

LOST IDENTITY:
THE ASSIMILATION OF DIGITAL LIBRARIES INTO THE WEB

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by
Carl Jay Lagoze
February 2010

© 2010 Carl Jay Lagoze

LOST IDENTITY:
THE ASSIMILATION OF DIGITAL LIBRARIES INTO THE WEB

Carl Jay Lagoze, Ph. D.

Cornell University 2010

The idea of Digital Libraries emerged in the early 1990s from a vision of a “library of the future”, without walls and open 24 hours a day. These digital libraries would leverage the substantial investments of federal funding in the Internet and advanced computing for the benefit of the entire population. The world’s knowledge would be a key press away for everyone no matter where their location. This vision led to substantial levels of funding from federal agencies, foundations, and other organizations for research into fundamental technical problems related to networked information and deployment of the results of this research in numerous digital library applications. The result was a number of exciting and influential technical innovations.

But, the attempt to transplant the library to the online environment met with some unexpected obstacles. The funding agencies and many of the members of the digital library research community mainly focused on the technical issues related to online information. In general, they assumed that the new technology would be applied in a largely traditional (library) context, and largely ignored the profound social, economic, cultural, and political impact of turning “books (and other information resources) into bytes”. The extent of this impact was demonstrated by the concurrent evolution of the World Wide Web, a networked information system not bound by legacy institutional conventions and practices or funding agency mandates and, therefore, able to organically evolve in response to the profoundly democratizing effect of putting

information online. This has provided the context for the recent revolution in the web known as Web 2.0, a participatory information environment that contradicts most of the core assumptions of the traditional library information environment. The overwhelming adoption of the Web 2.0 model for both popular culture and serious information exchange and the increased evidence of the efficacy of this model for activities such as learning and scholarship call into question the viability of the library information model and the digital libraries that were meant to instantiate that model online.

In this dissertation I examine the almost two decade history of digital library research and analyze the relevance of the library information model, or meme, in relationship to the transformative Web 2.0 meme. I use my research results in digital library infrastructure and technology over this period as both a lens for viewing this historical relationship and a mirror for revealing its various facets. This analysis is particularly relevant as I, and fellow members of the research community, begin to engage in large-scale cyberinfrastructure projects that need to move beyond the largely technical focus of earlier digital library initiatives and recognize the sociotechnical nature of the work that lies ahead.

BIBLIOGRAPHIC SKETCH

Carl Lagoze was born on April 17, 1953 in New York City, the second son, after his brother Howard, of Eli and Rita Lagoze. After a brief interlude in Pittsburgh, the family settled in Cheltenham PA, a suburb bordering Philadelphia. Following his graduation from Cheltenham High School, which he attended during the turbulent and thrilling late 1960's, he matriculated at Cornell University in what then seemed a foreign land on the edge of the arctic. He graduated in 1975 with a Bachelors of Science in Urban Studies, gaining much insight into the complexity of urban environments but, as a side-effect, learning about programming from a couple of computer science courses and independent study on an urban simulation gaming project with Douglas Van Houweling (then a faculty member in Government at Cornell, and now CEO of Internet2).

Intrigued by programming and loving Ithaca for its unique alternative culture of the mid 1970's and magnificent countryside where he could hike, ski, and bike ride to his heart's delight, he accepted an offer to stay on the project as a research programmer. During the late 1970's and through the mid-1980's he stayed in Ithaca, progressing through a number of programming and system administration positions at Cornell, simultaneously taking courses in the CS department to get some academic depth in the area that increasingly felt like his career. Convinced that more formal depth in this career choice made sense, he took a break from Ithaca in 1986-1987 to gain his Masters in Software Engineering at the Wang Institute of Boston University, a unique educational experiment in Tyngsboro, MA. He returned from that to a research programming position in the CS department at Cornell with Tim Teitelbaum and at, GammaTech, the spin-off of Teitelbaum's and Tom Reps' program analysis and code

synthesis research, during which he began to appreciate the rigor and excitement of research, and took a summer off to fulfill his dream of bicycling across the U.S.

None of this excitement, however, matched that of the birth of his daughter, Lucy Lagoze, in 1993. Little did he know at the time of her entrance to the world on November 23, eyes wide open, of the fantastic person that would emerge and the joy and wisdom he and others would experience from her.

