Metadata Standards

Trajectories and Enactment in the Life of an Ontology

In S. L. Star & M. Lampland (Eds.), Formalizing Practices: Reckoning with Standards, Numbers and Models in Science and Everyday Life
(accepted 2006; forthcoming 2007)

Florence Millerand and Geoffrey C. Bowker
Université de Montréal and Santa Clara University

Introduction

There has been an explosion of scientific data over the past twenty years, as more and more sciences deploy remote sensing technologies and data-intensive techniques such as MRIs. Just as we are increasingly monitoring and surveilling each other, we are increasingly tracking the processes of environmental change on every scale from the cell to the ecosystem. Further – and again the parallel with our forms of social control is uncanny[1] – we are with the development of cyberinfrastructure (or, in its kinder, European coinage 'e-science') working through the possibilities of working with data collected in multiple, heterogeneous settings and using sophisticated computational techniques to ask questions across these settings. We cannot answer, it is claimed, the big questions of the day without this facility – data from one discipline cannot build picture of species loss which can inform policy, just as data from one source cannot profile a population in order to discipline it.

The question is posed, therefore of the preservation, access and sharing of scientific data[2]. In the traditional model of scientific research, data is wrapped into a paper which produces a generalizable truth – after which the scaffolding can be kicked away and the timeless truth can stand on its own. There has been relatively little active holding of very long term datasets and little data reuse. In the current context, to the contrary, there is an emphasis on what is oxymoronically called 'raw' data – which can be gathered together, analyzed, visualized and theorized to produce new syntheses. This is particularly so for the case of environmental data[3], where theories need to range over multiple temporal and spatial scales. It has been traditional in ecology for individuals or small groups to collect data in short term projects (the length of a funding cycle) over small areas (one square meter)[4]; this is no longer sufficient to the task[5]. Ecosystems change in larger chunks of time (indeed, they follow multiple complex rhythms) and over wider areas of space than traditionally conceived[6]. Thus researchers need to be able to use datasets constructed by others, for different purposes; and they need to be able not only to reach some kind of ontological accord between the disciplines (allowing kinds and classifications to be shared) but also to be able to trust data produced by others – the traditional 'invisible college'[7] becomes a teeming city with multiple linguistic communities.

One of the major challenges for the development of a scientific cyberinfrastructure aiming to foster collaboration and data sharing through information networks is to ensure the frictionless circulation of data across diverse technical platforms, organizational environments, disciplines and institutions. Or, to use the term of the art, to ensure 'interoperability'. A central problem here is that the storage, access to and evaluation of the validity of data are extremely dependent on the ways in which the data has been collected, labeled and stored. While it may be possible for two colleagues in a discipline to share information about their data with a simple longhand note, there is unquestionably a need for more documentation in the case of pluridisciplinary teams working over multiple sites and scales. To deal with long-term questions, for example, a given dataset may have been collected in one context using a home-grown set of protocols, often deploying outdated instruments and terminologies. The task of making that data available across disciplines and over time is in general an unfunded mandate – it requires a special kind of altruism to carefully code your data in ways beyond what you need for its immediate use. It is easy to see, therefore, why

assorted technofixes are being discussed, debated and to some extent deployed in order to address this problem set.

The resultant standards – most often conceived as being simple technical solutions – are being developed to permit the interconnection of systems and thence the free flow of data. The capacity for distributed, collective scientific work practice is posited on the existence of shared information infrastructures and collaborative platforms. These in turn require some base of shared standards. Although they are largely ignored and invisible (buried in an infrastructure, wrapped in a black box) these standards nonetheless constitute the necessary base for distributed cognitive work. In order to understand the modalities of collaboration in collective work – scientific and other – we need to understand standards. In particular we need to understand the forms and functions of 'metadata' – data about data – standards.[8]

Here we examine an infrastructure development project for an ecological research community – the American network for long-term ecological research (LTER) – which is endeavoring to standardize its data management through the adoption of a shared metadata standard called the Ecological Metadata Language (EML). This is one of a suite of XML's – there is VRML (Virtual Reality Markup Language) and even another EML (Educational Modeling Language). This standardization process began in 1996, at the level of the LTER network. It crystallized in 2001 with the adoption of the EML standard by the community. It has since been the subject of a controversy which can be characterized as "mission successful" (by completing the implementation cycle) or "succes to come" (by staying the course).

It is precisely these divergent visions which are the object of this paper. We do not so much want to know the success conditions for the implementation of an information management standard in an organization, so much as to know from what *time* and according to which *point of view* the success or failure of the implementation are judged. We will draw on an ethnographic study of the community to explore the alignment of diverse trajectories[9] in standards development. We will explore this process as one of the enactment of a standard.

1. The development of information infrastructures at the intersection of social worlds

As for scientific activity in general, the development of information infrastructures for the sciences requires the cooperation of a heterogeneous set of actors – in this case domain experts, IT specialists, informatics researchers and funding agencies. There is no linear narrative to be told: "The time of innovations depends on the geometry of the actors, not on the calendar"[10]. In other words, we cannot track a single life cycle (development, deployment, death) but must pay attention to the diverse temporalities of the actors. This perspective allows us to better grasp how the existence and even the reality of projects varies over time, in line with the engagement or disengagement of actors in the development of these projects or objects. Thus although a technical object may exist in prototype form, it can be considered more or less real only to the extent that certain groups of actors rally or not to its cause[11].

