4th International Digital Curation Conference
December 2008

Data Stewardship: Environmental Data Curation and a Web-of-Repositories

Karen S. Baker and Lynn Yarmey
Scripps Institution of Oceanography, University of California, San Diego, USA

<July><2008>

Abstract

Scientific researchers today frequently package measurements and associated metadata as digital datasets in anticipation of storage in data repositories. Through the lens of environmental data stewardship, we consider the data repository as an organizational element central to data curation. One aspect of non-commercial repositories, their *distance-from-origin* of the data, is explored in terms of near and remote categories. Three idealized repository types are distinguished – local, center, and archive - paralleling research, resource, and reference collection categories respectively. Repository type characteristics such as scope, structure, and goals are discussed. Repository similarities in terms of roles, activities and responsibilities are also examined. Data stewardship is related to care of research data and responsible scientific communication supported by an infrastructure that coordinates curation activities; data curation is defined as a set of repeated and repeatable activities focusing on tending data and creating data products within a particular arena. The concept of "*sphere-of-context*" is introduced as an aid to distinguishing repository types. Conceptualizing a "*web-of-repositories*" accommodates a variety of repository types and represents an ecologically inclusive approach to data curation.

The 4th International Digital Curation Conference taking place in Edinburgh, Scotland over 1-3 December 2008 will address the theme *Radical Sharing: Transforming Science?*.

1.0 - Introduction

A point in time and a geographic location together typically define the origin of a field measurement which, when recorded, becomes data. A measurement represents an observation within a specified context as viewed from a particular perspective. The measurement 'context' refers in part to the properties of the broader physical environment in space and time and is recorded in the accompanying metadata. The context (and thus the metadata) also includes the technical and social environments comprised of instruments, people, traditions and organizational entities associated with obtaining the measurement as well as the later processing, storage, use, transport and reuse of the resulting data.

Scientific field measurements are primary data that are packaged with increasing frequency as digital datasets and housed in digital repositories. Regardless of research scope, the value of environmental data as a long-lasting product is increasing, even while the procedures, processes and roles that support the packaging and delivery of the data and metadata often remain behind the scenes, characterized by a complex production process. Through the lens of data stewardship we consider the data repository and its characteristics as a central element to data curation. We consider distinctive types of repositories that have different goals, participants, and "*spheres-of-context*" but that also have similarities in terms of roles, activities and responsibilities. We investigate "*distance-from-origin*" as a factor central to data curation activities and repository development.

2.0 - Background

In science, the journal publication is a traditional, recognized and specifically disseminated product of field research. With the advent of computers and a pervasive technical infrastructure available to a population intrigued by and familiar with all things digital, the dataset emerges as a new type of publication, a product with wide accessibility. However, methodologies regarding primary data and their transformations - processing, quality, filtering, description, and publication - remain understudied; theories, practices, and metadata standards addressing these questions are under development. Uncertainties about the data and its context affect the use and reuse, whether in the context of future scientific research, policymaking, or public education. One approach to addressing this gap in understanding and communication is the notion of data stewardship, carried out through interrelated data curation activities. The authors are long-time participants in scientific research activities associated with local-scale environmental field data acquisition, so the perspectives presented are those of an information manager and a data analyst.

2.1 - Considering data stewardship

Stewardship is a term that involves tending a community element not owned solely by one person. Appearing in the work of ecologists such as Aldo Leopold in connection with land stewardship (Leopold, 1968), the term is equally relevant when considering data products and the responsibility associated with their care and dissemination across multiple arenas. Unlike the finite and physically accessible nature of land, data are limitless and indirectly experienced (Eaton and Bawden, 1991), though the resources available to support the data are limited. In addition, land stewardship largely enjoys semantic agreement on familiar categories of land types (e.g. mountains, valleys, plains) and biome types (e.g. marine, watershed, grasslands). The units for measuring size (e.g. square units, hectares, satellite pixels) have been addressed and standardized. With data, however, types and descriptive standards for datasets and data collections are the subject of contemporary scholarly discussion and research. The advent of new measurement capabilities (e.g. autonomous vehicles and streaming technology), the increasing recognition of the value of long-term sampling strategies (e.g. time-series measurements and interdisciplinary studies) and the importance of data sharing initiatives (e.g.

master catalogues and interoperability efforts) all create unfamiliar data arrangements requiring identification of and agreement on new categories and descriptive standards as well as expanded dynamic infrastructures for both local and large-scale data endeavors.

