Sustainable Value and Generativity in the Ecological Metadata Language (EML) Platform: Toward New Knowledge and Investigations

Ben Li
University of Oulu, Department of Information Processing Science
banjili@gmail.com

Abstract

This paper examines Ecological Metadata Language (EML) as a generative platform facilitating new ecological research. It reflects on literature about the EML platform, and on the EML platform itself. First, it identifies a substantial gap in literature about use of the EML platform for intended research. Second, it identifies some strengths and weaknesses of the EML platform to support research about variance, process, and configurational theories. Third, it examines the EML platform's strengths and weaknesses in mediating values, particularly those concerning new kinds of ecological research envisioned in EML literature. Finally, it contributes some brief directions for future research, including: expanding notions of valuable (meta)data, of use and of users; articulating clear value; and exploring the morphology of (meta)data.

1. Introduction

Knowledge flows among communities have been of sustained information systems (IS) research interest. Research has been limited due to lack of access to settings in which stakeholders, ISes, and external realities align well for systematic investigation. This paper presents part of a case where variables appear to favour highly generative IS-mediated knowledge flows, but expected generative outcomes remain elusive.

This paper examines Ecological Metadata Language (EML) as a generative platform for ecological research. A key goal of the EML platform—an assemblage including the EML standard and tools that implement that standard for data entry, search, and retrieval—is to enable datasets to be openly described, published, and indexed like research publications in order that they may collectively inform research in new ways that individual datasets could not. This paper reflects on literature about the EML platform, as well as on the EML platform itself (standards, software, data, stakeholders), in order to better understand the EML platform's generativity in Zittrain's sense [1]. Specifically, this paper pursues the question: How does the EML platform support new kinds of research? The main contributions are ways to consider generativity in the EML platform, from a perspective underrepresented in the literature.

This paper proceeds as follows. First, it identifies a substantial gap in the literature about the EML platform concerning its use for ecological research. Second, it identifies some strengths and weaknesses of the EML platform to support research about variance, process, and configurational theories that stakeholder ecologists seek to collaboratively pursue. Third, it examines the EML platform through a value lens. Fourth, it offers some brief directions for future research for both IS and ecological stakeholders.

2. Reviewing the EML platform

Since the EML specification was conceived in the 1990s, much published work has focused on making ecological (meta)data more useful through connectivity and discoverability (e.g. [2]). Baker et al. [3] are among many to distinguish technical and social aspects of collaboration through the EML platform at different scales. They frame social dimensions such as articulation work to implement particular EML technologies and data management practices. The above is underpinned itself by technical work involving programming silicon, building computer networks and databases, etc. The EML community is thereby supported as an enduring set of relationships between changing people, changing research concerns, and changing technologies. A full literature review would be too lengthy to provide here.

Missing is a substantial examination of reusing data for research. Literature about long-term ecological research (LTER) networks' large-scale collaboration (e.g., [4], [5]) rarely mention the EML platform, and only infrequently refer to metadata. The EML platform may be so engrained in practice that authors routinely forget to mention it, or it may be so peripheral that it has no bearing on practice.

978-1-4799-2504-9/14 $31.00 © 2014 IEEE
DOI 10.1109/HICSS.2014.440

For all the discussion about "integrated routine and knowledge work" and a "spiral model of IS development" [3], the LTER information management literature of today (e.g., [6]) reads quite like literature from the 1990s, focusing on organizations that generate information and adopt the new IS to expose their (meta)data to others. Curation of (meta)data as discussed in LTER from the supply-side [7] does not actively consider *for whom* the (meta)data is curated, nor for what purposes. For example, when [7] was authored in 2006, quotes were of stakeholders discussing use of shared (meta)data in the prospective future tense, rather than as current practice. Rare is literature about reusers of exposed (meta)data to write scientific papers (e.g. [8]).

Exportability of metadata is touted as enabling automated discovery, harvesting, reuse, etc. [9] Reusability *by (re)users* is not always a priority. Gathering data for reuse reveals practical barriers:

- Exported XML files violate standards, e.g. [1]: `<definition</definition>`.
- EML tools that do not export XML or structured data (e.g. ILTER DEIMS[2]; OBIS[3]) but do support metadata written in "Imperial Aramaic (700-300 BCE)".
- Broken links from directories (e.g., [4]) to abandoned domains now owned by squatters (e.g., www.lter-tern.org formerly "Terrestrial Ecosystem Research Network", now "自殺は本当？矢島祥子"; which machine translates to "Suicide is true? Yajima Sachiko").
- Almost complete lack discoverability of (meta)data repositories across languages.

Challenges with using the EML platform appear more broadly in ecology as well. In 2011, [10] documents several unresolved challenges when attempting to mobilize metadata. In [11], authors developed their own use case about metadata purpose and alignment issues. Presumably no suitable case was available of climate researchers who struggled to manage and document their climate (meta)data—such as at the University of East Anglia's Climatic Research Unit which drew international headlines for such an issue in 2009 [10], [12]—*and* whose researchers sought a formal metadata IS to make data transparently available to other potential users.

A 2012 paper [13] describing the participatory development of a second-generation ecological (meta)data platform to handle new data sources surveys stakeholders' views and *potential uses* of an EML platform. In their "lessons learned" section, authors and developers do not refer to knowledge about stakeholders' *use* of first generation systems.

