

## Chapter 1

### Lost Identity

*“The future ain’t what it used to be”*

- Yogi Berra<sup>1</sup>

The digital library was imagined as the “library of the future”, but increasingly the digital library seems to be loosing its identity in the emerging *participatory culture* [251] of Web 2.0.

---

Beginning with Paul Otlet’s visionary work [392] in the 1930’s, librarians, joined later by computer and information scientists, have been exploring the potential of new information technology (IT) to enhance and expand the functions of future libraries [78, 105, 256, 346, 428]. Enthusiasm about the possible transformative effect of IT on the library increased towards the end of the 20<sup>th</sup> century in response to rapid advances in computing and networking and breakthroughs in the field of information retrieval [429]. This led to a number of early prototypes and implementations that were referred to by a variety of names including “electronic libraries” [11] or “electronic publishing” [183]. These efforts matured in the 1990s into the current notion of *digital libraries* (DL).

Digital libraries subsequently emerged as an active research field in the 1990’s due to DARPA funding of the Computer Science Technical Reports Project (CSTR) [139]

---

<sup>1</sup> [http://en.wikiquote.org/wiki/Yogi\\_Berra](http://en.wikiquote.org/wiki/Yogi_Berra)

and from a series of workshops and reports [211, 257] that laid the foundation for two well-funded inter-agency (NSF, DARPA, NASA, NEH, Library of Congress, and NLM) Digital Libraries Initiatives (DLI-1 and DLI-2) [219, 379, 380]. These funding programs encouraged the growth of an active research community, composed mostly of computer and information scientists, but also including librarians, archivists, social scientists, and experts from a number of specialized disciplines. With the establishment of a number of digital library journals and conferences, the mechanisms of scholarly communication upon which community identities are built [473], digital libraries had matured by the turn of the century into a well-defined scholarly field.

When measured as a research initiative (e.g., scientific integrity, impact), the results of digital library funding and associated activities have been notably successful.

Significant results of digital library work include the PageRank algorithm [96] that evolved into the Google search engine, OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) [314], Dublin Core [2], OpenURL [12], DSpace [438], Fedora [402], LOCKSS [419], and many others. The results of digital library research are also evident in the widespread deployment of disciplinary archives such as arXiv<sup>2</sup> and others [226, 276, 280], institutional repositories [143, 253, 438], and large archiving efforts such as Brewster Kahle's Internet Archive<sup>3</sup>. I, and many of my colleagues in the field, contend that these results more than justify the level of federal and foundation funding in DL research and deployment.

However, the motivation for digital library activities, including the DLI funding programs, extended beyond individual research results. With an almost messianic

---

<sup>2</sup> <http://arxiv.org>

<sup>3</sup> <http://www.archive.org>

enthusiasm fueled by a considerable sum of research money, members of the international digital library community envisioned the creation of a global network of “Libraries of the Future” [346] federated together via a common interoperability fabric<sup>4</sup> [330, 394].

When measured in this more ambitious context, the success of the digital library effort is considerably less certain. This dissertation examines the reasons for this. My goal is not to devalue the successes of DL research or suggest that the DL community of researchers and funders should have approached the problem differently. The benefits of hindsight make this unfair. However, an analysis of the factors that mitigated the broader impact of DL work might lead to better understanding of the issues underlying the development of information infrastructure. A key issue examined here is the ability to define the technology of information infrastructure from above, the approach taken by the DL funding agencies<sup>5</sup> and the researchers that they supported, versus emergence organically from within, as was the case with the web. Understanding this and related issues may improve our approach to current large-scale infrastructure

---

<sup>4</sup> This vision was based, in part, on the successful experience that the lead funding agencies, especially NSF and DARPA, had with earlier development and deployment of the Internet [331]. The federated digital libraries concept was similar in form to the Internet that joined together a set of heterogeneous nodes into a global fabric with common protocols and standards. According to Bill Arms “The DARPA program officer for the Digital Libraries Initiative once observed that the only reason DARPA funded digital libraries was to stimulate research in interoperability” [26].