What interests us here is the point of contact between the different trajectories of the human actors as well as the non-human actors in the process of standardization at

work in the heart of the LTER research community. What happens in this process which leads one group of actors to formulate an alternative history to the 'success story'? Which trajectories interact with each other, and how do they adjust accordingly? How are certain trajectories redefined?

An immediate problem is to know which trajectories to follow,[12] since the choice of following any one in particular over another can lead to a different understanding of the social, technical and organizational configuration of the study. Infrastructure studies is a useful source here, for it shines analytical light on rarely-studied phenomena – such as the 'invisible' work carried out in the background by actors whose performance is considered so much the better to the extent that it is self-effacing, invisible.[13]

We conceptualize the EML standard as a support for the coordination of different social worlds - domain researchers, standards development teams, information managers concerned with its implementation. EML *a priori* as a solution to a set of technical problems - a solution from which will issue the one good tool which can be used by all. However, this technical standard and its implementation process in fact speak directly to the organization of scientific work: it assumes specific configurations of actors, tools and data. In order to explore this dimension, we shall deploy the concept of enactment, developed by Karl Weick.[14] In this tradition, Jane Fountain invites us to distinguish between an 'objective' technology – that is to say a set of technical, material and computing components such as the Internet – and an 'enacted' technology – that is to say the technology on the ground as it is perceived, conceived and used in practice, in a particular context. In this view, the way in which actors enact technical configurations such as standards depends directly on their imbrication in cognitive, social, cultural and institutional structures. Organizational arrangements (characterized by routines, standards, norms, politics) mediate the enactment of technologies, which in return contribute to the refashioning of these arrangements.

We propose, therefore, to examine enactment in action – to trace two sets of histories of a single process of standardization, through restoring the artifacts, actors and narratives to the context whence they emerged. This perspective permits a better understanding of the social and organizational dynamics at the heart of projects for the development of large scale information infrastructures.

2. The LTER Research Community and the EML Standard

The LTER program constitutes a distributed, heterogeneous network of more than 1800 research scientists and students. Formed in 1980, the network currently consists of 26 sites or research stations[15] (ironically, some 'long term' sites have already closed; and more have been added). Each is arranged around a particular biome – for example a hot desert region, a coastal estuary, a temperate pine forest or an Arctic tundra – in the continental United States and Antarctica. A 27th site is charged with the administration and coordination of the group. The program's mission is to further understanding of environmental change through interdisciplinary collaboration and long term research projects.

One of the chief challenges of LTER is to move beyond the 'plot' of traditional ecoscience to analyze change at the scale of a continent and beyond the six year funding cycle or 30 year career cycle of the scientist to create baselines of data spanning multiple decades. While the preservation of data over time, and their storing in conditions

appropriate to their present and future use, has always been a priority within the different sites of the LTER network, there has been a new urgency with the development of a cyberinfrastructure projects aiming to encourage data sharing across the community.

Each of the 26 sites in the network takes responsibility for the management of research data produced locally; and each in general has its own information system (its own databases). An information manager is charged with the development and maintenance of local infrastructures. Across the network, then, data are stored autonomously by the different sites – a fact which renders the search for and access to data relatively complex and laborious (which in turn naturally militates against the network realizing its mission). Accordingly, a project was put into place in 1996 to initiate a networked information infrastructure permitting the federation of the local databases and thus data exchange.

The project has encountered three major challenges: the heterogeneity of the data which circulates through the research community; their wide dispersal; and the multiple systems of coding and storing.[16] Beyond the diversity of data attached to a given scientific project, there can also be an extreme disparity in their organization and formatting, depending on the collection protocols adopted. For example, data to measure the amount of chlorophyll present in a sample of sea water will be organized into separate files corresponding to the number of trips made; whereas the same measures taken over a year in a given lake will be held in one single file. Further, local cataloguing cultures generally use information (or metadata) which is not necessarily understandable outside of a given research project, site or discipline. Thus 'special' (personalized) measurement units can be created for the analytic needs of a particular research project (for example the ungeneralized unit: "number of leaves per change of height in a plant"). In this context, the LTER network office soon saw the need for a set of data standards. Or, to be more precise, the need for standardized methods for the collection of metadata has been taken as the preferred solution for data interoperability.

In an ideal world, the metadata contain all the details necessary for all possible secondary users of a dataset (an ideal solution which evokes Spinoza's problem – to know a single fact about the world, you need to know every fact about the world). These include detailed and diverse information, such as the names of the researchers who collected the data, the title of the project they were working on, the project summary, key words, the type of biome, sampling techniques, the calibration of the measuring tools at the time of data collection. By extension, the possibility of complex analyses drawing on physical, chemical, biological data within a given geographical area will depend on the quality of the metadata. (We should note at this point that 'metadata' is not necessarily the only possible solution; however it has presented itself – organizationally and intellectually – as central to this community.)