Data stewardship provides a conceptual framework for envisioning the flow of data amongst and between arenas. It provides an entree to the notion of collective practices. The movement of data and development of datasets has been conceptualized in a variety of ways - often as steps in an assembly line or less prescriptively as "life stages" (Carlson and Anderson, 2007). Recently, the concept of the data lifecycle (IASIST, 2007) has opened up the data process with inclusion of sub-cycles and eddies. The notion of a data lifecycle has been described as more of a multi-loop cycle (Higgins, 2008; Lord and MacDonald, 2003; ARL, 2006; NRC, 2007) than a single-cycle pipeline beginning at data collection and ending with archive. In some ways, the term 'lifecycle' is a misnomer, however, in that an idealized data object is persistent and never 'dies'. An implicit assumption - perhaps a goal of the information age - is that responsible data stewardship will ensure data immortality; a preserved measurement, dataset or data collection will, through complete and accurate contextual description, be useful and accessible far into the future. It is an ambitious goal of mythic proportions, where myth is used in the sense of a generally understood cultural goal rather than in the sense of a mistaken understanding (Campbell, 1988; Baker and Stocks, 2007).

2.2 - Considering data curation

In an Association of Libraries report (ARL, 2006), data curation is defined as involving "ways of organizing, displaying, and repurposing preserved data." Lord and MacDonald (2003; p12) provide a series of models illustrating the development of the data curation roles, activities, and products over time. They also provide a working definition involving different levels of data curation:

> The activity of, managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose. Higher levels of curation will also involve maintaining links with annotation and with other published materials.

For this paper, data curation is focused on the set of practices that comprise tending the data record within a single repository while remaining cognizant of a broader context of related repositories gathered under a data stewardship umbrella. From a single repository perspective, data curation may be seen as a black box with input and output, that is, with associated data ingestion and delivery of data products. Taking a look inside the black box reveals a multi-activity process involving diverse organizational, technical and social elements that are brought together through collaboration and infrastructure-building (Lord and McDonald, 2003; ARL, 2006). Detailing these activities entails considering not only the flows and cycles within a single box but also the lifecycles of related scientific research endeavors, partnerships, and institutions as integral elements to data stewardship. Articulation and discussion of scientific practices, data, and data curation as well as strategic relationship building with other repositories are important, ongoing community activities (CUL, 2008; Karasti and Baker, in press; Carlson and Anderson, 2007; Wallis et al., 2007; Cragin and Shankar, 2006; Birnholtz and Bietz, 2003; Zimmerman, 2003).

2.3 Considering scientific practices

While automated, large-scale environmental observatory-style measurements are becoming more abundant, enormous value lies in manually controlled field-based sampling planned in the context of a particular research question. Manually taken measurements are ideally collected following predetermined sampling plans, however field conditions are not always ideal and sampling is commonly influenced by expert judgment about unanticipated acquisition events. Capturing specific elements of all aspects of the field experience through metadata is essential as this record is the permanent connection between the data and its origin. The lack of documented narrative describing local data taking and handling procedures

suggests that the variety of insights and choices that influence non-automated data sampling decisions may not be articulated in full but rather are passed on interpersonally through shared experience and field-based mentoring. It is generally recognized that metadata creation is a stumbling block in data description efforts. This may not be solely a matter of reluctance to document but rather a lack of time, vocabulary and a missing link to broader frameworks in the current nascent period of what may be considered a 'pre-federation' transition from individual-based to community-based research. The design of community agreements regarding category crafting, best practices, standards, and appropriate technological supports are also required.

Scientific investigations have expanded from point-based, intermittent measurements to having any number of spatial and temporal characteristic combinations. The multitude of combinations of these characteristics result not only in more data but also in more types of data. Adding contextual characteristics further increases the diversity of descriptive needs, the semantic challenges to negotiate and the difficulty of standardizing both. Indeed as new language and tools emerge that are meant to ease these difficulties, a significant measure of time is spent evaluating classification and technological approaches that often are discovered to be inappropriate. Elaborating on data characteristic combinations, maintaining consistent practices over time and conforming to changing community data and metadata standards requires resources and time that have traditionally been dedicated to scientific analysis.

3.0 - Organizational Arrangements

Repositories may be described technically in terms of their physical storage capacity, though in practice a working repository is a complex organizational unit that requires continuing management. An early definition of digital repositories was presented in terms of four elements that differentiate a repository from other collections (Heery and Anderson, 2005):

* content is deposited in a repository, whether by the content creator, owner or third party
* the repository architecture manages content as well as metadata
* the repository offers a minimum set of basic services e.g. put, get, search, access control
* the repository must be sustainable and trusted, well-supported and well-managed

The definition provides an important guide but encompasses elements that can be recognized as fluid. Even in subsequent years filled with accumulated experience, explicit specification of what constitutes 'basic services', 'trusted', 'well-managed' and even 'content' remains elusive and subject to discussion, research, and redefinition.

