Non tabular wg notes 2019-08-20

Attendees:
M. Gastil-Buhl
Tim Whiteaker
John Porter
Corinna Gries
Greg Maurer
Stace Beaulieu
Jason Downing
Li Kui
Sarah Elmendorf
Mariel Campbell, UNM Museum of Southwestern Biology
Teresa Mayfield, UNM Museum of Southwestern Biology
Renée Brown
Kris Hall
Mark Servilla
Chris Turner
Hsun-Yi Hsieh
Mary Martin

Background:

Discussions at 2019 Science Council & IM/Esip Meetings re: non-tabular data
How to link up
Specimen data (ie in museum) plus environ where collected; how to link

Categories for data:

- Type - images, video, audio, model output, spatial, genomic data
- Approximate size annually?
- Active accessible storage - not directly accessible
- Other existing technology for access - netCDF, opendap, ERDDAP, WFS, WMS, Globus+box?
- Other existing repositories
- Indexing - area, time, species, behavior, trait etc.
- How should they be retrievable
- Other storage place and linking is necessary
- Existing metadata
- Raw data vs. publishable data
- Best practice vs. resolving storage and technology

==========

Linking part not hard. Best practice.

Seeing new kinds of data.

Mark Servilla

Pasta does not do as many quality checks when data not tabular.

Stace

At NES data high volume often stored locally. Best practice of interest is how to provide identifiers so locally stored data can be found thru repo.

Types of high volume data (see google doc)

Plankton imagery. ISIIS instrument. 8 to 16 Tb from one instrument, one cruise.

ID related to a specimen.

Ocean model data. Rel to Chris Turner too. Requested 12Tb and already full.

Acoustic data - now handled thru other repos. Need to keep good metadata.

Gastil

Also metabarcoding data (forgot about that). But not just the data itself - we can do that as a huge .zip - but then we lose all the relationships between the analysis and the raw data - photoquadrats (indexed) - currently only catalog results. Underwater videography with timestamped observations - only analysis cataloged. Volume high enough we don't even try to store it here…

Like idea of cataloging them by the challenges they represent. We have web page for browsing photo quadrats - then they can request 1 by one (not the modern repository practice).

Mariel Campbell

(audio unclear)

Museum Consortium online database. Pipeline… genomic data…

Talking to Mark Servilla and Kris Hall (SEV) … evolutionary modeling animals, plants

At UNM

Integrating data streams. Same LTER site and same types of data.

Collection manager for genomic resources.

Archive to link to environmental data.

Issues:

Really high volume data

Link to other data in other places

John Porter

Both large volume and diff to index: drone data

Video or large numbers of images.

Some things where databases will acetp: ie generic data w/o env data

Some types of data so specialized that there is not already a db that wants them.

(maybe gopro that category)

Mark Servilla

What is expectation for lower limit of “big” data? 100 Gb? 10 Tb? Higher?

Currently purchasing infrastructure at UNM.

JP

Storing easier than retrieving by timestamp or something.

Gridded-ocean remote sensing. OpenDap & such.

Tim W at BLE

BLE will have 24 automated cameras and video. 35 Tb annually. Plan is to archive to tape backups at the university, and archive analyses of the media at EDI.

.

Stace

16 Tb in first year (mainly due to model data and plankton imaging data)… Initially estimated 100 TB for 5 years, but expecting more….

Expect on order what Tim said but maybe more because model data.

Jason D at BNZ

Still photos timeseries archives. Recently began drone.

Photo collection not fit pasta because how to describe for search.

4 to 8 Tb or less.

Li K at SBC

Photoquad data. Scale 1 Gb per survey. Have done about (ongoing, annual) 20 surveys. 20 Gb each year at least next 2 years.

Kris H at SEV

On smaller end at this point. Not a lot. Just getting understanding this type data. Sound files, drone imagery in future.

Chris T at

Large volume. Imagery. Modeling. Also data outside of LTER project.

100s of Gb acoustic data per year (non LTER)

Hsun-Yi H at KBS

Collecting drone images. Have a few from last year. Not more than 500 Gb/year if continue at this pace.

Have non-LTEr data we collect every 2 weeks. Long-term (just not lter).

