

## Chapter 13

### **Lessons for Cyberinfrastructure Projects<sup>147</sup>**

Trying to extrapolate into the future based on what has occurred over the past two decades with digital libraries and the web ignores the fact that our analysis of the past benefits from 20/20 hindsight. However, it is possible to understand the factors that constrained our thinking in the past and interfered with our ability to see the future as it unfolded before us. This understanding may at least make it possible for us to be more flexible in the matter in which we approach similar problems in the future. This is especially important as we enter into a new phase of large-scale cyberinfrastructure projects that in many ways resemble the digital libraries initiatives described in this dissertation. These similarities include a mixture of core research and application, the development of and deployment of infrastructure, the involvement of multiple communities and disciplines, and the need for sustainability of both technology and organizational structures.

This is particularly germane to my future research since I am Principal Investigator on the Cornell portion of a 10 year, \$20 million grant from the Sustainable Digital Data Preservation and Access Network Partners (DataNet) Program [9] in the Office of Cyberinfrastructure at the NSF. The solicitation for the DataNet program describes its purpose:

---

<sup>147</sup> The content of this chapter benefited from conversations from the following colleagues on the Data Conservancy Project: Christine Borgman (UCLA), Sayeed Choudhury (Johns Hopkins), Mary Marlino (UCAR), Carole Palmer (UIUC).

Science and engineering research and education are increasingly digital and increasingly data intensive. Digital data are not only the output of research but provide input to new hypotheses, enabling new scientific insights in driving innovation. Therein lies one of the major challenges on this scientific generation: how to develop the new methods, management structures and technologies to manage the diversity, size, and complexity of current and future data sets and data streams. This solicitation addresses that challenge by creating a set of exemplar national and global data research infrastructure organizations (dubbed DataNet Partners) that provide unique opportunities to communities of researchers to advance science and/or engineering research and learning. [9]

At this date (August 2009), the DataNet Program has approved funding for two projects, including the project I am involved in, and plans include funding up to five projects, all large collaborations that will subsequently need to collaborate with each other.

The particular project that I am co-PI in is called the *Data Conservancy* [10, 120] and begins in September 2009. It is a collaboration between Johns Hopkins University, Cornell University, University of Illinois at Urbana-Champaign, University of California at Los Angeles, Fedora Commons, the Encyclopedia of Life, and many others. Quoting from the proposal text, "the Data Conservancy embraces a shared vision: data curation is not an end, but rather a means to collect, organize, validate and preserve data to address the grand research challenges that face society." Furthermore the proposal states, "the overarching goal of Data Conservancy is to support new forms of inquiry and learning to meet these challenges through the creation, implementation, and sustained management of an integrated and comprehensive data integration strategy."

The infrastructure proposed in the Data Conservancy is based on the notion of the *observation* [487] and its commonality across scientific disciplines. As explained in the proposal text:

... *observations* are the foundation of all scientific studies, and are the closest approximation to facts. Observations are objective measurements of entities at a particular location and time, which are gathered through a myriad of mechanisms that range from sophisticated telescopes mapping the galaxies to citizen scientists logging birds that visit a backyard bird feeder. All scientific observations share the same semantic template: they consist of an *object/event/phenomenon* captured via some *observing method* at a *location/time* and recorded as some *database entry/spectrum/image*. Developing a model of observations that can be generalized across disciplines and extended for specific instances is a key challenge and expected innovative result of The Data Conservancy. [10] (emphasis in original)

The project plans to deploy an eScience infrastructure based on this model that leverages a variety of technical components including Fedora and OAI-ORE, both of which are described in this dissertation.

The remainder of this chapter suggests a number of guiding principles for the Data Conservancy Project and other similar projects as they move forward in their work in the coming years. These principles are based on the analysis of digital library research projects outlined in this dissertation and hopefully represent some lessons we can learn from that previous experience.