<sup>5</sup> A legitimate question to ponder is why the “imposition from above” model was successful in the context of the Internet, but not in DLs. A look at the history on the Internet [331] reveals a key factor that initial deployment and ramp-up occurred within a tightly scoped community, academic institutions and (primarily defense-related) research labs. The infrastructure had a long percolation period in this context before its subsequent mass popularization. This is quite different than the DL infrastructure work, which from the beginning was motivated by visions of widespread grassroots dissemination inspired by scenarios such as that articulated by then Vice President Gore in his “schoolchild in Carthage, Tennessee plugs into the Library of Congress” speeches (<http://www.ibiblio.org/icky/speech2.html>).

efforts, such as cyberinfrastructure and eScholarship. It also may help us redirect the intellectual energy of future DL work in new directions.

This analysis is based on the current (2009) status of the broader DL vision. Except for a few small-scale prototypes or instances based on limited functionality technologies (e.g., metadata sharing), the global network of digital libraries has not emerged<sup>6</sup>. The term “digital library” is notably absent from popular usage (“I’ll look that up in the digital library” is not an often-heard phrase). Support for digital library research has almost vanished from U.S. federal funding programs [370] despite the fact that, according to Stephan Griffin (the main NSF program manager for DLI funding), key research problems, notionally in the digital library realm, remain relevant:

There are qualitatively altogether new types of opportunities associated with creation, access, and use of large-scale, distributed, digital content stores that can be exploited by advanced networking and computing technologies. Better tools and more robust access frameworks are needed to realize these, and discussion and resolution of intellectual, social, and legal issues associated with selecting content and making it available must proceed in a constructive fashion [219].

Rather than fund this work under the rubric of digital libraries, the NSF has chosen to classify it as “cyberinfrastructure” research [33, 381]. Finally, and perhaps at the core of these other factors, many of the fundamental assumptions in digital libraries about how information is organized, controlled, and managed seem increasingly out-of-date. The institutionally-based DL model (i.e., a global information network composed of

---

<sup>6</sup> Some might disagree with this statement, arguing that the web is the realization of this vision. The distinction between the notion of a digital library and the general web is examined in depth throughout this dissertation.

discrete, interoperating institutional units<sup>7</sup>) and its attendant focus on professional management and control has been eclipsed by a revolutionary Web 2.0 information model that emphasizes participation rather than control and the “wisdom of crowds” [450] rather than professional guidance.

After almost two decades of active research and funding, what has happened to the notion of the digital library? Have it and its underlying concepts been assimilated into the broader notion of the web? Have its core assumptions of information control and management lost favor and perhaps become outmoded in the face of a more flexible and dynamic web information model?

In this dissertation, I explore these questions in depth, using the results of my sixteen years of digital library research as both a mirror of the trajectory of DL research and a lens for understanding that trajectory. Although my work has been primarily technical (focusing mainly on interoperability architectures, protocols, and standards), I take a broader approach here, examining digital libraries (and more generally networked information environments) as inherently *sociotechnical* [57, 318, 469] undertakings. This approach acknowledges that the evolution of new technology, such as digital libraries, is influenced by the context of pre-existing and mutually developing other technologies and, more fundamentally, by social, political, and economic norms. This perspective underlies theories and frameworks such as Social Construction of Technology (SCOT) [58], Social Shaping of Technology (SST) [355], Information

---

<sup>7</sup>The notion of an information infrastructure based on discrete institutions (e.g., libraries, museums, archives) is a carry over from traditional libraries, which because of the high transaction costs of handling physical resources, had to be proximate to their patrons [433]. The subsequent need for cooperation among these discrete units, which enabled cost saving through shared cataloging or improved services to users through interlibrary loan, led to the development of expressive interoperability standards [146, 213, 336, 452].

Ecology [378], Actor-Network Theory (ANT) [108, 325, 326], and Activity Theory [179, 180, 377].

As articulated by Van House [468], the sociotechnical perspective is particularly appropriate for analysis of information technologies (e.g., digital libraries) because of the manner in which the creation and exchange of information is so deeply embedded in almost all human *activities*. This notion of an *activity*, the action of some subject motivated by the transformation of an object toward some desired state [281], is formalized in Chapter 4, in which I describe the utility of frameworks such as Activity Theory to explain the complexity of the network relationships in information activities. As I will show, the introduction of new technology such as online (digital information) causes considerable disruption to the multiple factors that mediate information activities.

