Annual LTER Information Management Committee Report (2019-2020)

Date: April 20, 2020

Name of Committee: Information Management Committee (IMC)

Names and affiliations of the Committee co-chairs: Suzanne Remillard (AND), Stevan Earl (CAP)

Current IMC members: Dan Bahauddin (CDR), Stace Beaulieu (NES), Sven Bohm (KBS), Emery Boose (HFR), Renée Brown (MCM), James Conners (CCE, PAL), Jason Downing (BNZ), Stevan Earl (CAP), Sarah Elmendorf (NWT), Mark Gahler (NTL), Hap Garrit (PIE), Gastil Gastil-Buhl (MCR), Corinna Gries (EDI/NTL), Kris Hall (SEV), Don Henshaw (AND; retired 2020-01-03), Jamie Hollingsworth (BNZ), Hsun-yi Hsieh (KBS), Darren James (JRN), Li Kui (SBC), Jim Laundre (ARC), Miguel Leon (LUQ), Mary Martin (HBR), Greg Maurer (JRN), An Nguyen (BLE), Margaret O'Brien (EDI/SBC), John Porter (VCR), Suzanne Remillard (AND), Adam Sapp (GCE), Chris Turner (NGA), Kristin Vanderbilt (FCE/EDI), Tim Whiteaker (BLE), Yang Xia (KNZ).

How membership is determined: Membership and roles are outlined in the LTER IMC Bylaws (v.3 ratified 2018-04-27). Membership includes an Information Manager (IM) from each site that serves as the primary site IM contact. Additional members may be identified by individual sites, by the LTER Network Office (LNO), and by Environmental Data Initiative (EDI) from among site or project personnel who are involved with information management. The Information Management Executive Committee (IM-Exec) is the steering committee for the IMC. IM-Exec members, including the IM-Exec Chair or co-chairs and a representative to the LTER Network Executive Board (EB), are nominated and ratified by the IMC committee as outlined in the IMC Bylaws.

Current Information Management Executive Committee (IM-Exec) members (LTER site, term end-date, and role (if relevant) in parentheses): Brown (MCM, 2021), Downing (BNZ, 2020), Earl (CAP, 2021, co-chair), Remillard (AND, 2021, co-chair), and Whiteaker (BLE, 2021). Ex officio members include Bahauddin (CDR, 2021, EB Representative), Marty Downs (LNO), and Vanderbilt (EDI/FCE) who participate in and provide reports during the monthly IM-Exec meetings.

Meeting frequency:

- The IMC holds an annual in-person meeting typically in conjunction with a related meeting (e.g., LTER All Scientists' Meetings (ASM), Earth Science Information Partners (ESIP) Meeting, or Ecological Society of America (ESA) Meeting). All LTER IMs are strongly encouraged to attend the annual IMC meeting.
- In addition to the annual in-person meeting, the IMC meets monthly through Virtual Water Cooler (VWC) meetings, which are video conferences to discuss key Network topics. IMs from all LTER sites are encouraged strongly to participate in these and the meetings are typically well-attended.

- IM-Exec meets monthly via video conference to plan events, discuss key Network topics, and coordinate cross-site activities.

Major activities or accomplishments for the year:

- The IMC addressed a wide range of topics during their monthly VWCs. Topics of note during this past year included: preparing for the LTER Decadal Review; identifying best practices for reporting data citations and maintaining bibliographic information; annual meeting planning; tools and approaches to managing geospatial data; managing the DataBits publication; incorporating new features afforded by an update to the Ecological Metadata Language (EML 2.2); and presentations by representatives from LTER Synthesis Working Groups. This year, a new venue was established called the Information Management Knowledge Exchange (IMKE), in which a subset of VWCs focus specifically on exchanging practical knowledge among the IMC regarding information management tools, techniques, and best practices. Agendas and notes from all meetings are posted to the IMC VWC repository on GitHub (https://github.com/lter/IMC_VWC).
- The 2019 annual LTER IMC meeting was held on July 15, 2019 in Tacoma, WA in association with the 2019 ESIP Summer Meeting. Meeting materials and reports are available through the LNO document archive (https://lternet.edu/wp-content/uploads/2020/04/2019IMCAnnualMeetingReport.pdf). Highlights and products of the meeting include:
    - IMC members, EDI representatives, and other guests gathered to discuss issues related to information management.
    - The meeting centered on two breakout sessions that featured small groups working through guided discussions about engagement and collaboration within the Network. A morning breakout session focused on approaches to stimulate more efficient and productive engagement between site Information Managers and scientists. The morning discussion identified educational- and marketing-styled approaches to engage with site investigators and to promote the value of publishing data. An afternoon breakout session focused on identifying new tools and platforms for promoting collaboration, and documenting and archiving IMC resources. Outcomes for the engagement discussion identified ways to let PIs know what the IM can do for them and their research can benefit. Outcomes for the collaboration discussion identified an approach to work with the LNO and a planned community platform to promote collaboration with and among the IMC, and to distribute and archive materials through cloud services (Google Drive, GitHub).
    - DataBits editors elected: Bohm (KBS; editor) and Nguyen (BLE; co-editor).
    - Met with NSF Program Officer Peter McCartney, also attending the 2019 ESIP Summer Meeting. McCartney discussed his visions of and for information management in the LTER and ecological sciences broadly. On the topic of data repositories (e.g., EDI, which provides critical infrastructure and resources for managing and archiving LTER data), he advised that to preserve repositories, researchers need to use and cite

them. He also emphasized that LTER IMs should be involved in data synthesis, analysis, and visualization efforts to further promote and make use of centralized resources like EDI and DataONE.
    - IMC members organized six sessions at the ESIP meeting, including:
        - *A Metabase Database Built on Usage Patterns in the LTER Network*, organized by Whiteaker (BLE), O'Brien (EDI), and Gastil-Buhl (MCR).
        - *EnviroSensing: Sensor Data, Technology, and Best Practices*, organized by Brown (MCM) and Strachan (Univ. of Nevada Reno).
        - *The Information Management Code Registry: Software Solutions for Information Management Needs*, organized by Smith (EDI) and Vanderbilt (FCE/EDI).
        - *Getting Stuff Done with R, Python, and Jupyter Notebooks*, organized by Gastil-Buhl (MCR), Beaulieu (NES), Porter (VCR), Smith (EDI), and Turner (NGA).
        - *Location, Location, Location: Enabling Data Discovery by Place*, organized by Porter (VCR) and Vanderbilt (FCE/EDI).
        - *Preparing climate and hydrological time series data for submission to CUAHSI*, organized by Gries (EDI/NTL), Seul (CUAHSI), Sapp (GCE), O'Brien (EDI/SBC), Remillard (AND), and Henshaw (AND).

- The ClimDB/HydroDB 2.0 database project, also known as the Next Generation ClimDB (https://github.com/lter/Clim-HydroDB-2.0), has a formal recommendation from the ClimDB/HydroDB working group with input from the LTER IMC to implement the next phase of ClimDB/HydroDB. ClimDB/HydroDB 2.0 centers on harmonizing climate data from all LTER sites and some USFS sites into the CUAHSI Observations Data Model (ODM) format, and archiving that output in the EDI data repository. Additionally, the old ClimDB/HydroDB database that includes data from LTER, USFS, TERN, and a few other sites will be archived in EDI and CUAHSI. The ClimDB/HydroDB working group is preparing a final recommendation to share with the LTER Executive Board.
- Published the summer 2019 issue of *DataBits, The Newsletter of Ecological Information Management* (https://lternet.edu/wp-content/uploads/2019/06/DataBits_Summer2019.pdf).