A traditional repository typically includes a technical infrastructure and an overarching vision subscribed to by participants. Frequently, an organizational status ensures continuous support. In addition, there is an emergent understanding of repository infrastructure as involving human and technical resources (Atkins, 2003; Edwards et al, 2007). With the number of organizational, social, and technical factors involved, it is of value early on in the growth of the curation concept to consider the diversity of existing practices, potential needs and possible arrangements that represent different repository types.

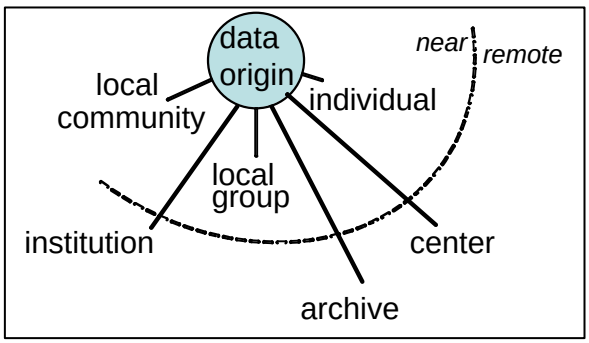


Figure 1 illustrates schematically one factor relevant to characterizing data repositories: their distance from the initial measurement point of data origin. The dashed line represents a symbolic boundary within a two-category classification based on distance-from-origin: local repositories

near the data origin and remote repositories farther from the origin. Near to the data origin are individuals, groups, and local communities with firsthand knowledge of and experience with the measurements and context. On the other side of the dashed line are organizational units more distant from the origin such as institutional repositories, community centers, and national archives. The distance here may be considered a *sociotechnical distance* having more to do with representation and communication issues as well as transformations and filter types than with physical distance or the number of repositories through which datasets have traveled (A.Gold, personal communication).

3.1 - Local repositories: near origin

Data capture carried out in a repository near to the researchers involved in data acquisition is defined as local. The activities and products from this part of data curation have been referred to as 'upstream' (Brandt, 2007). Data collected through manual physical sampling are often messy and quirky, requiring work before scientific interpretations can be made; local repositories develop in response to this agile data acquisition culture. An initial list of characteristics describing local repositories near the data origin is given in Table 1.

The primary driver of local repositories is data management in support of ongoing research. Taking responsibility for the capture of dynamic and frequently changing data, local repositories are designed with a focus on data acquisition. Changes in data and metadata are an expected part of the routine. For instance, data changes may involve processing shifts or quality control updates. To remain flexible in a digital environment subject to rapid change, participants at local repositories respond to the need for redesign and adaptability as new data types emerge. Local repository participants are closely connected with researchers and field-based research throughout the steps of the local scientific process and thus have a deep understanding of the data as well as the environmental and cultural contexts of the measurements. Local expertise facilitates the elicitation and capture of tacit knowledge about the data acquisition, sampling analysis, and quality procedures for creation of accurate and rich metadata records. These metadata become an effective frame for local best practices for data handling that emerge directly from local scientific practices and are essential to both immediate, effective data use and successful future data reuse. These best practices may be seen as an essential element of the standards development process.

Table 1 Environmental Research Data Repository Characteristics – Local and Remote

Characteristic	Local	Remote
Driver	Research	Service
Expertise	Data management	Collection management
Design focus	Acquisition, capture and use	Storage and reuse
Data state	Dynamic	Versioned
Design feature	Adaptability	Stability
Change mechanism	New data types and scientific practices	Widely-accepted data practices
Data knowledge type	Tacit, implicit, and explicit	Explicit
Standards contribution	Developing and enacting	Propagation

Close to the data origin, local repositories are in a unique position to initiate the data curation process, introducing it as part of local community practices (Lord and MacDonald, 2003; Heery and Anderson, 2005; Karasti et al, 2006). Local repository personnel have a unique translational role in bringing the broader notion of data stewardship to participants involved with the data at its origin. There are myriad activities that require mediation of local traditions as part of the broader notion of data curation and data stewardship: contributing to development and enactment of standards; making the case for data accessibility; co-designing data practices, information system practices, and scientific practices; and encouraging consideration of sustainable infrastructure during the scoping and planning phases of research.

These also promote community buy-in and create, through semantic negotiation, some of the common language needed for effective data curation and stewardship.

3.2 – Remote repositories: distant origin

Data curation activities more remote from the data origin and later in the data processing cycle are sometimes referred to as downstream (Gold, 2007; Nguyen, 2008). Remote repositories provide an expertise in data collection management. Their characteristics are summarized in the last column of Table 1. Remote repositories enable continuous data reuse and provide much needed services through continuing access and standardization. The initial focus is on dataset versioning or recordkeeping to index what has been captured and to address accession and long-term storage. The data objects here are buffered from the rapid ongoing local updates and versioning becomes a realistic means of tracking change. Data changes are less frequent and system updates are reserved for widely accepted data practices; both facilitate a stable repository design. Through developed expertise, explicit knowledge contained in data packages and collections can be mapped to broad community metadata standards.