Few authors have mentioned using the Metacat EML tool in their publications about *doing actual* ecology, e.g., [14], [15], which refer to individual data sets found on a Metacat server, rather than in publications *about the doing of* (long-term, spatially wide) ecology with Metacat and other tools, e.g. [16].

It is curious that DataONE [13] reports from its 2009 survey that "Over 70% of respondents either do not currently use any metadata standard to describe their data (50.8% n=676) or use their laboratory's own standard (20%, n=266)" without reporting anything at all about what proportion of its total 1329 respondants are consumers of metadata. Nor does DataONE provide data about metadata usage in other reports [17]. (Ironically, DataONE does not make its survey instruments available in its repository, inhibiting reuse of their data.) Data reuse would seem to be key context with respect to understanding qualities about the old platform that data non-experts find valuable or broken, and to test whether proposed features of the new platform would align with capabilities that users actually value, rather than aligning to features that the users claim to want (having been told so by data experts in the literature). It is telling that DataONE's "Usability & Assessment Working Group Charter" [18] frames usability in terms of operating technical tools, rather than the handiness of those tools for the purpose of conducting or finding data to support ecological research.

Absence of evidence of success of the EML platform to generate massive new cross-site, cross-time, system research is not evidence of absence of success. Reports of EML platform failures are less common than reports of successes, as expected [3]. But it is conspicuous that a decade after formal adoption, few discuss replicating award-winning research drawing together masses of data enabled by the EML platform. One might view the plethora of papers about extending the EML platform (e.g. [2]) as evidence of absence of success. Stakeholders would not spend years to find and make a place for the EML platform among the other research tools if it was already generative. If the EML platform has sustainable external interest, papers about the topic would be written by more than the same small community of ecology-IS authors, and would include new stakeholders, as with other IS innovations that become widely adopted.

[1] http://tropical.lternet.edu/knb/metacat/knb-lter-and.3234.5

[2] http://data.lter-europe.net/deims/

[3] http://iobis.org

[4] http://www.ilternet.edu/member-networks/loc-au-australia-lter/

This paper proposes that the gap between the EML platform's instantiation and widespread use arises from the systematic lack of consideration from the generative *reuser*'s perspective as outlined above, in favour of the *operator user*'s perspective. The next section argues that this is not reflected only in the implementation, but also in the EML standard.

3. The EML standard

EML has no explicit provisions to record specific theories or kinds of theories behind a dataset described by metadata, [39], nor does it record if or how (in terms of theory) a dataset has been (re)used. Not specifying theory through the metadata standard allows data arising from any theory to be described by the standard, but that omission does not actively encourage verifying that the standard accommodates data supported by, and supporting, the kinds of theory required to pursue the large temporal and spatial scale research envisioned to be supported by standardizing on EML.

By tacit default, EML gives attention to: when and where observations are made by the supplier; units, methods, and sampling techniques of observation; taxonomy of the organisms involved; the physical specifications of the data storage medium; supporting administrative bureaucracy responsible for a data set; and access rights to the dataset. That is, EML is about data that can be expressed in terms of quantifiable discreet location-bound phenomena that may be observed to change over time. Those are among things that ecology is about, and those dimensions of ecological research have been greatly aided by ISes, but long-term ecology is also about many other things.

For example, with respect to waterflea larvae, species identification depends on the theories employed by the observer. The observer can only discriminate based on morphology of a particular barb on a particular limb if employing a theory that the barb observation is significant. A novice observer might not know a theory about the significance of the barb, so might not notice it at all.

A low-level problem is that EML provides no way to encode the theory used to gather data. It does not reveal whether the lack of observation of a particular species of water flea is due to: actual absence of that species; non-use of a theory that would enable that species to be detected through differences from other similar species; inadequate application of that theory; use of a theory that suggests differences are insignificant at the required level of analysis, etc. Therefore, a potential reuser could not reliably use EML metadata alone to directly discriminate amongst data collected according to one set of theories from those collected according to another.

At a high level, the EML platform's tracking time and space (and bureaucracy!) as the most significant independent variables structurally disadvantages research about changes to organisms and ecologies arising from ensembles or systems of changes in temperature, nutrition, contaminants, etc. It favours theories expressed in direct relationships of variance.

Ecology is also interested in combinations of factors, and not just at a particular place and time, but how they cycle and repeat. IT has not traditionally been well adapted to data that are not matrices, procedures, or simple relations. A data series and a map are both matrices, a list of instructions can be viewed as a procedure, and sets of similar matrices reveal basic relations. Data managers complain about uncontrolled proliferation of poorly versioned Excel spreadsheets and image files, conflicting policy and procedures, and too many dependencies. But rarely do they complain about an office suite making it too easy for users to make, modify, and email too many versions of system models or diagrams, simulations, visualization models, etc. Yet the EML platform was intended to facilitate exactly that high-level ecological understanding: bringing together many low-level observations to gain systematic knowledge.

The kinds of data that enabled the formulation of the theory of natural selection as a gradual non-random variance of biological traits, simultaneously relating morphology of different species across different local physical conditions, would not be expressible in the EML platform in a way that would make those related variations searchable or visible. The timescale encoded as observation dates during Darwin's 5-year journey, and the latitude and longitude coordinates of the places of observation, required by EML for structured searches, are completely irrelevant to the theory. That many observations were only made for a single study period would today prejudice Darwin's observations and collections as being insufficiently long-term to provide great research value! (Subsequent researchers have doubtless found great re-use value in Darwin's data and collections, without the help of anything like the EML platform.)