Subcommittees or working groups (WGs):

- *IMC Website Improvement & REDesign (WIRED).* This WG was tasked with updating the IMC website (formerly http://im.lternet.edu). This year, as the website is being retired since upgrading the underlying technology is not feasible, the site content was migrated to various GitHub repositories within the LTER and EDI GitHub organization. The group plans to move some content under the new LTER website once the site platform is established.
- *EML Congruency Checker (ECC).* This EDI-led WG is tasked with implementing new or revised data-quality checks for the PASTA+ platform used by the LTER and other groups to submit data to the EDI data repository. The group implemented several new data checks this year that focus generally on ensuring

that the metadata concerning the size of data objects is consistent with the actual size of data objects that are submitted to the EDI data portal.

- *Drupal Ecological Information Management System (DEIMS).* This WG consists of LTER sites that employ the Drupal-based DEIMS tool as the foundation for their information management system. This year, they established a GitHub repository with code and documentation on migrating from Drupal 7 to Drupal 8 (https://github.com/lter/Deims7-8-Migration).
- *Semantics.* This EDI-led WG was established to address how EDI may improve data discoverability by adopting appropriate vocabularies, ontologies, or other registries of resources that have permanent identifiers. With the release of EML 2.2 and its support of semantic annotation, the group developed a semantic annotation primer that is included in the EML 2.2 documentation website.
- *ClimDB/HydroDB 2.0.* This WG is tasked with identifying long-term solutions for standardizing and archiving climate and hydrologic data from LTER, and other research sites and organizations. The group has developed a plan involving migration toward a format that draws from the CUAHSI ODM, and is preparing a proposed plan that will be brought before the LTER executive board.
- *Information Management Knowledge Exchange (IMKE).* This WG seeks to share knowledge about LTER information management tools, approaches, and best practices. The knowledge sharing occurs during select monthly VWCs. An example of the success of this format is one where a discussion of dataset attribution informed and inspired EDI to make improvements to how citations are generated in their data portal, resulting in their new Cite web service (https://environmentaldatainitiative.org/2020/02/12/cite-a-lightweight-citation-service-for-data-packages-in-the-edi-data-repository/).
- *Core Metabase.* This WG is tasked with developing LTER-core-metabase, a database schema and accompanying tools that can be implemented as the backbone (or enhancement to) the Information Management System (IMS) of LTER sites, or other groups or investigators. In the past year, version 1.0 of the schema was released, and since then the group has added support for the new semantic annotation, dataset license, funding information, and taxonomic identifier components of EML 2.2.
- *Zotero.* This WG seeks to develop best practices for using Zotero as a reference manager for LTER sites to facilitate sharing key products with the LTER Network bibliography, submitting product citations to Research.gov for NSF reporting, and enabling a searchable bibliography for an LTER site's website. This year, the group made slight improvements to the example JavaScript Zotero client, such as enabling dataset DOIs to be associated with publications in a Zotero library.
- *Non-tabular Data.* This WG is developing best practices for describing and archiving non-tabular data such as genomics data, videos, and software. The group has several best practices in draft form and will eventually formalize them in a GitHub repository with an associated GitHub pages website.

Planned activities for the coming year:

- The annual IMC meeting is planned to be held August 2, 2020 in conjunction with the annual meeting of The Ecological Society of America (ESA) (August 2-7,