Although they appear to be mere technicalities to some, metadata take on a central importance in the production of scientific theories in the degree to which they condition access to data, guarantee their integrity and delimit their interpretative uses. When you are dealing with comparative or long term studies, you find that material conditions change over time and space – instruments might become more accurate, for example, which indicates the need for precise documentation of their calibration. When combining data across the disciplines, metadata do more than provide a convenient label: they structure the conversation which ensues. An analogy here is with normal language use.

There are communities for whom 'casualty' means 'someone injured or killed' and others for whom it is simply an euphemism for 'fatality'. Or there are some for whom 'democracy' means 'rule of the people' and others for whom it is a euphemism for 'capitalism'. Unless you can calibrate across the communities (and it is a difficult act of imagination often to recognize how local and specific your usage is) then you cannot communicate.

The EML metadata description language is precisely a standardized language for the generation of metadata in the specific domain of the sciences of the environment – it is much better to use science in the plural here, since each domain has its own configuration of classifications, instruments, dates and places. EML is the standard which the LTER research community has decided to adopt when it engaged in the process of standardizing its scientific data management practices.

3. **Controversies**

The process of standardization which the LTER community is engaged in has two major objectives: the promotion of interdisciplinary collaboration through data sharing and the improvement of long term data preservation.[17] Although both these objectives and the deployment of EML to their end are generally agreed, conflicting voices could be heard at the moment of deployment.

Here, we examine two narratives, from two categories of actors, which tell radically different tales about the EML standard as successfully implemented or still a work in progress. The first comes from the standard's developers, which includes the experts who wrote the specifications for EML together with the coordinators of the LTER network. The second presents the point of view of those enacting the standard – that is to say the information managers whose task it is to implement it at a given site. At the time of this study, the 'success' narrative was carrying the day – it was already formalized, written up in reports; whereas the latter was diffuse and oral. (If in the context of policy work in John King's dictum, "numbers beat no numbers every time" – then in the context of computer science funding, written beats oral every time).

With respect to methods, interviews were conducted with both groups of actors, and participant observations were carried on throughout the standard implementation at two of the LTER sites. Detailed document analysis were systematically performed.

3.1 Narrative 1: "EML is a success: the entire LTER community has adopted it"

Version 1.0 of EML first saw the light of day in 1997 at the National Center for Ecological Analysis and Synthesis (NCEAS) in Santa Barbara. It was the product of a researcher in ecological informatics working with two doctoral students. EML responded in the first instance to internal preoccupations within NCEAS, which since its creation in 1995 has been addressing the absence of tools and techniques for analyzing and synthesizing environmental data. A grant was submitted to the National Science Foundation, which funded its development.

Technically, EML is based on two emergent standards (SGML and XML) which have, *grosso modo* been developed in order to turn networked information from simple text fields into searchable, combinable databases. Its content is drawn from the main data description types in use in the domain – such as those recognized by the Ecological Society of America, itself a pioneer in preserving datasets alongside of papers. Versions

1.0 to 1.4 cascaded out between 1997 and 1999. They were tested within NCEAS. Given the difficulties encountered in use, a major revision of the language was suggested (would that one could vary natural languages so simply) – and a second grant proposal was written and subsequently funded. The development team went from being three people to a collaboratory[18] - a collaborative platform based on voluntary participation and open to the whole community of environmental scientists. This open development model was not immediately successful, even if the team was able to attract some more developers, including – for the first time – a separately but synergistically funded information manager from the LTER. The development of EML went on apace, with several significant structural changes being made. Seventeen versions were produced between 1999 and 2002.

In 2001, the team reckoned that it had produced a stable version of EML (version 1.9, which was according anointed with the title EML2.0 beta). The team decided to present their product at the annual conference of LTER information managers, held that year in Madison. Discussion was lively but responses were extremely favorable – the information managers recognized the usefulness of such a standard for the LTER community and were moved to formally adopt EML. Version 2.0 was put into circulation; the LTER network scientific community (one of the most important communities in environmental research) adopted the standard. In short, the EML project was a resounding success.

3.2 Narrative 2: "EML is not (yet) a success: it needs to be developed further before it can be used"

At the period of the creation of EML in 1997, the LTER network sites already had in place systems for managing their scientific data. Depending on the site, these systems were more or less formalized – that is to say that they did not necessarily use the same standardized vocabularies, even if some of them broadly speaking used the data descriptors recognized by the American Society of Ecology – leading to the standard problem of almost compatibility. In 1996, the inauguration of a project to develop a network wide information system stimulated discussions about standardizing data management procedures and encouraged the development of a common tool set for the information managers of the community. However, there was still no central initiative covering the whole network. In 2001, the EML project received a favorable reception from the information managers, who made a consensus decision to adopt it. The implementation began.

While some sites began the work of implementing the standard relatively quickly, most of them ran into significant problems. The standard is complex and it is difficult to understand in its entirety. The technical tools intended to facilitate the standard's implementation proved unusable, i.e. incompatible with existing local practices and infrastructures. And in general there was just a huge amount of work to be done (on top of the normal workload) with minimal resources – some sites had to undertake a complete restructuring of their data management practices.