The EML platform's goal is not to re-discover existing theories, but (in part) to enable discovery of new theories, and to do new kinds of collaborative ecology that individual researchers and projects cannot. Enabling that kind of generativity within the EML platform requires slicing (meta)data along axes other than time, space, species, and bureaucracy; and new relationships and covariances among the (meta)data must be (re)user definable, storable, and searchable along relevant axes of discrimination.

Although much ecology work necessarily builds and extends local theories, it is long recognised that ecology

deals in complex systems and processes [19]. Following el Sawy et al. [20] who discuss ecodynamics in information systems, here we reuse an IS theory categorization that also fits ecological theories.

Variance theories concern individual variables that are independent, necessary, and sufficient predictors of outcomes. But these do not work well when elements and relationships both change under disequlibrium. For example, ecological theories about basic predator-prey and nutrition-growth relationships fit here, but those alone do not explain long-term predator-prey cycles or disturbances due to disease or other events. The EML platform's bias toward variance is understandable as arising from long-standing statistical approaches to data processing in ecological sciences [21].

Process theories explain how a phenomenon occurs over time and addresses discontinuous changes of state or phase changes. They do not work well to understand emergence holistically. Ecological theories about reaching or disrupting local equilibriums, invasive species, etc. fit here, but such theories are often highly specific to particular localities and contexts.

Configurational theories simultaneously consider patterns and combinations of elements, and patterns and combinations of outcomes, rather than reductionistically focus on specific elements or specific relations. But containing configurational theories' ability to consider 2^n permutations of n elements is considered a weakness. Although configuration theories do not inherently track shifts over time, configuration theories in conjunction with process theories, are capable of "modest generalization" strongly connecting context and conjecture with causality.

For example, theories about ecological succession with respect to changes in species structure over time would fit here. However, cross-scale ecological observation is often limited at the microscopic range by the sheer density and diversity of micro-organisms and lifecycles, and at the macroscopic range by the vast number of possible internal and external interactions of any defined system over time. (Tracking large quantities of things would seem to be a strength of information systems.)

The EML platform's competence at describing matrices (metadata about column headers and about GIS are heavily emphasized in the standard) and taxonomies disadvantage process and configurational theories. The EML platform does not include provisions for a taxonomy of processes or relationships analogous to taxonomies of organisms, and provides only limited support to describe simultaneous geographic configurations (biomes, maps), or for identifying co-variance. The default does not store or expose running system configurations, and users cannot systematically encode or search for high-level categories like "nitrogen fixing bacteria", "birds of prey", or "cycle" beyond haphazardly using the *keywords* field. It is perhaps not the role of the EML standard to categorize the entire field, but the standard could refer to and leverage more existing external standards as it does with GIS and imaging standards.

Standards designed for different purposes will have different scopes and foci [11], and will be more suited for work arising from different kinds of theories, such as historical path dependencies, current configurations and relationships.

Recall that main goals of LTER, for which the EML platform was built, are to "conduct research on ecological issues that can last decades and span huge geographical areas... on regional and continental scales… Research is located at specific sites chosen to represent major ecosystem types or natural biomes; It emphasizes the study of phenomena over long periods of time, based upon data collection in five core areas"[5]. Those sites and research programs have specific outcomes in mind for a single national (U.S.) research network. Early in EML's history, knowledge supporting process research was "othered" [22] in a paper aptly titled "Nongeospatial metadata for the ecological sciences", while geospatial configurational approaches (and roles) were prioritized over all other configurational approaches [23].

Expanding "significant integrative, cross-site, network-wide research" internationally emphasizes that individual sites must represent sites of particular ecological configurations, in addition to being sites with specific contexts, in order to generalize (within bounds) data, theories and knowledge from such sites. Variance theories alone do not address region- or continent- wide semi-synchronous realities, let alone discover systematic similarities across contexts.

Important discussions emerging about ecology—from ecology, policy, and other communities—concern wanted or unwanted large-scale innovations in the environment, such as disturbances with unclear impacts for the long term. If LTER is to contribute to those discussions by discovering, providing and contextualizing knowledge beyond its own local communities, it needs an approach to data that both *understands* systems and *facilitates* ecological work at scale, where local changes are too numerous or varied to address through variance theories alone and would benefit from process and configurational approaches. That potential to scale is considered a key valuable aspect of LTER [16], and requires a knowledge infrastructure supporting interdisciplinary approaches.

[5] http://www.lternet.edu/

4. Value and sustainability in LTER IS infrastructures

So far, this paper has discussed some practical considerations about mobilizing ecological data using the EML platform across broad scales, and about some of the underlying features of the EML platform facilitate work based on different kinds of theories. This section examines implications of the platform's features upon: a) values mediated by the EML platform, and b) sustainability of the EML platform.

Value and sustainability are easily identified operationally, but difficult to define and design. This paper adopts Bruntland's concept of sustainability [24], as the ability to simultaneously: a) continue current practices, and b) enable new practices, while maintaining benefits from current practices. Practice can be realized as almost any reoccurring event: a job, an industry, a publication cycle, a nutrient cycle, an annual bird migration, a training program, a pattern of engineering, etc., and more sustainable practices may displace less sustainable practices. Thus sustainability is compatible with ecological and IS perspectives (e.g., reviewed in [25]).