### **Understanding the complexity of infrastructure**

The notion of “infrastructure”, of which cyberinfrastructure is one instance, has been a dominant aspect of society since the beginning of the Industrial Revolution, and has over the last several decades attracted the attention of social scientists and historians. Friedlander’s excellent set of studies of the pre-Internet infrastructures [195-198] (e.g., railroads, electricity, telephones and telegraphs, and banking) provide an excellent introduction to the complexities and sociotechnical aspects of infrastructure development and acceptance. These complexities are summarized by Starr and Ruthleder [445] in their eight dimensions of infrastructure (as described by Borgman [79]). These dimensions are: the fact that it is embedded in other social and

technological structures, its invisibility when it is working properly, its visibility when it breaks, the process by which it is learned as part of membership of the group or organization, the manner in which it is linked with day-to-day work practices, the manner in which it is standardized and therefore can link with other standardized practices, and the manner in which it builds upon an installed base. All of these dimensions are evident in the web and digital libraries as instances of “information infrastructure”.

The term “cyberinfrastructure” was introduced into the US national funding agenda by the so-called Atkins Report [33]. A more recent report defines cyberinfrastructure as follows:

Cyberinfrastructure integrates hardware for computing, data and networks, digitally enabled sensors, observatories and experimental facilities, and an interoperable suite of software and middleware services and tools. Investments in interdisciplinary teams and cyberinfrastructure professionals with expertise in algorithm development, system operations, and applications development are also essential to exploit the full power of cyberinfrastructure to create, disseminate, and preserve scientific data, information, and knowledge. [381]

This definition is notable because of both the breadth of its technical vision and the absence of acknowledgement of social implications and complexities. Unfortunately, this bears some resemblance to the historical trajectory of digital library research in which social implications were either ignored or perceived as only relevant for after-the-fact evaluation of technical developments.

I agree with many of my colleagues in the Data Conservancy Project that the success of this and similar projects depends on the immediate and continued close collaboration between the technical experts, the computer scientists, and the social scientists, who have studied and understand the practices and workflows of the target communities and the manner in which proposed technologies conform to them.

Quoting Chris Borgman, whom I interviewed for this chapter, “good evaluation starts

at the beginning” before technical products are created and throughout their creation. This attention to “in-process evaluation” rejects the simplistic notion that building infrastructure is something that is planned and mechanical. Instead, as described in the excellent report “Understanding Infrastructure: Dynamics, Tensions, and Design”:

... the path between the technological and the social is not static and there is no one correct mapping. Robust cyberinfrastructure will develop only when social, organizational, and cultural issues are resolved in tandem with the creation of technology-based services. Sustained and proactive attention to these concerns will be critical to long-term success. [174]

This need for “sustained and proactive attention” will continue throughout the lifespan of the project and the commitment to collaboration across the social science/computing and information divide is essential to success. By involving the target communities in infrastructure development throughout the project lifecycle, it will be possible to continually adapt the developing infrastructure to the evolving needs of the stakeholder communities. It will be possible to mitigate notions that it was imposed by an external party, and instill the sense that it arose based on internally recognized needs.

### **Recognizing community diversity**

Receptivity to new technologies for scholarly communication and practice varies greatly across disciplines and scholarly communities. In many cases the level of acceptance depends on the manner in which the new technologies represent continuity by building on pre-existing practices and values. An example is high-energy physics where a “preprint culture” existed long before the arXiv<sup>148</sup> preprint server was created [456]. The exchange of preprints among authors and institutions was standard practice, and this practice extended to the institutional and library level in which these preprints

---

<sup>148</sup> <http://arxiv.org>

were indexed and collected. As a result, the arXiv, which is essentially a mapping of those traditional paper practices to the digital library environment, has evolved over the years into the first choice resource for scholarship in a number of fields of physics and mathematics with similar historical practices. Experience with other disciplines with very different historical practices is revealing. For example, the preprint model failed entirely when attempted in the early 2000s in chemistry<sup>149</sup>, and the concept had to be significantly altered before its take off as PubMed Central in biomedicine [272].

Our own work in this area [472, 473] combines ethnographic and bibliometric analysis as a means to understand the nature of scholarly communities, the correspondence of the structure of those communities to “communication cultures”, and the effects and influences of interdisciplinary activity on the nature of those cultures and their receptivity to new technologies. Our initial results indicate substantive distinguishing features at even the sub disciplinary (e.g., biochemistry, physical chemistry) level.