Drone data new. Still figuring out how to store and present to users.

Mariel Campbell

130 data collections. Close to 4 million records. At Texas Adv Computing center. (TAC)

Have capacity to add more. 100$/year/Tb.

Mary Martin at HBR

Recently. Traditionally 10s of Mb. Not ramping up; giant leap. Now Tb.
One dataset is a model output on order of couple Tb. Input data & code to generate.
Input is 1 or 2 orders of magnitude larger and requires supercomputer center to re-run.

Audio data (ie birds). 10 Tb/year. Do not know if people would want raw data in future. Seems should keep that. Analytical methods change. May analyze for different things (birds). Want to preserve. Can put summary data online quickly.

Renée at MCM

Starting to generate lots of images. Flights over the Dry Valleys. Landscape albedo.
Better way to link photos to data than just a giant zip file.
Also big project with a lot of dives. GoPro videos and still images.
Cameras. Similar to time-lapse.
Genomic data (currently housed at NCBI, others).
Models, model outputs.
Remote sensing, survey, and LiDAR data.
Size: not known.
Asked ability of EDI to store a 4Tb dataset.

Sarah E at NWT

Drone data from a post-doc. Not planned-for.
Time-lapse photography.
Genomic data expect will go to another repo. (Link to edi?)
Templates for metadata to go with drones. Don't know if getting info needed to re-use.
What is adequate metadata for drone data?

Mark Servilla

Yes maybe near a petabyte.
Volume of data exceeding current capacity.
Pasta designed around traditional data.
"This is great."
Have ideas how to scale our arch to support this data volume.
Next proposal.
Input appreciated.

John Porter

These new data demand change in paradigm.
Upload dataset; users want to download part they want.

Mark S

For those who have one-off datasets to archive, please contact us (EDI) because direct uploads not expected to work as for normal-sized datasets.

EOS system example, for remote sensing data. System for dealing with image time series. Could build upon those patterns.

Stace

Satellite data providers. For gridded data. NASA/NOAA choosing cloud storage / services.

Possible solution for serving georef video data, still matches old download data Like seafloor cruises. How to find snippet of drone data.

Seafloor data catalog may be applicable. NOAA solution good. I think this is the link to discover NOAA seafloor video, which I think may be relevant to metadata for discovering drone video: https://www.nodc.noaa.gov/oer/video/

Mark S

Sneaker-net still works. Option for tfr data.

Historic way pasta looks at data: publication-quality.

Pasta is not a dropbox or google drive place to “plop” data.

John P

Image data, esp time-referenced, diff to know what someone will want in future. Capture in richest form. Raw data.

How unique is (this kind of data) to us? Natural resources companies.

Time Series monitoring we may be main place doing this.

Corinna

Trying to capture a matrix to build.

1st page of (these) notes. ?not seeing it.

Type

Image / video / audio / model output

Size

Annually?

Thing like a drop-box. Big storage in bkgd. Apply for data & later receive it. NEON sends around hard drives.

Raw vs publish-able data. Classify.

Some data that already have good access mechanisms, such as NetCDF. Does it make sense to use EML or something else (for those)?

Mark S

Interoperability between repositories or database or dropbox. How to integrate across the board. Groups have tried w3ith varying success. How to link all these things together to take advantage of provided services.

Corinna

Indexing. What to link to, what gets an ID? Space and time.

Link to things stored elsewhere.
Some data streams that come with their own metadata streams, ie off the instrument.

John

What to index of time, place, taxa, behavior. Generalized model.
Granularity. What do you retrieve? Frames, videos, clips?

Scope of what can be done now and future.

Next steps:

What we can do for each type:

Best Practices
Store for future
… (I lost connection)
Find tech existing that could be hosted or re-purposed
Other existing repos (for specialized types of data)

Sarah

Become aware of other repos. Not re-invent. Create best practices.
Drone data nothing good for yet.

Stace

Has tried categorizing as Corinna suggests. Good idea.

…

Corinna

List data by category
Look for appropriate repos (where appropriate)
Write Best Practices

Sarah

Concern not knowing what metadata to ask from data collectors

Sept 3 Tuesday noon pacific next zoom.