Increasingly intrigued by computer science and the potential of information technologies to effect meaningful and beneficial change in society, Carl left Grammatech in 1992 for a IT position at Mann Library at Cornell, which was a leader among its peers in its approach to the dawning digital, networked information era. Fascinated by emerging Internet technologies such as Gopher, his eyes were opened by a demonstration at a conference by a student from Illinois, Mark Andreessen, of a client called Mosaic to something called the World Wide Web. He returned to the library enthusiastic to change the world of libraries based on web technology.

Exciting as the library was, the ever-impatient Carl found its legacy a little too burdensome. In early 1994 he accepted an offer from Dean Krafft to work with him and Jim Davis in the CS department at Cornell on a new digital library project called the Computer Science Technical Reports (CSTR) project. Mentored by the fascinating and brilliant Davis, Carl quickly began to find his legs as a researcher. He first gained national and international recognition through the CSTR project, and then received his first major grant in DLI-2 for Project Prism. During this time he established the Digital Library Research Group (DLRG) in Cornell CS, one of the important precursors to the Information Science Program. This grant was followed by a number of other large grants, which funded the results described in this dissertation. In addition, through his appointment as Senior Research Associate in Computing and

Information Science at Cornell he has taught courses in web design and web architecture and served on the committees of several Ph.D. students.

Throughout these exciting years of research Carl has benefitted from collaboration with a number of distinguished colleagues, many of who are identified in the acknowledgements section of this dissertation. However, his most important collaboration in terms of professional and personal enrichment has been with Sandy Payette, whom he married in 2006 and who continues to fill his and Lucy's life with wisdom, joy, and meaning.

For Lucy and Sandy:

From whom I am continually learning

ACKNOWLEDGMENTS

The work reported in this dissertation extends back over 15 years, a period of many rich and rewarding collaborations with some wonderful people. I owe all of them a tremendous debt of gratitude for the manner in which they have enriched my intellectual life, shared friendship, and supported me throughout these years. While I appreciate the opportunity to thank those people here, I fear errors of omission and I apologize in advance for those not mentioned here.

First and foremost, I must give my most heartfelt thanks and appreciation for the two people responsible for my entry and advancement in the field of digital libraries: Dean Krafft and Jim Davis. If it were not for a surprise phone call at my desk in Mann Library from Dean in early 1994 offering me a position in the newly formed CSTR project, and the valuable guidance from Jim in the early months of my research career, I don't think I would be where I am today. Thank you to both of you.

Thanks also to the three people most instrumental in encouraging me to proceed along this PhD route and pushing aside the bureaucratic barriers in the path of my rather unconventional "conformance" with Cornell graduate school residency and credit requirements: Bob Constable, Dan Huttenlocher, and the chair of my committee Geri Gay. The wonderful realization that all three of you distinguished people really believed in me and thought that I deserved this degree was a tremendous incentive to finish this through my months of hard work.

A number of other colleagues deserve special recognition. Herbert van de Sompel, with whom I collaborated for years in Open Archives Initiatives projects has been an important influence over my work in both his criticism and praise and has tolerated

my sometimes tempestuous personality when I get overly impassioned with my own ideas. Herbert has also been a wonderful friend through some difficult and trying times. Jane Hunter, who is one of the most productive, caring, and funny people that I have worked with, has contributed to my work in countless ways. Clifford Lynch has for years given sage advice and has been an ardent supporter of many of the projects I've been engaged in. Other notable people over the years include, in no certain order, Simeon Warner, Bill Arms, Stu Weibel, Tom Baker, Michael Nelson, Andreas Paepcke, and Paul Ginsparg. Also, many thanks to Rosemary Adessa who in her role as unit manager of the information science program at Cornell has provided endless advice and a tireless shoulder in support of my many trials and tribulations.

There are also a host of people outside of work, in my personal life, whose support was instrumental towards my finishing this task. First, chronologically and in importance, is my mother Rita Lagoze, who indeed was my first collaborator – she pushed and I headed for the light. Thanks Mom, you finally got “your son the doctor”. My enjoyment of life over my many years in Ithaca and my ability to face a lot of hard work with a sense of humor has been enhanced by many, many close and dear friends including Jeff Furman, Sara Hess, Marty Kaminsky, Oya Rieger, and Robert Rieger. Thanks also to Alan, who with careful listening and valuable advice helped me navigate the murky waters of life.