Clearly, then, the success of Data Conservancy and similar projects depends on continued and in-depth understanding of the languages, norms, and practices of the target disciplinary communities. In an interview about this chapter, Mary Marlino of the University Corporation for Atmospheric Research (UCAR) used the term “empathy” to characterize this process. Achieving such empathy requires an understanding and acceptance of practices that may seem archaic and sub-optimal, but which can not be erased by the immediate infusion of new technology.

Furhtermore, technical artifacts that are created by the project must simultaneously accomplish a level of interoperability sufficient for meaningful cross-disciplinary data activities, while at the same time accommodating a level of specificity sufficient to

---

<sup>149</sup> Chemistry Preprint Server (CPS) <http://www.sciencedirect.com/preprintarchive>

express and allow for disciplinary diversity. This diversity exists at the semantic level and workflow level. The late Jim Gray of Microsoft Research, before his untimely and mysterious disappearance while sailing<sup>150</sup>, suggested the notion of “20 questions” for requirements gathering<sup>151</sup> for data oriented infrastructure projects. While the technique is of course not complete, it provides shorthand guidance that technical infrastructure at a minimum should be able to answer the questions that domain scientists want to ask of their data while at the same time considering the wider questions of cross domain interoperability. Although the altruism of the latter appeals to some domain scientists, the importance of demonstrating the advantages of infrastructure to the specific domain scientist can not be over emphasized.

### **The danger of the “seduction of the known”**

The costs of projecting the past onto the future, or “horseless carriage thinking”, have been described throughout this dissertation. It led to digital libraries that looked very similar to their bricks and mortar predecessors.

In some cases this phenomenon is due to the effect of *institutional culture*, in which thinking and imagination are constrained by the practices of the past. In his groundbreaking work on disruptive innovation [122], Clayton Christensen describes how disrupted institutions fail to confront innovation because of the matter in which the resources and vision are limited by attention to existing customers and traditionally successful products.

In other cases, it is the result of what is called “path dependence” among infrastructure experts. As defined by the “Understanding Infrastructure” report: “path dependence

---

<sup>150</sup> <http://research.microsoft.com/en-us/um/people/gray/>

<sup>151</sup> <http://www.stccmop.org/node/909>

refers to the “lock-in” effects of choices among competing technologies. It is possible following widespread adoption, for inferior technologies to become so dominant that superior technologies cannot unseat them in the marketplace” [174]. For example, once a commitment to a railroad gauge is made, it is extremely expensive and impractical to modify that gauge even if there are persuasive reasons for the advantages (e.g., speed, safety) of adopting a new gauge. Similarly, the innovations that can be adopted by an information infrastructure organization such as a library are limited by their historical and resource commitment to their own “railroad gauge”; e.g., a cataloging standard, a library management system, etc.

Sayed Chaudhury, principal investigator of the Data Conservancy project, noted the “seduction of the known” that is prevalent in digital preservation projects.

Preservation has historically been conceived of as a service associated with the *institution*; such as a library, museum, or archive. However, as we begin to conceive of preservation of data in 2009, in projects such as the Data Conservancy, we need to recognize and conceive of solutions that are free of traditional institutional bindings and exploit distributed, networked computing and phenomena such as cloud computing. Chaudhury pointed to projects like SETI@home<sup>152</sup>, which demonstrate how large-scale problems can be approached in radical new ways that abandon reliance on traditional institutions.

In fact, many scholars are recognizing the manner in which network technologies affect and even undermine the justifications for many of our existing institutional frameworks. According to the well-known futurist Clay Shirky [433] this change lies in the massive reduction in transaction costs due to the movement from the physical to

---

<sup>152</sup> <http://setiathome.ssl.berkeley.edu/>

the online environment. According to Shirky, the traditional environment in which the management, storage, and acquisition of physical artifacts entailed large costs and investment, required institutional structures such as libraries, publishers, and archives with sufficient financial reserves and economies of scale to support such costly transactions. Although the digital environment is certainly not free – services such as curation and secure storage of valuable digital resources require expertise and long-term financial investment – Shirky notes that the transaction costs for dissemination and short-term storage of digital content are virtually zero. This has a dramatic impact on the justification for institutional frameworks that were built on the presumption of high transaction costs. Yochai Benkler in his excellent book “The Wealth of Networks” [48] presents a similar argument.