Numerous *ad hoc* solutions were brought to bear – for example, home-grown tools that some of the information managers shared amongst themselves to facilitate the work of the conversion of local systems into the format required by the EML standard.

The information managers organized two workshops in which the developpers participated devoted to the implementation of EML in 2003 and 2004. These led to the production of a synthetic 'best practices' document for EML implementation. These had a material impact on implementation at a number of sites. This in turn led to a five step implementation plan formulated conjointly by the information managers and the LTER network coordinators.

At the annual conference of information managers in 2005 at Montreal, progress was seen as somewhat mixed – EML implementation was seen as a complex and laborious process whose outcomes in terms of improvement of data management remained somewhat difficult to identify. EML was not yet a success – it had to be partially refactored in order to be usable.

4. Trajectories

On the one hand, then, we have a success story about the EML standard, highlighting its consensus adoption by the LTER research community. On the other hand, we have a very mixed picture – varying widely by site. While the first story moved into the 'happily ever after' phase in 2001, the second had barely gotten beyond 'once upon a time'.

A simplistic reading of these two narratives would say that the measure of success of a standardization process differs as a function of the different phases one is looking at (here the phases of conception and development of the standard in contrast to that of deployment and implementation). In other words, one could say that the information managers couldn't yet recognize the projects success at this point in time – they would only be able to see it once the project was finally completed. This evolutionist reading of technology development projects developing in an objective time frame really does not advance our understanding of what really happens during the periods of emergence, development, maturation, implementation and so forth. Further, it continues to privilege the second story over the first – considering the success of the standard as being always already assured, with full confirmation coming in the natural course of events. Thus it favors the invention of the standard over its innovation[19] - its deployment and enactment[20].

A temporal analysis of technology development projects should seek rather to account for their evolution in terms of the multiple temporalities into which they are integrated. It would then become possible to account, from the point of view of the actors, for the whole set of events – including the more troubled periods when folks do not want to talk about it (it does not sound good, for example, in the next funding application) while others still seek to find a voice (for example, because they are too low status to be heard; or if heard, they are not using a technical language the developers understand).

4.1 Multiple Trajectories

It is striking the degree to which all of the actors involved in the standardization process (EML developers, LTER network coordinators, information managers, domain researchers …) have supported – and continue to support – the "EML project". As mentioned before, neither metadata nor the EML standard were the only possible solution to ensure data interoperability through the network. Nevertheless, they all believe in the

idea of a metadata standard permitting the exchange and sharing of data throughout the LTER network and beyond. In this sense, it's not the case of the imposition of a standard by one group of actors (developers and coordinators) on a hostile, resistant group (information managers or domain researchers). The latter have always been highly supportive of the project, up to including status of EML implementation in the information management review's criteria for the sites.[21] It as at the moment of the actual implementation of the standard at a give site when critical problems emerge and discordant voices can be heard.

The recognition of these difficulties and the controversy which has ensued have contributed to bringing the status of EML as a useable standard into doubt. Two years after its adoption, an inquiry revealed that it had not yet been completely implemented in a single one of the 26 network sites, and that the tools developed explicitly for this purposes remained largely unused. EML seemed to be a standard in name only.

The juxtaposition of the two narratives above reveals the confrontation of two visions of the EML standard. An imaginary dialogue, inspired by Latour's history of the Aramis project[22] reveals the gulf between the two sets of actors:

> "EML 2.0 exists – the bulk of the work has been done, all we need to do is implement it" (the developers)
>
> "All we need to do …!? But a metadata standard is just a language. No matter how perfect it is, it only exists if it's being used – if it serves above all to integrate data" (information managers).

In other words, in 2001 the EML standard was a metadata standard without data.

We propose to read these differing perspectives on EML by restoring them to the trajectories of the actors concerned. Thus, from the point of its developers, the EML standard was above all one of research and development. The project goal was the creation of a standardized description language for metadata rather than its materialization. Its ambition was to make itself the reference standard in environmental sciences. From this perspective, the standard's development and its adoption by the wider research community of environmental scientists constitute the main success criteria for the project. From the point of view of the information managers, the EML standard represented a set of tools and practices for the better management of scientific data – notably by improving the quality of metadata produced within each of the sites. By this view, the successful incorporation of this new tool – and the new modes of practice which accompany it – within local sociotechnical infrastructures constitute the major success criteria for the EML project. Finally, from the point of view of the scientists belonging to the LTER research community, the EML standard is a technical tool which opens the door to multisite research endeavors through a better form of access to and sharing of data and which promises a better diffusion of data beyond the LTER network. From this perspective, the capacity to carry out multidisciplinary projects in very large datasets through a single interface constitutes the main criterion for the project's success.

4.2 Trajectory Alignment

We read the implementation work carried out by the information managers as a process of appropriation[23] of the standard in the course of which the work of trajectory alignment is done. The appropriation of the EML standard in the different sites worked out in effect as the adjustment of the technical tool to local contexts, and the adaptation of

pre-existing practices to new ways of working. Concretely, this entailed a real work of bricolage from the information managers seeking to incorporate a (generic) standard into a (local) context, which both gave it its purpose and permitted its use.

