where recordID is recorded

How to discover all recordIDs that are associated with an occurrenceID

Pay it forward:

Improve extension file data

Communicate the importance of occurrenceID

Create recordID directory

--

Something new learned:

OpenRefine and BioVel
Mapping is complicated - date processing, concatenating fields, definitions
Other ways (other than IPT) to generate DwC-A
Easiest to have fields in proper names (alias header)

Paying forward:

Inform other database users about concepts at institution and other smaller collections for more efficient data gathering for sharing.
Better informed to be able to communicate with database managers/programmers
Providing data to others in standardized form (rather than data dump) with metadata

Still Confused:

--Ottawa--

Something new learned: Darwin Core

Important to get any data published to start moving forward.

Still need to figure out associating media images with records.

--
Royal Botanical Gardens

Goal to get specimen collections published.

Exporting from database results in 100+ columns, 18 of which align with Darwin Core - will have to work on mapping.

Will train staff on IPT and data cleaning methods (OpenRefine) learned at workshop.

Comment (Gainesville)- identifiers and linked data papers/projects http://www.biomedcentral.com/1471-2105/15/257
http://www.biscicol.org/triplifier/
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0114069

--Wed Jan 14 2015 330pm--

--Future of IPT--

IPT Documentation

IPT Wiki
https://code.google.com/p/gbif-providertoolkit/wiki/IPT2ManualNotes?tm=6

IPT FAQ (Common Issues)
https://code.google.com/p/gbif-providertoolkit/wiki/FAQ

How to get help

IPT Mailing list and archives
http://lists.gbif.org/mailman/listinfo/ipt/
IPT interest group
http://community.gbif.org/pg/groups/3529/

Collaboration Areas Software Development

Issue and enhancement request filing
Collaborate in IPT development/coding
Collaborate in the development of associated tools (i.e., data quality tools by Canada, Brazil, Belgium…)

Collaboration Areas Deployment

Promotion
Documentation
Training
Install and use!

Translation interface
Translation documentation

Areas of Improvement

IPT 2.2 can be configured to assign DOIs to datasets, auto-generate a complete citation string for a dataset using them, and also assign a machine readable license to datasets. To promote citation, the interface has been redesigned to make it easier for users to understand how to download, use, and cite a dataset. Ultimately this will enable publishers to better track the usage of their datasets.

Improved dataset view:
- Easy navigation
- Displays information selectively
- Richer in-line help

Enhanced dataset version managment:
- All version stored and available if desired
- Major/minor version control
- Additional options prior to publishing

Dataset DOI management:
- The 2.2 version will support:
- DataCite
- EZID
- Associated to major dataset versions
- They direct to the dataset details page for that version

Easy and Consistent Citation:
- Automatically generated in a consistent format
- Metadata component automatically updated
- Supports DOI and versioning

Consistent Licensing:

- Includes the 3 license types that GBIF plans to support in the near future
- Machine readable
- Free text still possible
- Additional licenses possible for advanced users

The Future of Data Publishing - Expanding Types:

A sampling event uses a particular samplingProtocol with sampleSize and sampleSizeUnit, etc. and can record one or more taxa, each of which has a measurement (quantity and quantityType) associated with it.

The Future of Data Publishing - Dawn of DwC-A

Since GBIF pioneered the development and adoption of the Darwin Core Archive (DwC-A) as a flexible data exchange standard, the W3C (World Wide Web Consortium) has been working on a general exchange model for data represented as comma-separated values (CSV). The requirements of the Darwin Core Archive have been submitted to this working group and accepted as a use case. The outcomes of this work are likely to become widely-adopted by many research domains and applications and should be considered as a long-term replacement for DwC-A. The GBIF Secretariat will coordinate review of the CSV recommendations and develop a strategy for future adoption and use in GBIF software, including GBIF.org and the IPT release 3.0.

--Participant List--

Ottawa Participants : Tuesday January 13, 2015

Heather Cole	AAFC
Anissa Lybaert	AAFC
Iyad Kandalaft	AAFC
Satpal Bilkhu	AAFC
Allan Jones	AAFC
Joel Sachs	AAFC
James Macklin	AAFC
Wen Chen	AAFC
Tyler Smith	AAFC
Christian Gendreau	Canadensys
David Shorthouse	Canadensys
Kyle Martins	McGill
Glen Newton	AAFC
Diana Sawatzky	U of Manitoba
Richard DI	AAFC
Donald Moses	UPEI
Mathieu Ouellet	MPO
Jenny Chiu	DFO/MPO

Cobus Visagie	AAFC
Shannon Asencio	Cdn. Museum of Nature
Amanda Ward	AAFC
Gisele Mitrow	AAFC
Jenn McPhee	RBG
Tammy Elliott	McGill
Nadia Cavallin	RBG
Kelly Sendall	Royal BC Museum
James Smith	Royal BC Museum
Bryan Brunet	University of Alberta
Carolyn Babcock	AAFC
Dicky Yu	AAFC
Stephen Darbyshire	AAFC
Mark St. John	msjsci.com
Val Tait	CMN
Jennifer Wilkinson	AAFC

GAINESVILLE - 14 Jan - Unconference options

Finish yesterday's exercise of publishing a DwC-A √	4 people
Work with your own datasets √	
Data quality exercise with some of your datasets	8 people
SQL demos (x2) Greg √ & Laura	More than half of the room
New IPT version (v 2.2)	6 people
Open Refine √	

√ = DONE!

OpenRefine Notes

-- Download --
http://openrefine.org/download.html

-- Documentation on GREL - Google refine expression language --
https://github.com/OpenRefine/OpenRefine/wiki/Google-refine-expression-language

-- Find duplicates --
On a column header: Facet -> Customized facets -> Duplicates facet

-- Concatenate 2 columns --
On a column header -> Edit column -> Add column based on this column
Expression:
value + " " + cells["My other column header"].value

-- Get data from another 'project' --
On a column header with the id referencing another project -> Edit
column -> Add column based on this column
cell.cross("other project", "id")[0].cells["column to get data from"].value

-- Transpose columns into rows --
On the first column header -> Transpose -> Transpose cells across into rows

Select 'Transpose into':Two new columns
Check ' Fill down in other columns'