Value, on the other hand, is much more difficult to define in general, but knowledge about flows to and from the EML platform helps to relate value and sustainability as complements.

a) If sustainability of ecological research depends on finding and reusing data (not just publishing it), then the EML platform mediates value between those who contribute (meta)data and those who reuse it. Therefore, value must be linked to the (meta)data itself, and to how it enters and exits the EML platform.

b) If the value provided by the EML platform depends on its ability to aggregate and present (meta)data supporting knowledge, then value must be linked to sources of (meta)data and their connections.

Researchers generatively using (meta)data from the EML platform to conduct and publish new knowledge derive value from, and may be sustained by, the EML platform. In turn, researchers who contribute (meta)data and knowledge to the EML platform may provide value to the platform and sustain it. Thus, researchers and the EML platform can gain value through the flow of (meta)data. (Meta)data itself can gain or lose potential to yield value by association with other data, and through (re)use.

4.1 Diving into value and sustainability

This short paper cannot resolve fundamental philosophical issues of value. Instead, it focuses on the few underlying (subjective and objective) accessible bases for value with respect to ecological (meta)data, reflecting Goulder and Kennedy's [26] analysis of the value of ecological features that underlie (meta)data discussed here. That view is largely compatible with networked IS infrastructure notions of value (e.g., [27]).

4.1.1 Establishing context. Things derive value from context, since value is in reference to specific (potential and avoided) tangible and intangible interactions. A main objective of EML is to describe data, to explain where and how it is collected, and what it is about, explicitly in order to give the data and the knowledge it supports more possible contexts of combinations and reuse, i.e. more interactions and generative opportunities, to amplify the value gained from the original data collection expense [31]. The largest LTER metadata repository is the *Metacat* metadata server operated by KNB[6], which harvests metadata from a handful (almost all) other Metacat servers in the International LTER Network, so it is reasonable to examine how much value KNB's EML platform extracts and provides through metadata.

Over 19,500 of the 27,500 XML files in KNB's EML repository contain metadata descriptions, including XML formatting, 16,000 characters or less in length. Those 20 to 400 words of (usually English) prose and bureaucracy cannot provide much unique descriptive information despite requirements by public research funding schemes to make data reusable. Further, two-thirds of EML files (17,363 files) recorded by KNB appear to be near duplicates, with sequential file names and nearly identical contents and descriptions, sometimes differing only by geographic coordinates and/or dates. They add little unique information. (The EML standard does not specify how to indicate that a particular EML file is generated or copied and pasted from another.) Thus the files provide little value through their (non-)unique ability to establish context for their underlying data. This is empirically verified by the high compressibility of the text of 27,476 public KNB metadata records (excluding XML formatting tags) as compared to the text of titles and abstracts from 22,805 bibliographic entries for the International LTER Network's collective research outputs. See Table 1. (An unknown number of metadata records are not made public.)

Discoverability of context via the EML platform is limited: even for datasets supported by a variance theory, it is difficult to discover data sets in the context of other phenomena that vary on seven-year cycles; or in other contexts where particular interactions or relationships occur. Those kinds of data contexts and combinations are especially relevant to emerging research and policy questions such as those outlined by major international

[6] http://knb.ecoinformatics.org/

organizations (e.g., [28], [29]). These are not faults of the EML platform which merely echoes how data is documented and indexed in ecology. But that also points to missed generative and knowledge flow opportunities supporting sustainability.

Table 1. Information content of metadata.

	Text size (bytes)	gzipped size (bytes)	Compression ratio
ILTER Titles and abstracts	2,850,890	892,134	0.313
KNB (all)	284,251,397	18,026,803	0.063
KNB (near duplicates)	263,755,126	4,707,207	0.018

4.1.2 Defining objectives and preferences. A primary goal of the EML platform, through tools like the Metacat (meta)data server, was to provide a new way to consider long-term ecological knowledge to inform societal actions [30], via publications and other outputs and interactions. The value of a (meta)data set and the EML platform itself therefore arises through use which sustains both research and social practices. Google Scholar records 463 instances of "Metacat" in the scholarly literature, including many about unrelated metacats such as Hofstadter's automated analogy approach. Of the 146 scholarly outputs about "Metacat" since 2009, there are as many or more papers about enhancing EML tools as papers drawing on EML data. Since there are approximately 900 ecology journals in which results from synthesizing EML (meta)data could have been published, the impact of Metacat on generating ecological publications has not been wide-reaching.

Similarly, Metacat instances at metacat.lternet.edu and at knb.ecoinformatics.org collectively index approximately 32,000 items of metadata, yet citations of data from those sources are modest. According to Google Scholar, "knb.ecoinformatics.org" appears in 338 publications in total, including publications that refer to KNB's metadata platform. Similarly, "metacat.lternet.edu" appears 33 times in reference to the US LTER's metadata portal, and another EML metadata portal, "data.gbif.org", appears 610 times. In all three instances, the domain names given must appear in citations per the portals' usage agreements. By comparison, since Google Scholar's general availability in November 2004, it has been mentioned nearly 1.5 million times in publications.