The trajectory of EML according to the first narrative was born of the main descriptors of ecological data in use in the domain, became a research and development project at NCEAS and then the metadata standard of reference for environmental sciences. It seemed to take a turn or a certain reorientation as it began to circulate in the LTER network. From that moment, the description of the EML project as one of conception, development, deployment and implementation ceased to work. In the implementation phase there was re-development work, which led to a reconsideration of the conceptual basis of the work and then some more re-development for re-implementation and so forth. The EML project was changing – the set of trajectories had to be realigned.

What happened then in this implementation phase of the standard which necessitated more … implementation work? The information managers spontaneously responded that there was a lack of tools permitting the conversion of local metadata systems into the EML format. Equally, they complained that there was a weak understanding of real implementation processes from the network coordinators, who seemed to them to have unrealistic expectations. The problem was that data management practices are not solely dependent on the types of technical infrastructure: they are also and above all intimately linked to the nature of the research projects being studies, to the disciplinary and organizational cultures of the sites – in short to the local structure of scientific work.

The following two interview extracts illustrate on the one hand the local and contingent nature of the scientific work being done and on the other the complexity of the information managers' task of cataloguing research data:

(1) I was getting nutrient data and my units came in as micromoles with the micron symbol and capital M, micromoles. When I started having to go into EML, which does not have that unit, I had to figure out well what actually is this unit. And in digging deeper and going to our lab that processed these data I found out its not micromoles its micromoles/liter. And I am not a chemist so it just didn't mean anything to me. You know I am just organizing and posting this type of data, and so it really opened my eyes that I have a bigger issue here (35.26) than I thought you know because here we've got people reporting things as micromoles which is not proper. But that is just the way the work is done, and shared, and no one ever questioned it. That's kind of interesting. So I started dataset by dataset trying to retrofit everything back into you know EML. And I have this ongoing list of these custom units that I am compiling, making my best guess at and then I am going either to the actual you now my PI or a collaborator that gave me those data, and having to sit down with them to say can you please verify, if you were going to describe this unit in EML as a custom unit does this make sense. Are you reporting it the proper way. Are you calling this the attribute what it would universally be called, that kind of thing... (IM_L.)

This first extract provides an example of the locally situated work that doesn't yet scale as a joint measurement unit within the framework of the shared conventions of a community of practice – in this case the LTER network. The 'retrofitting' referred to is the occasion for a lot of cyberinfrastructure disasters – data that was understood well enough by a local group often has to be completely revisited for a wider community. Designers persist in seeing the conversion task as easy, but this is only on the assumption that the data being fed in is clean and consistent.

> (2) Micromoles Per Liter and Micromolar are measurement units for concentration. Technically, both are micromoles per liter, and so equivalent in magnitude. [But] their scopes are different, because micromoles per liter can be used for a particulate or dissolved constituent, and micromolar is correctly used only for dissolved. So they are not exactly interchangeable. Micromoles Per Liter and Millimoles Per Cubic Meter are equivalent in magnitude, but different disciplines have preferences for one or the other. [Also], if you happen to be in open ocean, you would run into micromoles per kilogram and micromoles per cubic meter, which are similarly equivalent only at sea level, because interconversion depends on pressure… (IM_M.)

This second extract gives an example of measurement units which are *a priori* identical but which mean different things in different disciplinary environments.

Taking a step back, the general problem can be characterized thusly. In the grand old days of the nineteenth through early twentieth centuries, when scientific certainty was at its zenith, it seemed as if there were a clear and consistent classification of and hierarchy between the sciences. The most famous example is Auguste Comte's classification of science into a classificatory tree going from mathematics through physics and chemistry down (or up, depending on your inclination) to sociology. Each part of each discipline was divided into statics and dynamics. This was also the period of the discovery of the principle (not, be it noted the fact) of division of labor: Charles Babbage mirrored his computer on factory production techniques – making a complex task easy by splitting it into a set of serial subtasks. Together, classificatory principle with the division of labor created a picture of scientists as workers in a giant collective enterprise – in Poincaré's terms, workers lay bricks in the cathedral of science. Some indeed saw the end of the period of 'heroic science' as the principle of the division of labor emerged.

We are running nowadays into the question of whether cathedrals need blueprints.[24] What could possibly guarantee that all of these bricks would fit together into a seamless whole? There are two options – in close parallel with 'blind watchmaker' positions. Either there was a higher entity (the universal scientific method and the positivist classification of science in this case) ensuring that they all fit; or there was a constant, contingent, local process of partial fitting and constant disordering which could still, in the long term, guarantee that at any one 'join' there was a fit (though there could not possibly be across the set of joins). Both scientists and information systems designers have been working largely from the former assumption. When they face the reality of the latter they are constantly surprised – they have not been following the scientific method as faithfully as they thought (their databases are dirty, units are ambiguous) and it does

not all fit into clearly designated, separable chunks (two disciplines might both claim 'control' over the same measurement unit).

4.3 The Case of the Dictionary as Articulation Strategy

At the start of 2005, with the tools created by the developers not yet being used by information managers and the implementation dragging on, the network coordinators began development of a new tool intended to accelerate implementation of the standard among the sites. In parallel – and partly in reaction to this project – some information managers initiated the development of a 'house' tool: a repertoire of measurement units.