A brief summary of the content of this dissertation is as follows. Due to a variety of factors – including the primary source of funding (the NSF Directorate for Computer and Information Science and Engineering CISE), the self-selection and funder-determined selection of the members of the digital library community (i.e., primarily technical), and the strong influence of the traditional library information paradigm as a default organizational frame for the research efforts – digital library research has primarily focused on technical issues<sup>8</sup>. In general, this research took the pre-existing institutionally-based information model of the traditional library for granted, and approached the problem as the construction of new technical foundations for this existing framework. As I will describe, this produced a variety of research endeavors and outcomes thereof, such as repositories and metadata, that are technically enhanced

---

<sup>8</sup> There have been some notable exceptions to this including the work of Bishop [62], Borgman [79], and Van House [469].

facsimiles of traditional library metaphors, such as collections and catalog records<sup>9</sup>. Johansen refers to this as “horseless carriage thinking” [252]; the modeling of contemporary innovations on familiar metaphors, and simultaneously constraining them by reusing those metaphors.

I argue that the attempt to deploy an information infrastructure that essentially retrofit new technology on traditional information models failed to recognize the enormous impact of virtually ubiquitous availability to online information and the complex interaction of that availability with the broader social context. The magnitude of this impact is manifested in the World Wide Web, which emerged and evolved during roughly the same time period as contemporary digital library research<sup>10</sup>. In contrast to the relatively organized and funding agency-driven nature of DL research, the web’s growth in scale and functionality has been the result of the combined efforts of a decentralized, almost anarchistic, community of entrepreneurs and open source advocates, with indirect guidance from the standardization efforts of the World Wide Web Consortium<sup>11</sup> (W3C) and the Internet Engineering Task Force<sup>12</sup> (IETF). Constrained only by minimal technical standards and free of any historical legacy, the web has fostered a spirited atmosphere of innovation that continues to accelerate in an

---

<sup>9</sup> When examined closely, as I do in later chapters, their origins are visible under the surface in the manner of a *pentimento*, a term used in the art world. A *pentimento* indicates evidence of previous work in a painting, indicating earlier thinking of the artist, as they evolved the final work (for more information see <http://en.wikipedia.org/wiki/Pentimento>).

<sup>10</sup> The web was invented by Tim Berners-Lee in 1989 during his tenure at CERN [52, 55]. It emerged into the public attention with the release of the Mosaic browser in 1993 [359]. The web rapidly exploded in scale and impact to reach its current status as a veritable mirror of and symbol of contemporary society in all its forms. Hendler, et al. [232] go so far as to claim that “the Web is the most used and one of the most transformative applications in the history of computing, even of human communications”.

<sup>11</sup> <http://www.w3.org>

<sup>12</sup> <http://www.ietf.org>

almost viral fashion. It has organically evolved from a collection of hyperlinked documents (Web 1.0), which bore some resemblance to pre-existing information paradigms, into a dramatically different socially-shaped, dynamic, and participatory information environment (Web 2.0), which as I will show contradicts many pre-existing notions about the nature of information.

Although digital library research has continually relied on the core web technologies – HTTP for network interactions, HTML (and later XML) for document markup, and URLs for resource identification – the DL community has by-and-large treated the web as a technical phenomenon, and has generally ignored the sociotechnical nature of its development. Throughout roughly the first half of DL research (1990's) many members of the DL community, especially those connected with libraries, dismissed the web as a serious information space [220], or tried to incorporate it into the practices of the conventional library (e.g., catalog web pages) [391]. The profound changes in the nature of information in Web 2.0 have only recently impacted digital library work, and as the line between *digital* and *traditional* libraries has become increasingly blurred (virtually every library has digital content), initiatives such as Library 2.0 [113, 373] (applying Web 2.0 information principles in the library context) have recently gained popularity<sup>13</sup>.