It is possible that the EML platform and portals are not mentioned because they have become invisible infrastructure in ecological research, in the same way that authors no longer generally list the model of workstation used to compute results. But it is professionally unlikely that they would all use data without citing it as required by the data providers.

From a handful of exceptional examples of spatially and temporally broad research (most cited in [5]) which only rarely mention the EML standard or platform, it is difficult to ascertain the value provided by the platform through use.

We might look at adoption of the EML platform itself as a value-generating innovation. A Google search for default strings appearing in the Metacat web interface revealed fewer than two dozen distinct public Metacat servers worldwide. The ILTER network includes 40 national network members, and thousands of participating researchers. The bugzilla for Metacat[7] lists 19 authors of 148 bugs between 2001 and 2013, while the bugzilla for EML[8] lists 20 authors of 73 bugs. In total, there are 27 different authors of bugs, only four of whom each made one single appearance. Either the Metacat and EML originators have discovered how to produce an exceptionally complete and bug-free standard, and an equally spectacular server platform, or few stakeholders outside the developers uses EML or Metacat enough to contribute suggestions or bug reports. Inspecting the lists of registered "reporters" of bugs for Metacat and EML shows that of the 236 reporters for EML, 234 are also reporters for Metacat.

Examining the Metacat documentation reveals that the Metacat software is open but potentially difficult to adopt because of old dependencies. It recommends Java and Windows or Linux versions all from the mid-2000s, which are difficult to obtain and difficult to deploy securely.

Many of the research and policy potentials of (meta)data are positive externalities accruing to those who directly collect or (re)use the (meta)data. As such, the value of (meta)data is difficult to assess without hindsight. Yet some sense of valuation of (meta)data is required in order to plan ecological research and IS investments with that (meta)data.

The social value from shared LTER data relates to informing our ability to continue to be intervening in ecological systems. We prefer knowing about sustainable ways to intervene, and to help express those ideas in terms of social, policy, commercial dimensions. Even though we are explicitly not discussing monetary value of ecological (meta)data in most cases, researchers prefer this value to return sustainably through research funds and support.

For some participants, particularly those winding down their careers or research projects, the intangible value of knowing that their data may help (unknown)

[7] https://projects.ecoinformatics.org/ecoinfo/projects/metacat-5

[8] https://projects.ecoinformatics.org/ecoinfo/projects/eml-2/issues

researchers in the future is sufficient motivation to spend weeks or months to describe and publish their data. For others, (meta)data and its infrastructure may provide sustainable routes to advance a career, to secure financing, etc. Since (meta)data can be consumed (to the exclusion of a second-mover's ability to author the first publication) with respect to specific fields and topics, transitions of already opened data between rival and non-rival contexts may be important to considering value of (meta)data.

We could also consider the present value, and potential for future use, or infer value from cost paid to get to the (meta)data, the cost of time to input that (meta)data, and the cost to maintain that data through generations of ISes. Or through option value: the value of preserving unique combinations and configurations of datasets and collections to be available for future use. Or in terms of a price paid to prevent the (meta)data from becoming unavailable.

Since the IS infrastructure facilitates data to be more easily considered in the context of some kinds of research theories and approaches and than others, the value of the (meta)data is also linked to the IS infrastructure housing it, and to knowledge networks.

These are all potentially interesting ways to investigate and explicate the value of (meta)data, but the current EML platform does not document cost, value, reuse, or even interest in reuse, in an accessible way. It does not expose the value it may generate.

We can consider value propositions of the EML platform as an infrastructure and as something in development. A value the EML platform claims to be offered is the ability to participate in long-term, with goals including improved ability to inform research and social outcomes. A value claimed occasionally by external stakeholders appears to arise from individual snapshots of data to support non-LTER analyses. And a value claimed by internal stakeholders is publishing about publishing data, gaining enhanced EML infrastructures, publication records, etc.

But based on how the EML platform is used, there appears to be a mismatch between the *prospective* value that EML originators want to offer, and the *actual* or *perceived* value available through the IS that implements it. The EML platform as a mediator of value could become more generative by enabling more and more kinds of contributors and (re)users to use the platform to intermediate more kinds of value not explicitly designed into the EML platform.

On the research question, public evidence shows that the EML platform offers only limited support for new kinds of research. Several ways to orient the EML platform toward more generative roles have been identified. The next and final section of this paper takes a high-level view of some opportunities.

5. How (IS) researchers might consider value in metadata, to design for enhanced sustainability

So far, this paper has argued the following. First, studies of the EML platform as an IS infrastructure have focused on supply of (meta)data, rather than on (meta)data reusers. Evidence suggests little practical reuse. Second, the EML platform prioritizes particular kinds of theory and value over others. Long-term variation systematically disfacilitates other relevant ways of organizing ecological (meta)data to pursue process and configurational theories and knowledge. Third, the EML platform does not appear to be sustainably generative. It disfacilitates data (re)users from adding value back for other users. This final section reflects on key points to sketch some directions toward a more sustainable collaborative infrastructure. Most of these ideas are individually not new, and many would apply to other collaborative knowledge and information systems. As a case study for further analysis, the EML platform offers opportunities to study expansion in several dimensions.