The experience of parenting my dearest daughter Lucy Lagoze over the past 15 years has given meaning to my life that cannot be matched by my career achievements. Thank you Lucy for being a constant source of joy, pride, and inspiration. I couldn't have done it without you.

Saving the most important acknowledgment for the end, I owe so much to my soul mate, best friend, intellectual inspiration, pillar of strength, most valuable critic, and

wife Sandy Payette. As you have heard me say over and over again, I lead a blessed life, and the completion of this dissertation is further proof of that. But the most blessed event in my life is the time I met you and the continual blessing of sharing my life with you fills me with constant joy.

TABLE OF CONTENTS

BIBLIOGRAPHIC SKETCH	iii
DEDICATION.....	vi
ACKNOWLEDGMENTS.....	vii
TABLE OF CONTENTS.....	x
LIST OF FIGURES.....	xiii
LIST OF TABLES.....	xv
Lost Identity	1
Digital Libraries and the Web: Origins, Impact, and Evolution	18
Origins of the “digital library”	18
Influence of the library on digital library technology	25
Coexistence of digital libraries and the web	33
Digital libraries and the evolving web	34
Chapter Wrap-up	38
A Meme-based Analysis of Digital Libraries and the Web	41
The library meme	44
The library made digital	51
The web as a technical artifact	54
The Web 1.0 meme	58
The Web 2.0 meme	61
Conflict among the memes.....	65
Chapter Wrap-up	76
An Network-Centric Approach for Examining Disruption.....	79
Actor-Network Theory	81
Information ecologies.....	84
Activity Theory	86
Chapter Wrap-up	100

Review of Related Work	101
Technologies for interoperability in networked information systems.....	101
Historical overviews of digital library research	115
Impact of the Web 1.0 to Web 2.0 transition	118
Digital libraries as sociotechnical systems	121
Introduction to Chapters 7-12.....	126
Making Global Digital Libraries Work.....	130
Preface	130
Acknowledgments	133
Introduction	134
NCSTRL – The test bed for a globally distributed digital library.....	138
Dienst architecture.....	138
The evolution of a distributed digital library: early experience	146
Connectivity regions and distributed collection service.....	150
Conclusions	156
Accommodating Simplicity and Complexity in Metadata.....	158
Preface	158
Acknowledgements	160
Realities for all occasions.....	161
A world of document-like objects	166
Confounding the simple model	171
Agents of change	177
Is it all worth it?.....	182
An Architecture for Complex Objects and their Relationships.....	184
Preface	184
Acknowledgments	185
Introduction	186
Background	188
Fedora model for complex objects	191
Relationships in Fedora	201
Results	208
Conclusion.....	212
Metadata Aggregation and “Automated Digital Libraries”	214
Preface	214
Acknowledgments	217
Introduction	218
Related work.....	223
Metadata providers	224
Provider management.....	227

Ingest processing	233
Metadata storage and OAI exposure	236
Search	240
Conclusion	244
Representing Contextualized Information in the NSDL	246
Preface	246
Acknowledgements	247
Introduction	248
Related work.....	250
The need for context and reuse.....	252
A suite of contextualized NSDL services.....	254
Design and information model	256
Results from implementation of the NSDL data repository	258
Conclusions	263
A Web-Based Resource Model for Scholarship 2.0.....	265
Preface	265
Introduction	267
The architecture of the World Wide Web	270
Scholarly documents – Pre-Web to Web 2.0	274
OAI-ORE: Identifying and describing compound objects	280
Deployment, experimentation, and implementation	287
Related work.....	295
Conclusion.....	298
Lessons for Cyberinfrastructure Projects.....	300
Understanding the complexity of infrastructure.....	302
Recognizing community diversity.....	304
The danger of the “seduction of the known”.....	306
Understanding the difference between text and data.....	309
Rapid prototyping and moving targets	310
Concluding Remarks and Observations.....	312
REFERENCES	316