2020) in Salt Lake City, Utah. The meeting's theme is *Harnessing the ecological data revolution*. This being the 40th year of the LTER Network, the LTER plans to have a large presence at this meeting to highlight the past 40 years of LTER.
    - Members of the IMC have been encouraged to submit abstracts for sessions and posters. The IM-Exec and IMC will plan a fruitful annual IMC meeting, and coordinate IMC participation.
- DataBits articles will take the form of individual blog posts on either the LTER or EDI website as appropriate, with articles collected subsequently into an annual issue.
- Working groups with an agenda for the coming year (outlined below) will work to meet identified goals. New working groups may form to address issues not identified at the time of this writing.
    - *IMC Website Improvement & REDesign (WIRED).* The WG will explore ways to develop an information management section under the upcoming new LTER website platform once the platform is ready.
    - *EML Congruency Checker (ECC).* This WG will consider several proposed data-quality checks for inclusion in PASTA+ for uploading data.
    - *Semantics.* This EDI-led working group is considering options and strategies for developing a workshop around this topic.
    - *ClimDB/HydroDB 2.0.* This WG will work to further plan and implement strategies identified for maintaining ClimDB/HydroDB.
    - *Information Management Knowledge Exchange (IMKE).* In the coming year, select IMC VWC meetings will continue to facilitate sharing of IM tips and strategies. Rather than a formal working group, IM-Exec is responsible for planning these events.
    - *LTER Core Metabase (LCM).* This WG is developing scripts to populate LCM from EML, enabling sites to migrate to LCM by supplying existing EML documents from their data catalog.
    - *Non-tabular Data.* This WG will finalize several best practices related to non-tabular data and incorporate them into EDI's data package best practices GitHub repository (https://github.com/EDIorg/data-package-best-practices) and associated website.

Upcoming changes in leadership, purpose, or process: Goals and processes will remain unchanged, and the current composition of LTER IMC leadership (i.e., IM-Exec) will remain unchanged until the IMC annual meeting (August 2020) when Downing's (BNZ) term ends and a replacement member is elected. IM-Exec is cognizant that the terms of four of the current five members, including both co-chairs, are scheduled to end summer 2021. IM-Exec is considering whether possible actions, such as adjusting the terms of a subset of current members, may be required to maintain a sufficient level of continuity beyond 2021, and will take steps if and as appropriate.

Specific question for LTER Science Council feedback:
Scientific approaches that draw upon collaboration and integration of long-term, multi-site data across broad spatial and temporal scales are critical to meeting ecological and environmental grand challenges. While most LTER data and associated metadata are

available through the Environmental Data Initiative, DataONE, or other relevant repository, assembling and synthesizing data from across networks, locations, and studies can be challenging. One area where data scientists and information managers have focused efforts to help facilitate data aggregation is around climate and hydrologic data. LTER and U.S. Forest Service (USFS) scientists and information managers developed ClimDB/HydroDB (https://climhy.lternet.edu/) as one approach to improving access to cross-site data. ClimDB/HydroDB is a web harvester and data warehouse that provides uniform access to common daily streamflow and meteorological data through a single portal. ClimDB/HydroDB proved successful as a bridge technology between rigid data distribution models and early service-oriented architectures, and has been critical to the success of numerous research efforts. However, the ClimDB/HydroDB approach relied upon scientific interest, sizable organizational commitment, and participation incentives. Though very successful and well used and cited for over a decade, resources were not dedicated to updating the infrastructure, the change in LTER leadership organization had no support for the current infrastructure, and contribution was no longer a requirement for LTER sites. Thus, contributions to ClimDB/HydroDB gradually decreased, eroding its utility. However, LTER and USFS practitioners have developed a new model and approach to ClimDB/HydroDB that capitalizes on existing technologies to streamline the flow, aggregation, and storage of data.

Even with the new model and approach, some overhead is unavoidably required to maintain this resource. The new approach may not be successful without interest and commitment of resources. The LTER IMC is interested to understand how the Science Council values this resource, and whether they are open to asking their respective sites to contribute the small, but real, effort required to maintain and populate this Network- and community- wide resource.

The initial successes of ClimDB/HydroDB demonstrate the utility and value of such resources to the scientific community. Yet, ClimDB/HydroDB focuses on a narrow type of data (i.e., predominantly sensor-derived, climate and hydrologic data). A separate effort by the Environmental Data Initiative using a different model focused on community population data suggests that information managers working with domain scientists can adapt these approaches to data of any type or domain. Given that possibility, the LTER IMC is interested to know the data types and/or domains where the Science Council would be most interested to focus data aggregation/synthesis efforts, and whether they are open to contributing the resources and expertise needed to bring such products to fruition.