As a link to the larger domain and multidisciplinary communities, remote repositories offer disparate local communities a common language as well as a broader organizational context. Assembling data from diverse sources allows local data to be translated to the larger context of global environments and multidisciplinary arenas. Data exchange protocols are required for data objects to be transported between repositories. A stable data object container and standardization of both container and content are necessary for frequent and reliable data exchange. Relationships between local and remote repositories facilitate not only data flow (often but not exclusively from origin or upstream to downstream) but also additionally the bi-directional flow of contextual perspectives and understanding. Remote repositories represent a framework for local repositories that promotes readiness for future integrative developments in data stewardship, curation and exchange.

3.3 - Example repository types

Considering a typology of repositories has been suggested as facilitating communication between repositories (Heery and Anderson, 2005). Though there are many types of *repositories* performing data curation in both local and remote categories, we open discussion of repository differences by considering the types of data *collections* defined by the National Science Board (2005). Three collection categories are identified: research, resource, and reference. The categories are not distinct and may be recognized as representing points across a landscape of possible configurations (Cragin and Shankar, 2006). Taking these non-exclusive categories along with organizational elements, we consider three idealized repository types for purposes of discussion: local, center, and archive (paralleling the research, resource, and reference collection categories respectively). The types are similarly recognized as under-defined in that community discussion has yet to mature with respect to categorization and naming; defining characteristics below are meant to prompt further discussion of various types of configurations and models.

An expanded set of characteristics is summarized in Table 2. Three exemplar repository types are shown: one near origin (local) and two remote from origin (center and archive). Repository differences are reflected in what may be described as repository goals serving different audiences with differing stakeholder interests that define that audience's task outputs: local management is attuned to data use for planned research, centers to current data reuse within the discipline, and archives to future data reuse. Data scope and related objects that typify the three repository types are the field program and its data, the discipline and datasets, and long-term programs and data collections, respectively. The near origin information data unit is the measurement while remote origin repositories are organized around data packages and collections. The repositories have differing collaborative configurations – partnered individuals, collaborative communities, and public service – that contribute to metadata in different ways highlighting methodological, contextual, and preservation description

correspondingly. There are significant ramifications to the differing funding arrangements that underscore the value of data flowing from local to archive repositories for long-term storage.

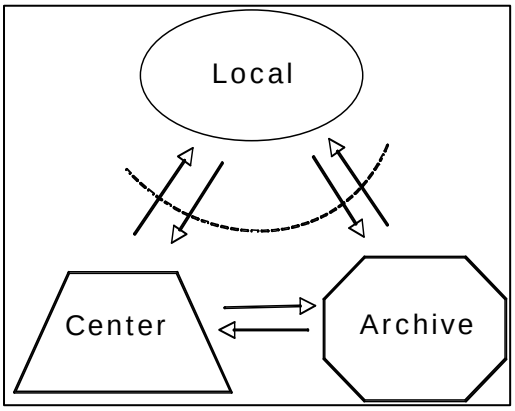
Table 2 Environmental Research Data Repository Type Characteristics

Repository Characteristic Category	Repository Characteristic	Repository Type		
		Local	**Center**	**Archive**
Position	Distance-From-Origin	Near	Remote Distance	Remote Distance & Time
Audience	Stakeholder Interest	Planned Research	Domain Use	Future Use
Scope	Data Scope	Field Program	Disciplinary	Themed
	Collection Type	Research	Resource	Reference
Standards	Metadata	Methodological	Contextual	Preservation
	Data Sharing	Informal	Formal	Public Standards
Curation & Maintenance	Task Focus (input)	Producing/Capturing	Registering	Preserving
	Task Focus (output)	Data Use	Current Data Reuse	Future Data Reuse
	Collaborative Configuration	Partnered	Collaborative Community	Public Service
	Data Object/Content	Data Values	Dataset	Data Collection
	Data Unit/Content	Measurement	Package	Collection
Budget Issues	Funding	Short-Term	Variable	Long-Term
Persistence & Reliability	Data Preserving	No	Variable	Yes

The local repository specializes in accommodating legacy scientific practices as well as developing augmentations and adapted methods for mediating existing data practices to contemporary arrangements and for interfacing with other repositories. Data centers making aggregated datasets available on community scale are the first to see and respond to the changing scope of comprehensive science questions. Archives serve to preserve materials as a part of the historical record with careful attention to data provenance.

