Vtc notes 2019-09-03 non tabular data wg

This document is in google:
https://docs.google.com/document/d/1im1Ufu3x3D5sf2h_BJ3j8KeAACCuaNkCp4uDXvo0V-k/edit

Attending:
Stace, Gastil, Sarah, Mary, Hap, Corinna, Li, Greg, Mark, Kris, Mariel Campbell, Tim, An

Below is an attempt to summarize the information into different working groups with a few issues listed that need to be resolved. Obviously, every working group needs to identify and define the issues in more detail. I have tried to classify the data in the spreadsheet with the color coding here. The division between groups 2 and 3 are not entirely clear to me and I may have color coded them wrong.

1. Models and their input and output (red)
    a. input data are too big to archive
    b. output data are too big to archive
        i. archive only output from certain runs (published, useful to others?)
        ii. archive only summary data used in publication
    c. archiving the model itself
        i. is on a community maintained website/git repo
        ii. does the author agree with publishing
        iii. if model is archived, how to deal with environment setting to make it run again
    d. what metadata are needed
2. Species, specimen, image, audio, genetic sequence data (green)
    a. this would be everything that fits into the natural history collections concept
    b. most of these data would reside in a different repository and then be linked through GUIDs to environmental data in EDI.
    c. summary data from analyses would be in EDI
    d. EDI metadata would provide the link to the other repository where the raw data could be found.
3. remote sensing in Lidar, drone, (under water photoquads, zooscan?) (yellow)
    a. this would be everything that uses multi-spectral imagery or motion analysis to measure a proxy for species/communities/landuse, i.e. somewhat larger units than single species/individual counts
    b. most of the raw images would be in an offline storage/other repository
    c. what metadata/data should be accessible in EDI

    d. should there be any policies on how long to keep raw binary data files that need proprietary software for processing/viewing/converting to other formats
    e. Jane Wyngaard jwyngaar@nd.edu. is the drone specialist at ESIP
4. Data that should go into EDI only if additional functionality is developed
    a. If EDI were to provide such additional functionality, e.g., pre-view shape files, rasters, access netCDF files natively, etc., what is the prioritized wish list.

Resources :

https://sites.psu.edu/dcnworkshops/primers/?utm_source=newsletter&utm_medium=email&utm_campaign=dataone_webinar_series_data_curation_primers_expanding_the_community_curation_toolkit&utm_term=2019-09-03

Looking at spreadsheet, …

MCR GoPro - a lot of empty video time between observations that get further analyzed.
Photoquads have a high level of indexed data about both the analysis process and product.
NOAA northwest fisheries (suggested by Stace for … I forgot what).
Greg mauer has 3D image data from drones at JRN.
Hap PIE is pursuing structure by motion to get growth of marsh plants. Beginning phases.

Corinna begins to categorize the spreadsheet with colors.
Blue means large raw data likely only local?

Tim mentions putting metadata in pasta but data elsewhere.

Stace at NES asks if investigators are hosting their own servers, then meta in pasta, pointing to local server.

Sarah Elmendorf mentioned large datasets stored offline (external hard disks).
Sneakernet to send data to users?

Stace suggesting data-not-in-pasta pkgs.
Threadserver - handles model output. 3D ocean circ and coupled biogeochem models 3D+time.

NES has data going into R2R of course.

Green for data going to repos that have doi's.

Specialized data w/ proprietary sw. To use data from raw form to processed.

NES Attune NxT Flow Cytometer - look at spectra?

NES ZooScan - scans a whole sheet and sw pulls out individuals. Ultimate data products are counts of species and morphometry of organisms. EcoTaxa.

KBS - large spatial raster

JRN - campaign data. Flown every 6 months at several sites. Lots of imagery. 10 Gb per campaign per site. Used to create 3D point clouds using structure for motion method. Georeferenced, stitched into mosaics. 3D models of vegetation.
One research group is trying to put the drone data into NetCDF.
3D point clouds come out as LAZ file
Metadata inside these files.
Ecosystem hydrological modeling. More complicated than the usual dataset. Parameter filesets fed into models, diff versions of models, diff configurations. Model outputs. Point-based models produce time series simulations. Spatially distributed models produce a raster.
Parameter sets and outputs can go into EDI. Assuming a model binary could go in as an otherEntity?

Corinna - what are these models?
Greg - SOILWAT & ECOTONE, … diff researchers have diff versions of it.
Corinna - are those model versions maintained somewhere, accessibly?
Greg - depends. Some are very version-specific. At least one of these models is definitely not on a public repo. Email modeler to get it.
Important to keep model and parameter set together in a repo for reproducible results.

Greg - GIS data not as reusable from EDI. Can go in, and JRN is putting them in.

Corinna - does HydroShare have such tools for GIS data?

Tim - HS has capability to preview NetCDF or other formats. More visualizations can be added. Users add tools. Hook data to tool.

Hap at PIE
LiDAR. 60 to 100 Gb each. So far PIE has put the resulting DEMs into pasta.
Depending which company, the raw data files, flightlines, air-ground last-return data… all on external hard drives. Mailed to distribute. Not sure how to pkg if EDI could handle.

Other LTER sites have lidar data. Could work together. NEON is publishing the raw files. Obtuse to find, so much of it. Discoverability.