5.1 Expand notions of valuable (meta)data

Beyond matrices and instructions, socially valuable metadata about individual data, and about sets of data and sets of metadata may include: theories, reuse, recombination, search, visualizations, disruptions, system context, etc. Any individual dimension or set of those dimensions may be a useful discovery or access point to knowledge, including dimensions external to the data itself, such as categories of work. In ecology specifically, short-term studies can almost always be repeated to become low frequency long-term studies. The EML platform would benefit from new ways to detect past datasets that would benefit most from being updated or repeated. Kepler[9] is promising, enabling users to save and reuse data processing workflows with Metacat and other platforms.

5.2 Expand notions of use and users

Platform standards and large infrastructure must keep up with advances/changes in technology as well as advances in their use to support knowledge discovery and sharing. For example, wireless sensor networks, motes, immersive visualization, etc., were not prevalent when EML was designed, yet are now common in ecological research.

[9] https://kepler-project.org/

Like most platforms at the outset, the EML platform favors users with existing related capital and knowledge investments. In this case, such users are institutional members of nation-scale research and knowledge networks. Generative systems must facilitate and enable participation from many scales, and therefore more kinds of use than envisioned by (in this case, institutional) originators. It is valuable to ask whether a new generative platform would enable stakeholders to receive at least the same kinds of value and facilitate the same kinds of knowledge flows as systems and practices that would be replaced. We might pursue better ways to investigate whether the most successful or valuable outcomes of the previous infrastructure be facilitated or even possible with the new platform. We also want to investigate the processes inferred by stored observations about the environment represented in the platform [34].

5.3 Articulate clear value

EML platform originators provide clear value propositions for contributing informative descriptions of data, such as more citations without conducting new research work, but not for *reusing* data to develop knowledge. Finding groups of interesting data to reuse still requires substantial effort. However, the platform can help by avoiding value-destroying representations of stored (meta)data. For example, a single two-year data series about intertidal site temperatures at Fogarty Creek[10] need not be described eight times as each a one year series, or appear 21 times in lists or search results. Platform authors could demonstrate (not just model) an entire cycle of value flows in and around the platform, to encourage participation beyond grant compliance or metrics purposes, perhaps taking ecosystem service value [32] as a model.

5.4 Explore the morphology of (meta)data

As of mid-2013, there was no facility to gather, let alone calculate or visualize properties of (re)user selected sets of (meta)data in the EML platform. The ILTER Network could not discover what knowledge is supported by its own (meta)data. However, new global-scale ecology questions require access to (meta)data at a global scale. Manually (re)indexing or (re)adding new metadata at the scale of 10^6 records would not be sustainable. Therefore, it would be worthwhile to discover advances to generate some new metadata and categories automatically, to at least surface examples like "datasets containing 7-year patterns involving coastal predators", and "datasets relevant to spills of produced water". That would require conceptual advances in both IS and ecology, beyond raw information content, to find or cluster interesting similarities among ecological (meta)data.

5.5 Links to key knowledge flow problems

Reflecting on broader knowledge flow research, EML provides interesting but mixed evidence.

First, the kinds of information encoded and transmitted via the EML platform are not very complex. The knowledge flows it facilitates are relatively basic and can be patched locally by the socially and geographically distant recipients [35]. The strong networks and collaborations manifesting physically at LTER network meetings show that complex knowledge flows require complex offline interactions, and that intermediate forms of value in knowledge are not expressed in the EML platform. The EML platform may generate collaborations, but as an address book rather than a (meta)data platform.

Second, this case highlights diverse motives and incentives for participating in knowledge flows [36]. While the US LTER network (among others) is mandated to participate on the supply side of the flows, they are not mandated to consume flowed knowledge to realize scientific and social value. This is complicated by the (policy) demand that individual researchers and organizations both produce and consume non-local data, without resources or EML facilities to participate in knowledge or document cycles. Elsewhere in the world, some data suppliers attempt take advantage of low local labor costs by joining the network and flooding the community with data, in hopes of attracting reuse [40].

Third, the EML platform provides some insight into the roles of knowledge and expertise clustering [37] with respect to stimulating research innovation (e.g., [38]). Virtually supplying documented practical knowledge about local unique or shared value-generating activities, without a persistent shared geographic anchor, does not alone stimulate new kinds of activity across the entire community. Within member networks and research sites, some senior researchers treat the requirement to attempt interdisciplinary knowledge flows as a structural hassle than a professional requirement.

5.6 Limitations

The EML platform arose through a unique set of circumstances and history as partially documented in [33]. "EML is settled" is a refrain informally heard when new and prospective international LTER participants are enrolled into the platform and network.

[10] Datasets FCKX00_XXXIBTNXTSR01_20080730.50.2 through FCKX00_XXXIBTNXTSR01_20101008.50.2

The intent of this paper is not to judge that process, but to highlight some practical concerns from a newcomers' perspective to improve and use that infrastructure to do interesting ecological work.

Many dedicated groups and individuals and their circumstances made the EML platform into what it is today, as reflected in the networks of authorship about that platform, and walls of assumed history only sparingly available to outsiders and new members. The data (re)user only sees the few latent details on the surface of the EML platform as it is presented.