4.0 – Discussion

A simplified illustration of relationships among three types of repositories is shown in Figure 2 where the local repository focuses on research collections, centers on resource collections, and archives on reference collections. This presentation contrasts with a linear conceptualization of data flow. The dashed line indicates the artificial boundary described earlier as data distance-from-origin. The bi-directional relationships between repository types are purposeful, underscoring that the communication of information through participants and the flow of data across machine exchanges are not necessarily uni-directional or predefined. In one scenario, participants at an archive may ask center participants for help in assessing collections while in another local participants may develop a new algorithm for checking data quality that migrates for use at the center level. There are many non-mutually exclusive paths, both existing and potential, in the lifecycle of a given data object. The dataflow for a particular object depends upon its use as well as its long-term value for reuse in disciplinary and interdisciplinary research. Data curation is necessarily distributed as no single person, technology or organization can support all data objects for all purposes including their final long-term maintenance. Data curation efforts involve human collaboration, negotiation, and



technological sharing arrangements to appropriately steward data objects amongst and between all appropriate repositories.

A framework proposed for a distributed network of computing arrangements in general by Kling and Scacchi (1982) provides a way of envisioning repository relations. We apply their web concept to repositories for visualization of a network or *web-of-repositories*. Figure 2 presents a simple multi-repository web model in contrast to more traditional discrete, chain, or pipeline repository models. Lord and MacDonald (2003) presents a broader notion of curation when they report that the term curation "covers a wider context than just archiving; it embraces the care of the record within scientific context and environment. This is particularly relevant for primary research data, as the term implies, is part of an ever-widening chain – indeed a chain which increasingly is a cycle." Data processing has been explored previously using pyramid, continuum, and package models (ARL, 2006; Treloar et al, 2007; Hunter, 2006). One strength of a web model is the inclusion of a wide range of activities and feedback loops. Some of these activities are discussed below. The web portrays in particular a non-hierarchical set of pathways that represents the complex set of data flows that occur in practice. The *web-of-repositories* supports the broader '*ecology of repositories'*, a sociotechnical framework that encompasses relations between repositories, data flow between repositories, and workflow issues (Heery and Anderson, 2005). The web as an element of the infrastructure requires new understandings of scientific practices, data practices, and *curation practices* as it enables distributed collective practices (DCP; Schmidt, 2000; King, 2006), scientific data collections (SDC), and the "conceptualizing of SDCs as distributed in nature and practice" (Cragin and Shankar, 2006).

Table 3 Data Curation Activities

Activity	Task/Responsibility	Role
Scope and System Design	Formulate research goal/foundation and determine scope and purpose; Methods and strategy development	Research planner
Gather	Acquisition (capture/record); quality assessment	Acquirer
Scientific Assess/Appraise	Determine relevance to planned research	Research assessor
Informatics Assess/Appraise	Consider relation to broader community efforts including standards-making	Informatics assessor
Describe	Description (definition & environmental contextualization)	Metadata provider
Ingest	Ingestion to build collection (capture, translation, organization & registration)	Receiver
Process	Quality control	Analyzer
Use	Consider for original question (local user) or new questions and issues (remote user)	User
Transform	Create derived products or synthesis	Mediator
Deliver	Enable public and/or other repository access through exchange or web access	Provider

4.1 – Data curation activities and roles

To expand further upon a basic understanding of repositories we consider the activities, tasks and roles associated with data curation (Table 3). Joint activities and user-oriented methods have been identified within the computer supported cooperative work community (CSCW) as well as the information systems design literature as coordinative practices key to communications and collaborative work (Schmidt, 2002). Soft systems methodology (Checkland, 1981; Checkland and Poulter, 2006), participatory design (Greenbaum and Kyng, 1991; Schuler and Namiok, 1993; Kensing and Blomberg, 1998) and information systems design (Isomäki and Pekkola, 2005; Barki and Boffo, 2007) are furthering contemporary understandings of users' roles with the introduction of key concepts – such as human activity systems and multiple perspectives. Table 3 presents the loosely associated, multi-activity tasks involved in carrying out data gathering and management as well as design, collection, assessment, description, processing, and delivery. This table extends an earlier table of

activities (Karasti and Baker, in press) adding tasks and roles. Table 3 is not intended to limit activities to a sequential workflow, but rather to explore the different tasks and roles associated with data, datasets, and data collections regardless of location. Each role represents a participant contributing to the data curation process whether in a local, center, or archive repository, similar to the human-in-the-loop identified as the reviewer who filters a flood of information or a participant in field sensor deployments who monitors data collection streams (Averman, 1999; Borgman et al, 2007; Wood and Stankovic, 2008).