Corinna - is there a place to register all lidar data ever taken?

Hap - historical (paper) ie diaries of colonial era, such as “haying out the marsh” land use data. Text documents. Library of information. Would be nice to have it discoverable.

Corinna - local historical society?

Hap - a list of historical documents at 20 places. History of the region.

Kronos sequences of historical info over decades. Images into context.

Mark Servilla - Hap could have local website with pointer to historical documents. Pasta … capability.

Hap - is purpose at EDI for discoverability no matter where stored.

Corinna - people would go to PIE looking for this kind of data.

Mark - do want to improve discoverability across other types of data. Some outliers of types of data. Multiple points of discoverability.

Hap - historic and current imagery. USGS T-sheets back to 1850’s with little metadata, like bad xerox copies. Boundary info. Not georeferenced. Need to scan (digitize) to apply that.
Image data.

Corinna - zip of image files, lat, lon and other information. Could have one entity per image.
Hap - directory structure by dates. Zip to preserve structure.

Hap - drone data not organized. Flying hither and nither. Will be trying “structure from motion” to classify marsh plants. A work in progress. Species based on structure and form of plant seen. Spartina alternifolia, Spartinia patens (hay-like), intermediate forms.

Corinna - raw data, analyzed results, csv output of …

Hap - stitched image and its interpretation. Some spectral data.

Greg - output from structure from motion is a 3-D point cloud like LiDAR data.

Hap - image embedded in data table. Videos of sedimentation in estuary. Landsat, Sentinel images. Avi movie showing sedimentation over time. Put into an Excel template.

There is also associated (tabular) data. 80 images. Tiff images. Zip files of images and video going into otherEntity. References to csv file for data results.

Stace at NES
I am exploring how NOAA handles Lidar data: "NCEI is the long term archive for coastal lidar data distributed publicly through the NOAA Office for Coastal Management's Digital Coast"
https://www.ngdc.noaa.gov/mgg/bathymetry/lidar.html

Genomics data
NCBII accession numbers, project numbers, created a Microsoft Excel descriptive file:
Paper where published, authors, accession numbers, dates, and environmental parameter data such as temperature, salinity, marsh habitat.
Excel descriptive file going in as otherEntity into EDI. (The spreadsheet is metadata.)
About 30 papers. Users can get sequence data from genbank. 10 or 12 columns.
How does EDI want to discover genomic data?
Our data is their metadata.

Mary at HBR
Lidar and hyperspectral drone. Audio bird studies about a Tb/year. Camera traps: Large quantities of jpg files. GIS data: shapefiles going into otherEntities.
Model macrosystems data. Inputs are larger than outputs. Not practical to host the inputs and code. Outputs sizeable.

Corinna - are there databases out there for audio files and camera trap data, counting species. Analysis done on the raw files. Output products are much smaller and tightly coupled to publications. But also want to archive the raw data. Future data users.

Mariel
Is this raw or spatial?

Mary
Metadata and analysis for every .wav file. A table of what was (found?) in that file.

Mariel
We can take that into the collections database. Observational data. Audio and video can be stored and linked to taxa found in them.

Stace:
"World's largest natural sound archive now online"
https://news.cornell.edu/stories/2013/01/worlds-largest-natural-sound-archive-now-online

Mariel
Arctos at __Texas computing center

The Cornell Lab.

Corinna - most of the natural history museums can do this.
iDiggBio

Mariel
iDiggBio is not a primary repository. They are an aggregator. They take data from primary repos. Most of the media is at the primary repo.
Arctos is a multi-institution repo. Would need an agreement to host thru that platform.

Corinna -
To Mary at HBR - ask your university if they have a natural history collection.
One of the associated universities.

NWT Sarah
Drone data. Same sorts of non-tabular data as other sites have listed.
Genomic, GIS,... Want our primary search at EDI.

Corinna - going towards using EDI as a pointer, a search node.

A best practice for genomic data, for collections of images (of varied type) including drone, lidar, species-counting, population community analysis images.

A best practice for model data. Not clear. Too many different types of models?

SEV
Voucher specimens and their data. Link thru EDI. Data records have each a url which can be linked?

NWT = Sarah
Phenocam - maybe National Phenological network? Phenocam Network. They may have apps.
Green-ness tabular, raw on local server?
Future data user might want raw files. Re-process. Off-label uses.

SBC
All types of data already discussed.

Tim at BLE
Frozen physical specimens? Linking.

Corinna

Break into sub-groups.
Best Practices:

- Voucher specimens - no takers
-

Tim - helpful to distill the different groups.
From BLE's perspective, our sci will return from arctic soon.

Hap - intent is … if data types can go into EDI,... do that. Discoverable via D1 to edi.
If too-large data and exist other repo, what is the process for discoverability, via D1?

Corinna - yes
My vision is if there exists a repo, such as animal audio files, then probably that would be a place people would look for them so appropriate there. If data files associated, sucha s env. Data, that could go in EDI. Then, how to link these to other repo. "The audio files are somewhere else…" 2 ways to discover.

Mark S
DataONE only if other repo a Member Node.

Corinna - link to other repo can go into metadata.

Mark S
There is not yet a Best Practice how to point to another repo from EDI.
Interoperable relationships between repos.

Corinna will summarize and ask which WGs each of us want to work on the BPs.