Previous attempts to draft this paper have met with informal resistance from LTER communities concerned about how the voice of the communities should be expressed. This paper argues that it is of scholarly and practical interest (and has been advised to that effect by several members of the LTER community) to raise these observations about the EML platform's capabilities with respect to its intended uses on two grounds. First, the composition of major LTER communities continues to diversify as it expands, and is already home to many new voices and perspectives. And second, the EML platform tools examined here are recommended to newcomers are not necessarily "settled" for adopters who must adapt. This paper is explicitly NOT written with the voice of any particular LTER community or member.

Most of the EML stakeholder community has not formally expressed views about the EML platform. As noted, the formal discussion has been dominated by EML platform originators to the exclusion of end (re)users and even data suppliers who face widely discussed but rarely published operational challenges with the EML platform. The critical views presented here align more closely with views held by newcomers outside the original US LTER network, than from within that settled network. This paper likely over-represents diverse views of the broader international LTER stakeholder community, and underrepresents views of the US LTER community for whom the EML platform has been the long-committed and mandated standard. As such, this paper compliments, rather than contradicts, the established body of US-focused scholarship on LTER ISes by contributing an alternative fresh perspective.

This paper is written (hopefully) plainly, and without any intent to hide anything "between the lines". Individuals with strong personal connections to LTER, the EML platform, or studies of those may (quite legitimately) remark that more nuance or alignment with each of their views would provide valuable context for interpretation.

6. Conclusions

This paper has reviewed several ways in which the EML platform has focused largely on suppliers of (meta)data, to the disadvantage of users of that (meta)data. It has also argued that the EML platform privileges knowledge based on variance theories over knowledge based on process or configurational theories. It has looked at the EML platform's potential to be generative with respect to addressing new systems questions not directly addressable through variance alone. And it has provided some suggestions to enable the EML platform to better link value and sustainability in its supporting and supported practices.

7. References

[1] Zittrain, J. L., "The generative internet," Harvard Law Review, vol. 119, 2006, pp. 1974-2040.

[2] Berkley, C., S. Bowers, M. B. Jones, J. S. Madin, and M. Schildhauer, "Improving data discovery for metadata repositories through semantic search," CISIS'09, 2009, pp. 1152-1159.

[3] Baker, K. S., and K. I. Stocks, "Building environmental information systems: Myths and interdisciplinary lessons," HICSS 40, 2007, pp. 253b-253b.

[4] Vihervaara, P., D. D'Amato, M. Forsius, P. Angelstam, C. Baessler, P. Balvanera, H. Boldgiv, P. Bourgeron, J. Dick, R. Kanka, S. Klotz, M. Maass, V. Melecis, P. Petřík, H. Shibata, J. Tang, J. Thompson, and S. Zacharias, "Using long-term ecosystem service and biodiversity data to study the impacts and adaptation options in response to climate change: insights from the global ILTER sites network," Current Opinion in Environmental Sustainability, vol. 5, no. 1, 2012, pp. 53-66.

[5] Kim, E. S., and Y. S., Kim, "Current status of Korea Long-Term Ecological Research (KLTER) Network activities compared with the framework activities of the Long-Term Ecological Research (LTER) Networks of the United States and China," J Ecol Field Biol, vol. 34, no. 1, 2011, pp. 19-29.

[6] San Gil, I., M. White, E. Melendez, and K. Vanderbilt, "Case studies of ecological integrative information systems: the Luquillo and Sevilleta information management systems," in S. Sánchez-Alonso, I. N. Athanasiadis (eds.) Metadata and Semantic Research. Springer Berlin Heidelberg, 2010, pp. 18-35.

[7] Karasti, H., K. Baker, and F. Millerand, "Infrastructure time: long-term matters in collaborative development," Computer Supported Cooperative Work (CSCW), vol. 19, no. 3-4, 2010, pp. 377-415.

[8] San Gil, I., "Metadata Driven Analysis: Global Warming and EcoTrends", August 25, 2009, http://www.scivee.tv/node/12384

[9] Hahsler, M., and S. Koch, "Discussion of a large-scale open source data collection methodology," HICSS 38, 2005, pg. 197b.

[10] Edwards, P. N., M. S. Mayernik, A. L. Batcheller, G. C. Bowker, and C. L. Borgman, "Science friction: Data, metadata,

and collaboration," Social Studies of Science, vol. 41, no. 5, 2011, pp. 667-690.

[11] Shu, Y., K. Taylor, P. Hapuarachchi, and C. Peters, "Modelling provenance in hydrologic science: a case study on streamflow forecasting," Journal of Hydroinformatics, vol. 14, no. 4, 2012, pp. 944-959.

[12] Queenborough, S. A., I. R. Cooke, and M. P. Schildhauer. "Do we need an EcoBank? The ecology of data-sharing," Bulletin of the British Ecological Society, vol. 41, no. 3, 2010, pp. 33-35.

[13] Michener, W. K., S. Allard, A. Budden, R. B. Cook, K. Douglass, M. Frame, S. Kelling, R. Koskela, C. Tenopir, and D. Vieglais, D. A. "Participatory design of DataONE—enabling cyberinfrastructure for the biological and environmental sciences," Ecological Informatics, vol. 11, 2012, pp. 5-15.

[14] Haynes, K. J., O. N. Bjørnstad, A. J. Allstadt, and A. M. Liebhold, "Geographical variation in the spatial synchrony of a forest-defoliating insect: isolation of environmental and spatial drivers," Proceedings of the Royal Society B: Biological Sciences, vol. 280, no. 1753, 2013.