4.2 – Sphere-of-context

We define 'context' as the common ground necessary for framing the broad issues of this discussion. The notion of context has been described as 'that elusive explanatory structure always invoked, never explained' (Galison, 2008). Certainly the broad contexts of 'e-science' and cyberinfrastructure apply to data curation but considering the many distinct realms of context illuminates the subtleties of data curation in practice.

In practice, data curation can be imagined as a shifting contextual window. From a multiple repository view, these contextual windows may be described as interrelated *spheres-of-context*. The activities and roles in Table 3 associated with curation can be applied at each node in the *web-of-repositories* albeit from a different perspective. As data objects are exchanged amongst spheres, the activity cycles repeat with new participants and goals. For example, we can loop through Table 3 from a local repository perspective by considering the activities in the context of data measurements, where data measurements are scoped, gathered, assessed, described, etc. and the roles are carried out by local participants. Locally, activities are often performed by members of the community best suited for each responsibility, for instance at a local repository the scientist may be the exclusive acquirer and scientific assessor, and retains full responsibility for the served data. Best practices concerning each activity (i.e. methods of description and appraisal) are greatly informed and influenced by community discussion with input from broader practices as translated through the local repository personnel. A local repository may also have a number of various data use policies representing different projects or contracts.

From a data center perspective, the context shifts to data packages and the activities repeat, the roles carried out by data center personnel. For manually collected environmental data, centers are a step removed from local-scale science though they maintain an understanding of the local *sphere-of-context*, as in the cases of field datasets that are passed along or streamed data that bypasses the local repository. In this sphere the repository as an organizational unit takes ownership and responsibility for the data package and the repository personnel are the primary participants.

Repeating this exercise at an archives facility, activities and roles are set in the context of long-term data collections. Archives may take on responsibility for their data collections, transferring decision-making responsibility from data providers in order to fulfill the long-term needs of collections over generations. Decisions and practices at this level are not necessarily linked to the original data capture *sphere-of-context* except through retained metadata. A single set of policies applies to all archived contents.

Imagine Figure 2 as having *n* repository nodes rather than three nodes. These repositories and their *spheres-of-context* overlap and blend but taken as a whole they encompass a distributed collective practice within the framework of data stewardship. With the common activities and roles, there exist many points of shared understanding amongst the different repositories. Creating not only relationships of inter-repository data object exchange, these understandings have the potential to create relationships of bi-directional learning and process exchange through collaboration and mediation within the diverse repository community.

5.0 - Conclusions

We visualize data not solely in relation to one *sphere-of-context* but simultaneously as moving over time through many repositories, with a cumulative value as defined by all participating repositories and *spheres-of-context*. This contrasts with a single repository curation view with all outside activities regarded as pre-ingestion or limited, non-standard data handling. Whether working with data or metadata organized within dataspaces, databases or data collections, repositories represent a definable set of activities and roles, practices and procedures as well as relations and products. Developing in response to different needs, priorities, and cultures, single repository activities support lifecycle activities and inform data practices; we shift into a broader context of a *web-of-repositories* and see data stewardship as the coordination of many such distributed efforts. A *web-of-repositories* presents a challenge to information infrastructure-building and standards development. Yet we have a model of web repositories at hand. Consider the complex network of libraries and library communications - from personal through local branch and city central as well as institutional and regional to national archival libraries. Though traditionally dealing with predominantly physical artifacts, the library community provides one model for the preservation of dataset artifacts.

Articulating the types and characteristics of lifecycle eddies within and between repositories is essential as communities of all sizes move forward with the development of appropriate organizational, social and technical responses to emergent data types and arrangements. Our intent is not to limit advancement of repositories or to prescribe a plan for current or future webs, but rather to contribute to awareness of a diversity of repository types and characteristics as a step towards semantic convergence thus setting the stage for future collaboration. With three artificial exemplar repository types, we initiate exploration into organizational arrangements and the overlapping areas between them while emphasizing the importance of a non-hierarchical, federated approach to data stewardship. In a *web-of-repositories*, the foundation is not built on a single “natural heir” to data but rather built on a diverse group of nodes in open social, technical and organizational communication which enables community-designed data exchange, category building and standards-making. In fostering an inclusive arena for repositories, we can learn from characterizing similarities and differences.