In the end, the web and the principles of Web 2.0 have arisen as the dominant information paradigm. New and engaging collaborative applications continue to emerge, capture the public attention, and then seamlessly transition to “serious”

---

<sup>13</sup> The reader should *not* interpret comments made as being directed towards the library as *institution*. I will leave discussions about the future of the library institution, and in particular, the research library, to those more informed on that subject, many of whom are members of the library community and who, based on personal communication, share my critique of many traditional practices.

information practices. Take, for example, Twitter, which at first appeared as a curious diversion for geeks, but which has lately been adopted as an important dissemination mechanism by established news organizations (e.g., The New York Times<sup>14</sup>) trying to survive in a rapidly changing information market. And, as recent events in Iran have shown, Twitter has even emerged as an important tool for international diplomacy [319].

Applying a concept developed by Clayton Christensen [122], the web can be classified as a *disruptive innovation* vis-à-vis (digital or physical) libraries. In the manner of other disruptive technologies the early web emerged with relatively low functionality on the fringe from the dominant (library-based) information paradigm. The web of the 1990's was a potpourri of junk and quality and the limited functionality of its toolset (e.g., search engines) made it difficult to “separate the wheat from the chaff”. As a result, it served mainly popular, mass-market information activities, not yet competing with the library for “serious” information work.

The positioning of the library and the web has fundamentally changed with the emergence of Web 2.0 [390], which is truly “disruptive” in the exact sense described by Christensen. It embodies innovation and agility, and its leading applications – Google, Wikipedia, Facebook, Twitter, and the like – are the “first stop” for almost all popular information seeking and an increasing number of serious information activities, such as scholarship. The advantages of its participatory information model have been demonstrated in many domains. In contrast, the “disrupted” library, locked into a legacy information model and maintaining an infrastructure in support of that model, is steadily losing “market share” and faces an uncertain future. It has become

---

<sup>14</sup> <http://twitter.com/nytimes>

the subject of studies that examine its survivability and the manner in which it needs to be reconceived in the Web 2.0 era [7, 113]. Ironically, these thoughts on how to reinvent the library abandon many of its core notions (e.g., cataloging) and adopt the information principles derived from Web 2.0 (e.g., collaboration).

What then, of the future of digital libraries and digital library research? Rather than presumptuously asserting some answer, I am working with my respected colleagues<sup>15</sup> many of who are asking the same question to articulate a community answer. I hope this dissertation provides some valuable insight as we examine those questions. As an alternative to an individual answer, I offer the following additional thoughts.

The name “digital libraries” has been controversial from the beginning. In a 1992 workshop, the well-known futurist Esther Dyson said; “What is the digital library? That term smacks of “filmed play,” “horseless carriage,” and the like. The digital library will be less like a library than we think, and more like itself” [191]. Even the form of the term has been controversial: the distinction between the set of “digital libraries” and the global “digital library” was, according to Bill Arms (personal communication), an active issue of discussion in the early days of DL funding and in the naming of the research program “Digital *Libraries* Initiative” itself.

Other community members noted from the beginning the need to distinguish digital library work from its pre-existing namesake. Indeed, the intent of the funders in their use of the name was to endow the new online environment with established library attributes such as trust and integrity, while encouraging innovation. Clifford Lynch, a recognized thought leader of the information community, stressed early on the need

---

<sup>15</sup> As Program Committee Chair for the 2010 Joint Conference on Digital Libraries, this is precisely what I am trying to do with the theme “Digital Libraries - 10 years past, 10 years forward, a 2020 Vision” (see <http://www.jcdl2010.org>).

for digital libraries to move beyond “simple information access”, characteristic of traditional libraries, to “environments for actually doing active work”. He noted that “the more they [digital libraries] move in this direction [collaborative work environments], the further they move away from the traditions of the libraries that are funding and developing many of them” [350].

Arguably, then, the notion of digital libraries could continue, retaining positive attributes of libraries while shedding some of the traditional constraints (after all, we still “dial” phone numbers even on our touch-screen mobile phones). But as linguists such as Lakoff [317] and Nunberg [387] state, names and the images they evoke are powerful devices. As I have already mentioned, they affect the manner in which the participants in an effort (in this case the digital library research community) frame their work and the products of it.