[15] Gil, M. A., "Unity through nonlinearity: A unimodal coral-nutrient interaction" Ecology, (in press), 2013.

[16] Lindenmayer, D. B., G. E. Likens, A. Andersen, D. Bowman, C. M. Bull, E. Burns, C. R. Dickman, A. A. Hoffmann, D. A. Keith, M. J. Liddell, A. J. Lowe, D. J. Metcalfe, S. R. Phinn, J. Russell-Smith, N. Thurgate, and G. M. Wardle, "Value of long-term ecological studies," Austral Ecology vol. 37, no. 7, 2012, pp. 745-757.

[17] Branch, B. D., C. Tenopir, S. Allard, K. Douglas, L. Wu, L., and M. Frame, "DataONE: Survey of Earth Scientists, To Share or Not to Share Data," in AGU Fall Meeting Abstracts, vol. 1, 2010, p. 1062.

[18] DataONE, "DataONE Usability and Assessment Working Group Charter," 2011, http://www.dataone.org/sites/all/documents/UandA_Charter.pdf

[19] Von Bertalanffy, L., General system theory: Foundations, development, applications, George Braziller, New York, 1968, pg. 289.

[20] el Sawy, O. A., A. Malhotra, Y. Park, and P. A. Pavlou, "Research Commentary—Seeking the Configurations of Digital Ecodynamics: It Takes Three to Tango," Information Systems Research, vol. 21, no. 4, 2010, pp. 835-848.

[21] Hochachka, W. M., R. Caruana, D. Fink, A. R. T. Munson, M. Riedewald, D. Sorokina, and S. Kelling, "Data-Mining Discovery of Pattern and Process in Ecological Systems," The Journal of Wildlife Management, vol. 71, no. 7, 2007, pp. 2427-2437.

[22] Riggins, S. H., "The rhetoric of othering," in The language and politics of exclusion: Others in discourse," Sage, Thousand Oaks, 1997, pp. 1-30.

[23] Michener, W. K., J. W. Brunt, J. J. Helly, T. B. Kirchner, and S. G. Stafford, "Nongeospatial metadata for the ecological sciences," Ecological Applications, vol. 7, no. 1, 1997, pp. 330-342.

[24] Brundtland, G. H., Our Common Future, Oxford University Press, Oxford, 1987.

[25] Chen, A. J., M. C. Boudreau, and R. T. Watson, "Information systems and ecological sustainability," Journal of Systems and Information Technology, vol. 10, no. 3, 2008, pp. 186-201.

[26] Goulder, L. H., and D. Kennedy, D. Valuing ecosystem services: philosophical bases and empirical methods. Nature's services: societal dependence on natural ecosystems, Island Press, Washington DC, 1997, pp. 23-48.

[27] Dong, S., S. X. Xu, and K. X. Zhu, "Information Technology in Supply Chains: The Value of IT-Enabled Resources Under Competition," Information Systems Research vol. 20, no. 1, 2009, pp. 18–32.

[28] ICSU, "Earth System Science for Global Sustainability: The Grand Challenges," International Council for Science, Paris, 2010.

[29] GEO, "The Global Earth Observation System of Systems (GEOSS) 10-Year Implementation Plan," 2005, http://earthobservations.org/docs/10-Year%20Implementation%20Plan.pdf

[30] LTER, "Network Overview", n.d., http://lternet.edu/network/

[31] Carpenter, S. R., S. W. Chisholm, C. J. Krebs, D. W. Schindler, and R. F. Wright, "Ecosystem experiments," Science, vol. 269, no. 5222, 1995, pp. 324-324.

[32] Daily, G. C., "Nature's services: Societal dependence on natural ecosystems," Island Pr. Washington DC, 1997, pp. 4-6.

[33] Gosz, J. R., R. B. Waide, and J. J. Magnuson, "Twenty-eight years of the US-LTER program: experience, results, and research questions," in Long-Term Ecological Research, Springer Netherlands, pp. 59-74, 2010.

[34] Yang, Q. and X. Wu, "10 challenging problems in data mining research," International Journal of Information Technology and Decision Making, vol. 5, no. 4, 2006, pp. 597-604.

[35] Sorenson, O., J. W. Rivkin, and L. Fleming, "Complexity, networks and knowledge flow," Research Policy, vol. 35, no. 7, 2006, pp. 994-1017.

[36] Lipparini, A., G. Lorenzoni, and S. Ferriani, "From core to periphery and back: A study on the deliberate shaping of knowledge flows in interfirm dyads and networks," Strategic Management Journal, 2013, DOI: 10.1002/smj.2110

[37] Porter, M. E., "Clusters and the new economics of competition," Harvard Business Review, vol. 76, no. 6, 1998, pp. 77-90.

[38] Liyanage, S., "Breeding innovation clusters through collaborative research networks," Technovation, vol. 15, no. 9, 1995, pp. 553-567.

[39] Li, B., "Searching for theory in metadata." iConference 2013 Proceedings, 2013, pp. 377-388.

[40] Li, B., "Toward an infrastructural approach to understanding participation in virtual communities". In H. Li (ed.) Virtual community participation and Motivation: Cross Disciplinary Theories, pp. 103-123. IGI Global. Hershey, PA, 2013.