Stewardship is a broad ecological concept pertinent to environmental data curation as well as to what may be considered our current pre-federation state of repositories. Data stewardship in particular is a unifying umbrella able to shelter diverse, interrelated activities at nodes throughout a web of repositories. Both “*distance-from-origin*” and “*spheres-of-context*” are concepts that provide some insight into differing repository types. Another key to understanding and further developing the stewardship framework is the recognition of a set of similar roles within the realm of each repository, possibly a ramification of having similar work elements (input, storage, output). Yet, in practice, the choices and implementations at each repository differ with respect to methods and processing (semantic, structural, syntactic) with the broader considerations of governing arrangements and formalizing knowledge. While this parallel repository development at first glance seems to create barriers to interoperability, it is part and parcel of a distributed network, i.e. a partnership with each repository specializing in a different aspect of or approach to the data and curation. With elaboration on data curation, a multi-faceted view becomes apparent, describing multiple repositories and spheres of context requiring coordination to create a robust, distributed and flexible vision of data stewardship in practice. Joint consideration of the full range of repository types and their *spheres-of-context* leads to a better understanding of collaborative requirements for the making, growth and diffusion of community standards. Data stewardship and data curation are neither problems to be solved nor solutions in and of themselves. Rather these concepts represent dynamic learning arenas. Through context-aware data curation arises the possibility of designing, federating and sustaining an interoperable “*web-of-repositories”*, fulfilling the ultimate goal of data stewardship.

Acknowledgements

Support is provided by NSF OCE #04-05069 CCE, OPP #02-17282 PAL, CalCOFI, and SBE/SES #04-33369 CIP. We acknowledge the contributions of Helena Karasti in foregrounding the experience of Information Managers and Florence Millerand in following the data as well as the roles associated with the data.

References

[report] ARL, Association of Research Libraries (2006). To Stand the Test of time: long-term stewardship of digital data sets in science and Engineering. A Report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe, September 26-27, 2006, VA.

[report] Atkins, D.E., Droegemeier, K.K., Feldman, S.I., Garcia-Molina, H., Klein, M.L., Messerschmitt, D.G., Messina, P., Ostriker, J.P., & Wright, M.H. (2003). Revolutionizing Science and Engineering Through Cyberinfrastructure, Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure [Web-document]. Available: http://www.communitytechnology.org/nsf_ci_report/ [Last referenced: 23.05.2006].

[thesis] Averman, C. (1999). Using "Human-in-the-loop" in an adaptive system: An evaluation study of the ConCall system. Masters thesis. http://www.handels.gu.se/epc/archive/00001335/01/averman.ia7400.pdf

[proceedings] Baker, K.S. & Stocks, K.I. (2007). Building Environmental Information Systems: Myths and Interdisciplinary Lessons. Proceedings of the 40th Hawaii International Conference on System Sciences. HICSS40, IEEE Computer Society, January 2007, Hawaii.

[journal article] Barki, H, Titah, R. & Boffo, C. (2007). Information system use-related activity: An expanded behavioral conceptualization of individual-level information system use. Information Systems Research, **18**(2): 173-192.

[proceedings] Birnholtz, J.P. & Bietz, M.J. (2003). Data at work: supporting sharing in science and engineering. November 2003. Group '03. Proceedings of the 2003 International ACM SIGGROUP.

[proceedings] Borgman, C.L., Wallis, J.C., Mayernik, M.S., & Pepe, A. (2007). Drowning in data: Digital library architecture to support scientific use of embedded sensor networks. In: Proceedings of the *ACM/IEEE Joint Conference on Digital Libraries 2007*, Vancouver, BC. http://polaris.gseis.ucla.edu/cborgman/pubs/JCDL07_117_fin.pd

[online journal] Brandt, D.S. (2007). Data, research, metadata, metaresearch," presentation at ACRL/STS, ALA annual meeting, June 2007.<http://www.ala.org/ala/acrlbucket/stsconferencepro/annual2007programs/brandt.pdf>. Accessed 9/12/07.

[book] Campbell, J. (1988). The power of myth. Anchor Books, NY.

[online journal] Carson, S. & Anderson, B. (2007). What are data? The many kinds of data and their implications for data re-use. Journal of Computer-Mediated Communication 12(2), article 15.

[book] Checkland, P. (1981). Systems Thinking, Systems Practice. Wiley, Chichester.

[book] Checkland, P. & Poulter, J. (2006). Learning for Action. A Short Definitive Account of Soft Systems Methodology and its Use for Practitioners, Teachers and Students. John Wiley and Sons, West Sussex, England.

[journal] Cragin, M.H. & K.Shankar (2006). Scientific Data Collections and Distributed Collective Practice. Computer Supported Cooperative Work 15:185-204.

[report] CUL (2008). Digital Research Data Curation: Overview of Issues, Current Activities, and Opportunities for the CUL. Cornell University Library Working Group. http://ecommons.library.cornell.edu/handle/1813/10903

[journal article] Eaton, J.J. & Bawden, D. (1991). What Kind of Resource is Information? International Journal of Information Management.