One of the principal difficulties which the information managers faced was tied to the complexity of the work of translating their metadata into the EML language – notably with respect to measurement units. On the one hand, the dictionary of measurement units which came with the EML standard essentially catalogued physical measurement units of ecological phenomena – while most LTER network sites were using biological and ecological measurement units. On the other hand, it is extremely difficult to describe in a standardized language 'special' or personalized measurement units – that is to say units developed for some specific purpose in a research project, and which only really make sense in the context of that project.

Faced with these difficulties, some information managers then began to exchange lists of measurement units (including local ones) used at their site, so as to compare their respective translations and to catch any inconsistencies. This quickly evolved into a project to transform these lists into an LTER-wide catalogue of units. The plan was to produce a dynamic, online tool available through the LTER intranet. The team, which until then had been made up solely of information managers, expanded to include a representative from the network office. They developed a prototype integrating the unit lists of six sites. This was presented in August 2005 to the annual conference of LTER information managers. It was represented both as an implementation aid for the EML standard and as an example of a successful collaboration between information managers and developers/coordinators.

Technically, this tool provided the information managers in the network access to definitions of measurement units in EML (including some specialized units), to propose corrections to the standard's unit dictionary, and to add definitions of other units. However, it did considerably more than facilitate conversion from one format to the next – it was above all a work of social coordination.

These local *ad hoc* initiatives and the network-wide projects can be seen as strategies for re-articulating the work – on the one hand between the information managers themselves and on the other between them and the developers/coordinators. The information managers did not need the tool of itself as much as they needed all this work of information mediation which the enactment of the standard brought to bear. How best to describe such and such a measurement unit? Is a given measure a local or a network one? Can this unit be added to the EML dictionary? The work of producing the dictionary of units became a tool for facilitating coordination and cooperation between different worlds. In other words, by creating a tool that could be used locally at each site and yet contribute to the improvement of the standard, the information managers created a boundary object[25] which capable of supporting this articulation work between enactors and developers.

5. Enactment

We propose using the concept of enactment to better understand the complexity of the work of the information managers. These latter not only ensured the implantation of the EML standard and the programming of some additional functionalities which could make it operational (its implementation) – they also worked on its interpretation. That is to say, they worked on its staging (in the theatrical sense of the term), which involved co-adapting the standard and local work practices. These mutual adaptations are better seen not as local resistances but as necessary adjustments without which the EML standard could not operate within the LTER community.

In what ways did the EML standard contribute to changes in the social worlds of the actors? And how did these worlds work to change the standards? These changes involved first the identities of the actors and their organizational structures and second the 'script' of the technical tool.

5.1 New Organizational Roles and Structures

Throughout infrastructuring in general and this standardization process in particular, the information managers as a community of practice became 'visible' within the LTER network[26]. In the same way, certain aspects of their work which until then had been little known and recognized became explicit. In 2004, when there was an efflorescence of house tools in particular sites and the information managers were being integrated with the developers, they achieved 'official' status as developers – that is to say they are on the list of credits attached to the standard's documentation. The complexity of their task (taking the local and rendering it into a universal language – a task that even the engineers of Babel found daunting – was recognized.

That said, even if the transformation of organizational models within the LTER community forced a reorganization of working patterns (from site-based to a federated structure), and even if the role of information managers was considerably transformed, their status as technicians whose task is to provide support and maintenance remains dominant in the network – notably among the domain scientists. Further, even if the team of experts recognized their contribution as developers of the standard, their contribution remains ambiguous to the extent that developers retain the tendency to see initiatives from the information managers as too local and not state of the art. Thus the synthetic 'best practices' document produced by the information managers was judged to have too many signs of its origin within the LTER community to be integrated into the standard's documentation. (And, one hesitates to say 'of course', of course, but as a matter of course the social and organizational innovation was similarly not included in the standard's documentation).

It remains the case that the set of actions carried out by the information managers during this standardization process revealed that another organizational configuration was possible – if only at the very basic level of resource allocation. If the development of a metadata standard for a research community like the LTER requires significant funding then so *a forteriori* does its enactment within a given setting. More concretely, the information managers contributed to putting into place a new representative structure – in this case a permanent committee formed equally of information managers and domain experts, whose mission is to ensure a representative and advisory role in the development

of integrated network information management practices: the Network Information Advisory Committee (NISAC). This committee came into being one year after the adoption of the EML standard (a somewhat lengthy gestation period!), when the initial sets of difficulties incited the information managers to initiate a dialogue with domain scientists. Out of this committee came the plan to implement the standard in different stages.

Beyond this new organizational structure, a new form of collaborative work at the intersection of local, site-based work and global (network) activity was experimented with successfully. The project of building a dictionary of measurement units constituted a veritable innovation (from below) within the LTER community, to the extent that one the one hand it opened the way for the transformation of a local initiative into a project for participatory design at the network level and on the other created a new collaboration space between two groups of actors who had not been directly associated before (information managers and developers/coordinators).

5.2 Redefinition of the Standard

If the standard itself has been the object of multiple versions over the course of its development – in part as a result of the new members integrated into the team – what was presented in 2001 to the LTER information managers was a black box – a final version. The box was reopened, and not always in the same way, across the set of sites.