LIST OF FIGURES

Figure 1 - Expansion of web functionality (from [130])	37
Figure 2 - Meme map	43
Figure 3 - Library meme map	45
Figure 4 - Library information flow	47
Figure 5 - Mapping of concepts to external artifacts in traditional library	52
Figure 6 - Digital library meme map	53
Figure 7 - Mapping of concepts to external artifacts in digital library	54
Figure 8 - Relationship of basic web architecture components [246]	56
Figure 9 - The web graph	57
Figure 10 - Web 1.0 meme map	59
Figure 11 - Web 2.0 Meme Map	63
Figure 12 - Web 2.0 information flow	70
Figure 13 - Simple mediation in an activity system	88
Figure 14 - Triple mediation in an activity system	89
Figure 15 - Library-centered research	91
Figure 16 - Pre-web publication	92
Figure 17 - Web 1.0 research	94
Figure 18 - Web 1.0 publication	95
Figure 19 - Internal disruption to an activity system	96
Figure 20 - Web 1.0 disruption	97
Figure 21 - Web 2.0 scholarly communication	98
Figure 22 - Web 2.0 disruption	99
Figure 23 - Research project timeline	126
Figure 24 - Dienst Services	140
Figure 25 - Dienst service interactions	145
Figure 26 - Simple distributed search with server failure	147
Figure 27 - Primary and secondary index servers	149
Figure 28 - Connectivity regions	151
Figure 29 - Interactions of CCS, RCS, and user interface server	154
Figure 30 - Mixing information from multiple communities	163
Figure 31 - Multiple views of the same content	164
Figure 32 - Flattening complex reality	169
Figure 33 - Uncontrolled qualification vs. interoperability	176
Figure 34 - A closer look at resource, entities, and their relationships	178
Figure 35 - Event aware descriptive data model	181
Figure 36 - Representational view of Fedora objects	193
Figure 37 - Fedora object with PID, properties, and datastreams	194
Figure 38 - Properties of a datastream component	195
Figure 39 - Fedora object with disseminator added	196

Figure 40 - Disseminators establish relationships to service definition objects.....	197
Figure 41 - Integrity datastreams - relationships, policy, and audit trail.....	200
Figure 42 - NSDL network overlay example	212
Figure 43 - NSDL metadata flow	221
Figure 44 - NSDL harvesting failure rate	231
Figure 45 - Harvest failure categories	232
Figure 46 - Modeling an aggregation	258
Figure 47 - Scholarship 2.0 meme map.....	268
Figure 48 - Identifier, resource, and representation (from [246])	271
Figure 49 - Web graph.....	272
Figure 50 - RDF triples and graph representation	273
Figure 51 - Expressing types in RDF	274
Figure 52 - Linked data cloud	275
Figure 53 - Aggregation of evidence of scholarship	276
Figure 54 - Pre-Web Scholarly Publication.....	277
Figure 55 - HTML splash page.	278
Figure 56 - Web graph with embedded compound object.	279
Figure 57 - Identification and description of an Aggregation.	281
Figure 58 - A Resource Map and Aggregation with 3 Aggregated Resources	283
Figure 59 - Citing a Resource in the context of an Aggregation.....	285
Figure 60 - Resource Map discovery from an Aggregation using Cool URIs	287
Figure 61 - JSTOR collection mapped to the OAI-ORE data model	289
Figure 62 - Screenshots of Word OAI-ORE plug-in.....	291
Figure 63 - The splash page dynamically rendered from Resource Map.....	293
Figure 64 - Activity system	313

LIST OF TABLES

Table 1 - Comparison of digital library and web architectures	25
Table 2 - Essential library elements compared.....	66
Table 3 - Employing the "dumb-down" principle	172
Table 4 - Example relations datastream	204
Table 5 - Object-representation relationship	205
Table 6 - Data type properties	205
Table 7 - Sample RDF query using iTQL	206
Table 8 - A query to build an OAI response	207
Table 9 - The query response as triples	208

Chapter 1

Lost Identity

“The future ain’t what it used to be”

- Yogi Berra¹

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 20th 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]

¹ 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 arXiv² and others [226, 276, 280], institutional repositories [143, 253, 438], and large archiving efforts such as Brewster Kahle's Internet Archive³. 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

² <http://arxiv.org>

³ <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 fabric⁴ [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 agencies⁵ 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

⁴ 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].

⁵ 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 emerged⁶. 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

⁶ 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 units⁷) 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

⁷The 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 issues⁸. 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

⁸ 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 records⁹. 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 research¹⁰. 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 Consortium¹¹ (W3C) and the Internet Engineering Task Force¹² (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

⁹ 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>).

¹⁰ 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”.

¹¹ <http://www.w3.org>

¹² <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 popularity¹³.