[report] Edwards, P.N., Jackson, S.J., Bowker, G.C. & Knobel, C.P., (2007) *Understanding Infrastructure: Dynamics, Tensions, and Design*. Ann Arbor, http://hdl.handle.net/2027.42/49353

[journal article] Galison, P. (2008). The Problems in History and Philosophy of Science. ISIS 99:111-124

[journal article] Gold, A. (2007). Cyberinfrastructure, Data, and Libraries, Part 2. D-Lib Magazine 13(9/10). http://www.dlib.org/dlib/september07/gold/09gold-pt2.html.

[book] Greenbaum, J. & Kyng, M. (1991). Design At Work: Cooperative Design of Computer Systems. Lawrence Erlbaum Associates, Hillsdale, NJ.

[report] Heery, R. & S. Anderson (2005). Digital Repositories Review UKOLN and AHDS: 33. http://www.jisc.ac.uk/uploaded_documents/digital-repositories-review-2005.pdf

[journal article] Higgins, S. (2008). The DCC Curation Lifecycle Model. IJDC 3:134-140. http://www.ijdc.net/ijdc/article/view/69/69
[report] Hunter, J. (2006). Scientific models – A user-oriented approach to the integration of scientific data and digital libraries. University of Queensland.
[report] IASSIST (2007). Conceptualizing the digital life cycle, IASSIST Communique. (http://iassistblog.org/?p=26)
[proceedings] Isomäki, H. & Pekkola, S. (2005). Nuances of Human-Centredness in Information Systems Development, Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05). Hawaii.
[journal article] Karasti, H., Baker, K.S., & Halkola, E. (2006). Enriching the notion of data curation in e-science: data managing and information infrastructuring in the Long Term Ecological Research (LTER) network. Computer Supported Cooperative Work 15, 321-358.
[proceedings] Karasti H. & Baker, K.S. (2008). Digital Data Practices and the Long Term Ecological Research Program Growing Global, Int Journal of Digital Curation, in press.
[journal article] Kensing, F. & Blomberg, J. (1998). Participatory Design: Issues and Concerns. Computer Supported Cooperative Work 7: 167-185.
[journal article] King, J.L. (2006). Modern Information Infrastructure in the Support of Distributed Collective Practice in Transport. Computer Supported Cooperative Work 15:111-121.
[journal article] Kling, R., & Scacchi, W. (1982). The web of computing: computer technology as social organization. Advances in Computers, 21, 1-90.
[book] Leopold, A. (1968). A Sand County Almanac. Oxford University Press.
[report] Lord, P. & Macdonald, A. (2003). e-Science Curation Report-Data Curation for e-Science in the UK: An Audit to Establish Requirements for Future Curation and Provision. Prepared for the JISC Committee for the Support of Research (JCSR). Twickenham, UK, The Digital Archiving Consultancy Limited. Available: http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf [Last referenced: 23.05.2006].
[report] NRC (2007) National Research Council. Environmental Data Management at NOAA: Archiving, Stewardship, and Access. National Academies Press. Washington DC.
[report] NSB (2005). National Science Board: Long-Lived Data Collections: Enabling Research and Education in the 21st Century. NSF NSB-05-40.
[report] Nguyen, T. (2008). On Keeping Data Open and Free. http://sciencecommons.org/weblog/archives/2008/04/23/nguyen-on-keeping-data-open-and-free/.
[journal article] Schmidt, K. (2000). Distributed Collective Practice: A CSCW Perspective. Presented at Conference on Distributed Collective Practice, September 19-22, 2000, Paris. http://www.itu.dk/people/schmidt/papers/dcp.paris2000.pdf
[journal article] Schmidt, K. (2002). Remarks on the complexity of cooperative work. In cooperation and Compexity, H.Benchekroun and P. Salembier (eds). RSTI, Hermes, Paris.
[book] Schuler, D. & Namiok, A. (1993). Participatory Design: Principles and Practices. Lawrence Erlbaum Associates, Hillsdale, NJ.
[online journal] Treloar, A., Groenewegen, D. & Harboe-Ree, C. (2007). The Data Curation Continuum: Managing Data Objects in Institutional Repositories. D-Lib Magazine 13(9/10).
[proceedings] Wallis, J.C., Borgman, C.L., Mayernik, M.S., & Pepe, A. (in press). Know thy sensor: Trust, data quality, and data integrity in scientific digital libraries. European Conference on Research and Advanced Technology for Digital Libraries 2007, Budapest, Hungary. [pdf]
[proceedings] Wood, A.D. & Stankovic, J.A. (2008). Human in the loop: distributed data streams for immersive cyber-physical systems. ACM SIGBED Review 5(1). http://portal.acm.org/citation.cfm?id=1366303&dl=GUIDE&coll=GUIDE
[thesis] Zimmerman, A.S. (2003). Data sharing and secondary use of scientific data: experiences of ecologists. Ph.D. Thesis, The University of Michigan, Ann Arbor.