This breakdown in traditional institutional boundaries has opened the door for the recognition of and involvement of scholarly contributions from outside the established university and research institutes. This phenomenon known as “citizen science” has led to projects like SETI@home<sup>153</sup> and Project FeederWatch<sup>154</sup>. Future cyberinfrastructure projects such as Data Conservancy must recognize the increased relevance of scholarly activities and observations that take place outside the institutions that previously contained them, and must consequently devise infrastructure that works across these highly distributed and individual citizen scientists. High-cost infrastructure that requires system administrator support is simply not viable in this type of environment.

---

<sup>153</sup> <http://setiathome.ssl.berkeley.edu/>

<sup>154</sup> <http://www.birds.cornell.edu/pfw/>

### **Understanding the difference between text and data**

During an interview about the content of this chapter, Carole Palmer of University of Illinois at Urbana-Champaign discussed the complexity of data, the manner in which they are distinct from textual digital objects, and the unknown consequences as we move data to a Web 2.0 online environment. Our experience with data use is quite different than that for text. As noted by Palmer, long before the appearance of the World Wide Web and the notion of online information there was a body of scholarship that provided a reasonable level of understanding about how scholars used and manipulated textual resources. Despite this wealth of knowledge, the movement of text to the online environment, and especially to the Web 2.0 environment that supports the deconstruction and reconstruction of that text, has had consequences that none of us imagined in the early days of digital library research.

The situation is quite different for data. This is because whereas data have always been a crucial ingredient in scientific explorations, until recently they were not treated as first-class objects in scholarly communication, in the same manner as the research papers that report on findings extracted from the data. This is rapidly changing. There are currently active discussions and exploration of implementing all core functions of scholarly communication – registration, certification, awareness, archiving, and rewarding [421] – for data sets. Increasingly, there is widespread recognition of the need for an infrastructure to facilitate discovery of shared data sets [426]. And, efforts at defining a standard citation format for data sets take for granted that they are primary scholarly artifacts [19].

Despite the fact that these changes are underway in the manner in which we view and handle data, according to Palmer we have very little scholarly evidence other than anecdotal about the manner in which scholars use, share, and maintain their data sets.

Furthermore, she states that it would be erroneous to base our assumptions of data behavior on book behavior or to extrapolate from Web behavior. Our only choice then is to focus on the few communities such as astronomers who have paid some attention to data practices and projects like the National Virtual Observatory<sup>155</sup>. Although the experiences of these communities will be extremely valuable as we move forward in the Data Conservancy project, we must be very careful about generalizing the discipline-specific practices. As described in the previous section, these generalizations proved incorrect in the area of online preprint dissemination. In the end, throughout our work we must be prepared for and constantly ready to react to significant changes in the use of data as the technologies that support its use and reuse evolve. We may indeed find that these changes are even more profound than that which occurred with text.

### **Rapid prototyping and moving targets**

Ultimately, the Data Conservancy Project and similar DataNet projects will be building technology amidst a moving target of contexts. This is not unlike the digital library projects that assumed a stable context of traditional library institutions, an Internet that was largely the domain of scholars and scientists, and an immature World Wide Web that was by and large a distributed document store; and then found themselves in a vastly different information environment that questioned and contradicted their fundamental assumptions. There is no way to avoid or foresee this. I agree with Sayeed Chaudhury who in our interview for this chapter stressed the importance of rapid prototyping and the need for a “advance and retreat” strategy in which we iteratively demonstrate new solutions with full knowledge that they may

---

<sup>155</sup> <http://www.us-vo.org/>

lead to dead ends. This may be trying on our impatience to find “the solution” or the need for our funders to demonstrate immediate results, but ultimately it is the only means by which we will flexibly absorb and integrate the inevitably changing context in which we work.