They also affect the perceptions that the external communities, the “users”, have of that work, which I refer to as “reverse horseless carriage thinking”. Digital libraries are frequently associated with notions of “traditional”, or “old-fashioned”. Y.T. Chien, the program officer at the NSF perhaps most responsible for the initiation of digital library funding, reflected on this in a 2004 paper describing the future challenges to digital library research [119]:

First and foremost [among the challenges to DL innovation] is the ill-formed public perception towards digital libraries. This is perhaps the most serious roadblock for DL’s future. The general public by and large continues to view a digital library as the electronic version of the traditional library – where you get to use books and other materials in electronic forms either online or from the local library, for free. The broader vision for the DL circa 1994 has hardly had much effect on that outdated perception.

Perhaps then it may be the appropriate time for a form of “rebranding” both as a symbol of changed context and new internal direction. A number of influential

members of the web community have called for the creation of a new scholarly field called “web science” [232]. Their vision of this field is notably interdisciplinary, recognizing the full sociotechnical impact of online information across traditional scholarly and societal boundaries. This is an attractive notion, but I have some hesitation of signing onto a name linked to an instance (“web”) rather than a concept (although in personal communication advocates of web science have argued that the “web” is indeed a concept). For now, I will leave the name question open and part of the broader community discussion.

---

The remainder of this dissertation is structured as follows. The first part, consisting of Chapter 2 through Chapter 4, analyzes the disruption of digital libraries by the web from three different perspectives.

Chapter 2 uses a historical approach. It describes the background behind the choice of the term “digital library”, and the manner in which that decision to link the emerging research area with a traditional notion (i.e., the library) has affected the trajectory of digital library work. Finally, it positions that work alongside the evolution of the web, which as mentioned was concurrent with the modern digital library initiative.

Chapter 3 uses a conceptual approach. It employs the notion of a *meme*, which captures the sociotechnical nature of the web and libraries, to deconstruct the nature of both entities into core principles, capabilities, and technologies. This deconstruction reveals the nature of the incompatibilities between them and the causes of the disruption.

Chapter 4 uses a network-centered approach. It integrates the analysis of the previous chapters into a number of the frameworks for analyzing technological change and

disruption that originate from the fields of Science, Technology, and Society (STS) and Workplace Studies. It focuses on one framework in particular, Activity Theory, as a mechanism for understanding the activity systems underlying scholarly research and publication in the library, Web 1.0, and Web 2.0 contexts, and for revealing the contradictions between those individual activity systems.

The dissertation then continues with Chapter 5 that summarizes work related to the four areas included in this dissertation: digital library interoperability architecture, retrospectives on digital library research, the Web 1.0 to 2.0 transition, and digital libraries as sociotechnical systems.

Chapter 6 introduces the second part of the dissertation, consisting of six chapters, each of which is constructed around a specific result from my sixteen years of digital library infrastructure research. The overall goal is to use this span of work to illustrate concepts presented in the initial chapters – the influence of the library meme on the nature the technical work within digital libraries and the efforts to break away from the constraints of that meme as the web information model increasingly diverged from the traditional library model. The core of each chapter is one of my published papers, which is preceded by a preface that positions that paper and work in the context of this dissertation. The subjects of these subsequent six chapters are as follows.

Chapter 7 describes the Dienst digital library architecture and its deployment in the Networked Computer Science Technical Reference Library (NCSTRL), which illustrate classic digital library components including metadata, repositories, portals, compound digital objects, and federated search.

Chapter 8 describes metadata in two forms: Dublin Core and ABC/Harmony. The former demonstrates traditional, library-based bibliographic principles, while the latter

shows the effort to accommodate metadata to the changed web information environment.

Chapter 9 describes the Fedora digital object and repository architecture, a state-of-the-art system that extends traditional library-based content management principles with service-oriented architecture and semantic web concepts.

Chapter 10 describes metadata harvesting (via the Open Archives Initiative Protocol for Metadata Harvesting – OAI-PMH) used as the foundation for the National Science Digital Library (NSDL), demonstrating problems that arise when library-based metadata practices are deployed in a distributed digital library.

Chapter 11 describes a new architecture for the NSDL that is resource rather than metadata-centric and that encodes semantic relationships and context among resources.

Chapter 12 describes the Open Archives Initiative Object Reuse and Exchange (OAI-ORE), a standard for modeling compound (aggregated) objects using web architecture and semantic web concepts, and for encoding and identifying those objects in common machine-readable formats. This work demonstrates integration of digital library content principles and the web.