Thus as we have seen certain lacunae in the standard were identified and the implementation plan itself changed to accommodate the different rhythms of integration of the different sites. More generally, the kinds of difficulties which come into play when you try to enact a standard generic enough that it could in principle apply to any kind of data.[27] The EML project included a definition of a certain role for researchers – that of describing their data in this new language.[28] Indeed, researchers have always been envisaged by the developers as future users of the standard. They both describe their data collections in EML terms and can carry out complex, integrative studies using vast dataset as a result of these standardized descriptions. It is interesting to note that the role of the information managers has never been mentioned (at least explicitly) in scenarios of EML use.

And yet, in practice, a number LTER researchers refused their roles – principally through lack of time (standards tend to be an unfunded mandate) and interest. It should be recognized that implementation of the standard can double the amount of work they need to do to enter their data. Moreover, the investment in time to learn the EML language (without mentioning that of learning the conversion tools) has constituted a point of no return for many. The information managers has taken on this role which was *a priori* destined for others.

One can certainly read this redistribution of roles as a coming from a transitional period – and thus imagine that the LTER network researchers will get up to speed with EML to the extent that the tools become easier to use and the standard is taken up within environmental science generally. However, the question of training researchers to use EML (and more widely to understand the new forms of information management associated with an integrated infrastructure) is at present hanging – and it seems likely that the information managers will continue to pick up the slack.

6. Conclusion

In one sense, then, the EML standard did not change anything. The division of labor remains the same (information managers are still in charge of the production of EML metadata); roles are stable (information managers contribute to the redevelopment of a standard for the LTER network, whilst developers work on the development of a standard for environmental science in general); local practices are confirmed (information managers share *ad hoc* solutions and in-house tools). And yet, the EML standard has changed the world. The actors' identities have changed (information managers recognized as developers and not merely implementers); new organizational structures are built (information managers are now represented in the NISAC committee); new forms of work are proposed (a collaborative space between the sites and the network has opened up).

The implementation of EML is not simply a case of upgrading an existing system. It consists above all in redefining the sociotechnical infrastructure which supports this articulation of technical, social and scientific practices. These redefinitions have significant consequences socially and organizationally. Because the tools are intimately imbricated in local work practices, and because the EML standard operates only within a given (social, technical, organizational) configuration, its enactment requires infrastructural changes.

This is why it does not make sense to see standards simply as things out there in the world which have a predetermined set of attributes. In information systems, standards are in constant flux – they have to migrate between communities and across platforms. Closure is a narrative which serves a purpose, not a fact which describes an event.

We slice the ontological pie the wrong way if we see software over here and organizational arrangements over there. Each standard in practice is made up of sets of technical specifications and organizational arrangements. As Latour has reminded is in another context, the question is how to distribute qualities between the two[29] – what needs to be specified technically and what can be solved organizationally is an open question, to which there is no one right answer. By assuming that specifications can exist outside of organizational contexts, we have already given the game away: leading to the continual need for the kind of innovation detailed in this chapter. And the innovation is always forgotten, since the same ontological mistake – made elsewhere, by other people - next time will again occasion its necessity. Indeed, a test for ontological errors in general is that one can say the same thing a hundred different ways over a span of years – there is no way in which the message can be heard *until* the organizational changes have taken place such that a reception is possible[30]. Both standards and ontologies (the one apparently technical and the realm of machines, the other apparently philosophical and the realm of ideas) need to be socially, organizationally bundled – not as a perpetual afterthought but as an integral necessity.

Acknowledgements:
The authors would like to thank our collaborators at UCSD, Karen Baker and David Ribes, both members of the Comparative Interoperability Project, SES0433369

Interoperability Strategies for Scientific Cyberinfrastructure: A Comparative Study, 2004-2007. For further information and publications visit*:* http://interoperability.ucsd.edu (there is a special prize for the millionth visitor). We also thank the LTER community participants for their interests and openess to this research.

---

[1] See Foucault, M. (1991). Governmentality. *The Foucault Effect: Studies in Governmentality*. G. Burchill, C. Gordon and P. Miller. Chicago, University of Chicago Press, Luke, T. W. (1999). Environmentality as Green Governmentality. Discourses of the Environment. E. Darier. Oxford, UK; Malden, Mass., Blackwell**:** 121-151.

[2] P. W. Arzberger, P. Schroeder, A. Beaulieu, G. C. Bowker, K. Casey, L. Laaksonen et al., « Promoting access to public research data for scientific, economic, and social developement », Data Science Journal, 3 (29), 2004, p. 135-152

[3] J. F. Franklin, C. S. Bledsoe & J. T. Callahan, « Contributions of the Long-Term Ecological Research program », BioScience, 40 (7), 1990, p. 509-515

[4] Lewontin, R. C. and R. C. Lewontin (2000). The triple helix: gene, organism, and environment. Cambridge, Mass., Harvard University Press.

[5] Bowker, G. C. (2006). Memory Practices in the Sciences. Cambridge, MA, MIT Press.

