VWC notes April 8, 2019

Attending:
Participants: Suzanne Remillard (AND), Don Henshaw (AND), Wade Sheldon (GCE), Margaret O'Brien (SBC/EDI), Tim Whiteaker (BLE), An Nguyen (BLE), Stevan Earl (CAP), Ed Johnson (CWT), Duane Costa (EDI), Kristin Vanderbilt (FCE/EDI), Darren James (JRN), Sven Bohm (KBS), Yang Xia (KNZ), Renée F. Brown (MCM), Elizabeth Dobbins (NGA), Chris Turner (NGA), Corinna Gries (NTL/EDI), Sarah Elmendorf (NWT), John Porter (VCR),James Conner (PAL/CCE), Adam Kennedy (AND), Lindsey Rustad (HBR, USFS), Emery Boose (HFR), Julien Brun (NCO), Dave Rugg (USFS), M. Gastil-Buhl (MCR) (notes)

Suzanne R - imexec. Due April 26 … started on g drive topics for esip meeting vs imc meeting.
NextGen ClimHydro may be a topic.
7 or 8 primarily ESIP and as many that could be IMC.

Suzanne R - intro
EDI organized a wkshp to discuss ClimHydroDB's future. IMs, EDI staff, scientists from USFS and LTER.

Don H - overview
Powerpoint VTC_8April2019.pptx

Suzanne
Wanted to evaluate needs to harmonize multi-site, multi-agency data. Considered infrastructure and data models. (see more on ppt)

Don
Silde 2 - current status: old tech, difficult to maintain, needs to be retired. Will archive current contents into EDI/pasta.
What standard framework good for clim and hyro data?
Considered CUAHSI.
Begin to prep clim/hydroDB data using the CUASHI ODM.
This to eval CUAHSI tools.
ODM to harmonize high-temporal and met data forward.
Will write a proposal for RFC to the IMC, EDI and others. Available at the summer meeting.

What ClimDB/HydroDB is.
Devel late 90's. Started at NTL, then AND, then LNO. Allowed each site to maintain their own data, submitted in a common exchange format csv file.
Primarily precip, air temp, stream discharge. Stdized atts and units.
Harvester you could trigger. Post data where harvester could reach.
Wade developed way to pull in USGS stream data and NWS data.
The DB has an interface to display (see ppt)

Some web services that no longer function.
See slide 5 for tally of participants and stations. >10 million daily values.
All daily data.

Yang currently manages ClimDB.

To archive, take as-is and put in pasta. Use GCE toolbox and AND tools to make the archive.
An L0 structure, one pkg per site. Multiple tables, one per station. Wide format (cols are measurements).
Generic EML abstract, methods. Sites would have opportunity to update data in clim hydro before this archive happens. Deadline TBD for submissions the shutdown.

Then put the data into CUASHI.
CUAHSI is a data values table, with sites and meas described in parent tables.
Mapped the vars to CUAHSI vars.
Also to put the ODM structure into EDI.

Slide 8 future
Use high temporal resolution data already in EDI.
Use ODM as framework to store common version of those data.
Harmonize params and units.
Once into ODM, then aggregate to hourly.

Lindsey -
Grace period, number of months, when sites could update their data in Clim/Hydro. Years in arrears.
Deadline likely after the summer.

Synthesis overview paper (publication) broad spectrum, high level synthesis, as motivation (a carrot) to encourage site involvement.

Wade
Different ppt file (blue bottom border on first slide)

Description of CUAHSI. Cyber infrastructure for sharing point-based measurements of variables over time.
A time-series is site + meas + method + source/contributor + quality level (raw/analyzed) This is better on the slide.

ODM Observations Data Model
Currently vers 1.1. A new version ODM2 around for years but not yet incorporated, adds GIS.

ERD star topology

Contribute via a csv file
Assemble 6 data tables, 5 are look-up tables that can be reused.
Data values table.
Defined contribution format: worksheet named DataValues.
The template is an Excel file.

ODM support was already built for GCE Toolbox for another project years ago with DataTurbine.

… see slides

For archiving the old/existing content, could get more specific methods from each site, or just use what is already stored in clim/hydro.

Slide: GCE Toolbox for L1 (ODM)
Will try a test one-time conversion of data using Toolbox, then generate L1 EML to describe a wide table for each met station.
Then use mapping tables to generate the ODM1.1 structure “skinny format” long format data table and its 5 lookup tables.
That would go into EDI as a L1 table, appropriate to contribute to CUAHSI.

Training for sites to do this on their own, or EDI could help sites use our (GCE’s?) workflow. Multiple pathways.

Refactoring ClimDB to ODM.

What should go in the “derived” metadata, a topic for discussion.

Margaret

Powerpoint with grey stripe background at top.

Level zero is any raw data. Ideally data is converted to a predefined model. (Example, the ecology community survey data harmonization model.)
Level one is a step where data has original granularity (not aggregated) but format is harmonized.
Level two can be multiple data products made with code.

ODM is good for Level one.

Step two CUASHI has done for us.

Phase 1: EDI Archives ClimDB/HydroDB Content
(slide)

Phase 2: Ongoing ODM-style Datasets (L1)
(slide)

Future Discussions
Different kinds of data, ad-hoc format.

There is a session at ESIP, a code session proposed by EDI and ESIP?

----- end of pwpt.

Discussion:

Lindsey
Tasked with spending a day in the CUASHI tools. User friendly. The tools are clunky as-is to download the data, a limit how many rows. Graphing is rudimentary.
Will ask CUAHSI their future direction.
Did have a library of R and phytho scripts.
Not quite plug-and-play.

Stevan
What do I (as a site IM) need to do? Just wait til these issues are resolved?
Suzanne
Yes can wait til WG finishes testing and offers the RFC.
Meanwhile can add to current ClimDB since it will get archived.
Wade
Think what your perspective (as an LTER site) what metadata would enhance re-usability of your data.
Tim
What format will discussion continue?
Mob
No slack nor git yet. Need to settle on the name. HydroMet? Not ClimDB2.0.
Lindsey
HydroNet name amenable to other stream data streams such as chemistry. Could be broad in name.
Mob
Profile data, concentration measurements. Beyond hydrology and met.
Suzanne
Not everyone has access to slack. Email someone in the WG.
Mob
There is already an EDI slack channel for data harmonization.
Wade
Wind-up to get test products is short so could generate a couple of candiate L1 and L0's for folks to hammer on. A practical thing to ask questions about.
Corinna
There is still an issue of every LTER site contributing even if they are getting data from a different (non-LTER) group. In CUAHSI you can define what the source is. Users commented that important to get all met data from one place. Not a definitive answer yet. Or, how many weather stations is a site going to contribute to this.
Don/Corinna
CZO will use ODM also.
Lindsey
Cannot brand as an LTER vs CZO site. GLEON was not clearly identified. In metadata be clear about the brand where the data are coming from. Branding is an issue for the forest service.
Corinna
Could name all the sites prefixed by LTER. But what about USFS sites.
Wade
Organizational scope is important to search/discover and group LTER data.
Suzanne
Could we go from our LTER site directly into CUAHSI and skip uploading the high-resolution data?
Because I question the usefulness of the hi-res data in pasta if not useable.

Mob
Long-format minimizes empty space. But because key-value pair, limited to the type that can be assigned, and needed info is stored in a separate table. Not always a place to put all the metadata that would go into an EML pkg in pastat.
Corinna
Historically, sites were not willing to reformat their own data into an L1 format.
For other types of data there is an important reason for the L0 format.
OK to only contribute L1 directly.
Mob
Maybe we need guidelines as to whether to contribute L0 or go straight to L1.