Chapter 14 takes a look forward to understand the manner in which experience with digital libraries can inform recently-funded cyberinfrastructure projects. The particular focus is the Data Conservancy project, of which I am the Cornell PI, which is an NSF-funded 10-year, \$20 million project to investigate data-centric eScience. Similar to digital libraries, the success of this and similar projects depends on a subtle integration of technology with social, economic, and political factors and an awareness of how the information and scholarly context in which the project exists is changing.

Chapter 14 concludes the dissertation with some wrap up remarks.

---

## **Methods**

A variety of methods were used in the research that is reported in this dissertation. The technical work described in Chapter 7 through Chapter 12 was the result of extensive community participation in the design, prototyping, and eventual deployment of the technical results. This is especially true for the work carried out under the auspices of the Open Archives Initiative, the Protocol for Metadata Harvesting and Object Reuse and Exchange. Both of these projects and the standards that resulted from them involved the formation and management of international, cross community technical and advisory committees, which were closely involved in the evaluation and eventual testing of alpha and beta versions of the work. This close participation of the target communities was vital to the drafting of standards and products that demonstrated them that eventually met the needs of a broad range of deployment scenarios.

Although they did not engage formal advisory and technical committees, the other technical projects reported in his dissertation - such as Dienst, the ABC/harmony work, and the NSDL work - were subject to widespread and long-term community exposure as a result of their global deployment. This deployment in real production scenarios played a significant role in their eventual refinement and validation.

The analysis in Chapter 2 and Chapter 3 of digital library history, the role of various communities in that history such as libraries, funding agencies, and the computer science community, and the relationship of that history to the history of the web is based on my long-term, and prominent role in that community. I have been funded by and the principal investigator of digital library grant funding from DARPA and the

NSF, as well as private funding from the Mellon foundation, since the beginning of my career in this area in 1994. As described elsewhere in this dissertation, that time period more or less corresponds to the entire history of modern digital library research. In addition, I have for over 10 years served on the program committees of almost all international digital library conferences, including the ACM/IEEE Joint Digital Library Conference. Notably, I am Program Chair of this Conference in 2010. I have also participated in and chaired a number of NSF and privately-funded digital library and related topic workshops throughout this time span. I have spoken internationally on digital library, eScience, cybreinfrastructure, and related topics throughout this period. Finally, I have taught an upper-level undergraduate/graduate course in the Information Science program at Cornell University on Web Information Systems, which covers digital libraries and the Semantic Web, for the past five years. This extensive, prominent, and long-term involvement in this community has given me a unique and intimate perspective on the research activity within it, the politics and process of its funding and organization, and the nature of its successes and failures, and is the foundation of the analysis here.

The analysis of digital libraries as sociotechnical systems and the use of Activity Theory, activity system diagrams, actor-network theory, and related frameworks in Chapter 4 is the result of a standard literature search in these areas. Throughout this literature search, I focused mainly on the application of these frameworks to information systems in general and to digital library applications in particular. As I note in the related work section, the majority of work in this area has been focused on the evaluation of particular digital library applications, rather than on the notion of digital libraries and the information model that they manifest as a whole. Based on my investigations, the use of them in this dissertation for this type of overall, comparative analysis is unique to this dissertation.

The use of memes in Chapter 3 as an analytical tool is similarly based on a standard literature search. As noted also in the related work section, meme maps as an illustrative tool have mainly been used in informal, business applications, and the use of them for comparative analysis as employed in this dissertation is original.

Finally, the analysis of the impact of this work on future cyberinfrastructure projects, as reported in Chapter 13, is the result of interviews with prominent colleagues who have played a major role in those still nascent projects. These interviews were carried out over the phone and the particular people involved were Christine Borgman, Presidential Chair & Professor of Information Studies at University of California Los Angeles, Sayeed Choudhury, Associate Dean of Libraries and the Hodson Director of the Digital Research and Curation Center at Johns Hopkins, Mary Marlino, Director of the National Center for Atmospheric Research ( NCAR) Library, and Carole Palmer, Professor and Director of CIRSS -- Center for Informatics Research in Science & Scholarship – at the University of Illinois at Urbana-Champaign.