[6] O'Neill, R. V. (2001). "Is it Time to Bury the Ecosystem Concept? (With Full Military Honors, Of Course!)." Ecology **82**(12): 3275-3284.

[7] Crane, D. (1972). Invisible colleges; diffusion of knowledge in scientific communities. Chicago, University of Chicago Press.

[8] W. K. Michener, J. W. Brunt, J. J. Helly, T. B. Kirchner & S. G. Stafford, « Nongeospatial metadata for the ecological eciences », Ecological Applications, 7 (1), 1997, p. 330-342

[9] A. Strauss, Continual permutations of action, New York, Aldine de Gruyter, 1993

[10] B. Latour, Aramis ou l'amour des techniques, Paris, La Découverte, 1993, p. 80.

[11] Ibid.

[12] J. H. Fujimura, « On methods, ontologies, and representation in the sociology of science: where do we stand? » in D. R. Maines ed, Social organization and social process, New York, Aldine de Gruyter, 1991, p. 207-248 ; S. L. Star, « Power, technologies and the phenomenology of standards: on being allergic to onions », in J. Law, ed, A sociology of monsters? Power, technology and the modern world, London, Routledge, 1991, p. 27-57 ; S. Timmermans, , « Mutual tuning of multiple trajectories », Symbolic Interaction, 21 (4), 1998, p. 425-440.

[13] G. C. Bowker & S. L. Star, Sorting things out : classification and its consequences, Cambridge, MIT Press, 1999 ; S. L. Star & G. C. Bowker, « How to infrastructure », in L. A. Lievrouw & S. Livingstone eds, Handbook of new media. Social shaping and consequences of ICTs, London, Thousand Oaks, New Delhi, Sage Publications, 2001, p. 151-162 ; S. L. Star & K. Ruhleder, « Steps toward an ecology of infrastructure: design and access for large information spaces », Informations Systems Research, 7 (1), 1996, p. 111-134.

[14] K. E. Weick, The social psychology of organizing, Addison-Wesley, 1979.

[15] J. E. Hobbie, S. R. Carpenter, N. B. Grimm, J. R. Gosz, T. R. Seastedt, « The US Long-Term Ecological Research Program », BioScience 53(1), 2003, p. 21-32.

[16] M. B. Jones, C. Berkley, J. Bojilova & M. Schildhauer, « Managing scientific metadata », IEEE Internet Computing, 5 (5), 2001, p. 59-68.

[17] J. E. Hobbie, S. R. Carpenter, N. B. Grimm, J. R. Gosz & T. R. Seastedt, « The US Long Term Ecological Research program », BioScience, 53 (1), 2003, p. 21-32.

[18] G. M. Olson et al., « Technology to support distributed team science: The first phase of the Upper Atmospheric Research Collaboratory (UARC) », in G. M. Olson, T. Malone & J. Smith eds, Coordination Theory and Collaboration Technology, Hillsdale, Lawrence Erlbaum Associates, 2001, p. 761-783.

[19] J. Schumpeter, The theory of economic development, Cambridge, Harvard University Press, 1934.

[20] K. S. Baker & F. Millerand, « Articulation Work Supporting Information Infrastructure Design: Coordination, Categorization, and Assessment in Practice », in Proceedings of the Hawaii international conference on system sciences (HICSS'40), 2007.

[21] K. S. Baker & H. Karasti, The long term information management trajectory: working to support data, science and technology, San Diego: SIO Report, 2005 ; H. Karasti & K. S. Baker, « Infrastructuring for the long-term: ecological information management », in Proceedings of the Hawaii international conference on system sciences (HICSS'37), 2004.

[22] Latour, op. cit.

[23] F. Millerand, « Usages des NTIC, les approches de la diffusion, de l'innovation et de l'appropriation (1re et 2e parties) », COMMposite, 99 (1), 1999.

[24] Turnbull, D. (1993). "The Ad Hoc Collective Work of Building Gothic Cathedrals with Templates, String, and Geometry." *Science, Technology & Human Values* **18**(3): 315-343.

[25] S. L. Star & J. Griesemer, « Institutional ecology, 'translations,' and boundary objects: amateurs and professionals in Berkeley's museum of vertebrate zoology, 1907-1939 », Social Studies of Science, 19, 1989, p. 387-420.

[26] H. Karasti, K.S. Baker & E. Halkola, « Enriching the Notion of Data Curation in e-Science: Data Managing and Information Infrastructuring in the Long Term Ecological Research (LTER) Network ». In M. Jirotka, R. Procter, T. Rodden & G. Bowker (eds). Computer Supported Cooperative Work: An International Journal. Special Issue: Collaboration in e-Research, 15(4), 2006, p. 321-358.

[27] M. Berg, Rationalizing medical work : decision-support techniques and medical practices, Cambridge, MIT Press, 1997.

[28] M. B. Jones, C. Berkley, J. Bojilova & M. Schildhauer, « Managing scientific metadata », IEEE Internet Computing, 5 (5), 2001, p. 59-68.

[29] Latour, B. (2005). *Reassembling the social: an introduction to actor-network theory*. Oxford; New York: Oxford University Press.

[30] Douglas, M. (1986). How institutions think. Syracuse, N.Y., Syracuse University Press.