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”

¹³ 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 Times¹⁴) 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

¹⁴ <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 colleagues¹⁵ 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

¹⁵ 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.

Chapter 2

Digital Libraries and the Web: Origins, Impact, and Evolution

In the previous chapter I claimed that the choice of the term “digital library” influenced the information model adopted by the digital library research community and shaped the technical components of digital library applications that realize that model. I further claimed that this shaping has had an impact on the coexistence of digital libraries with the remainder of the web information space.

This chapter explores those issues in greater detail. It begins by describing the motivations underlying the reuse of the term “library” to describe the new digital information research area. As will be described, those motivations reflected the expedient interests of each participant community and their respective definitions of a library. It next describes the influence of that term on the architectures and applications that were produced by the digital library research program over the subsequent years. The chapter continues with an explanation of how the architecture of those digital library applications impacts the technical coexistence of digital libraries with the mainstream web. Finally, the chapter describes how the evolution of the web from Web 1.0 to Web 2.0 has not only increased the technical incompatibilities between the two information environments, but has led to a fundamental conceptual difference in their information models.

Origins of the “digital library”

The decision in the early 1990’s to extend the notion of the library forward into the emerging digital information context was the result of the collective and distinctive

assumptions of three stakeholder communities: the funders, the technology-focused researchers, and the practitioner library community [82]. Each community responded to the opportunities offered by emerging networked computing technologies in a unique, opportunistic (and sometimes myopic) fashion. While they all agreed that “digital libraries” was an appropriate term for the new endeavor, they each had a different idea of the meaning of the term and different allegiances to the components of what they considered a library.

The following sections demonstrate this “interpretive flexibility” [468] by describing the different meanings attributed to digital libraries by the three major communities involved in the research effort.

Perspective of the digital library funders

The primary funders of digital library research in the U.S. were the NSF, within the Directorate for Computer and Information Science (CISE), DARPA, NASA, and NIH, with lesser contributions from NEH, IMLS (Institute of Museum and Library Services), and the Library of Congress. The notable characteristic of all the primary funders is their focus on technology-oriented science, in contrast to social science, humanities, or arts. This focus is reflected in a research program that funded mainly core computer science and its applications to networked information, with very little attention to network information as a sociotechnical phenomenon. In fact, the first phase of Digital Library Initiative (DLI) funding [379] was exclusively technical. This was moderated somewhat in the second phase [380]. Influenced by the results of a 1996 NSF workshop that called for increased research on the social aspects of digital libraries [83], the NSF included social science research in the DLI-2 solicitation.

An examination of documents published early in the digital library effort reveals the underlying assumptions and biases of the lead agencies that shaped the nature of the funding programs and their vision of the digital libraries that would emerge from it.

Although it appeared in a visionary 1988 document from Kahn and Cerf, the term “digital libraries” was introduced into the national research agenda in February 1994 in a report from a task force on Information Infrastructure Technologies and Applications (IITA) [257]. This report was commissioned by the newly funded High Performance Computing and Communications program, which was formed to leverage advances in computing and networking for the general social benefit [109].

The report defines digital libraries as follows:

[Digital libraries are] both technologies and applications which will lead to significant advances in the generation, storage, and use of digital information of different kinds across high speed networks. *A digital library is a knowledge center without walls, open 24 hours a day and accessible by way of a network.* Research areas range from advanced mass storage, online capture of multimedia data, intelligent filtering, knowledge navigation, effective user interfaces, system integration, to prototyping and technology demonstration. [257] (emphasis added)

The list of research areas enumerated in this statement is notable for its omission of the implications of eliminating the “walls”, or a being open “24 hours a day.” Clearly, the impression of the task force, or at least the only concern, was that the transfer of information to an online form raised only technical issues and that the larger social issues raised by this transfer were either inconsequential, unforeseen, or not worthy of study.

Another report from the same era, authored by Gladney and Fox, demonstrates prevailing thinking of the time that perhaps underlies this decision by the funders to focus only on technical issues.

The concept "library" has been refined over several centuries. It would be injudicious to depart from what people expect merely because a digital

service is replacing a material one. Except where explicit reasons suggest an improvement that is easily explained to ordinary users (e.g., in query services), library services should implement a familiar model¹⁶. [211]

Implicit in this text is the assumption that the nature of the institution and the information model it entails should remain as a stable overlay on a changed technical foundation. A digital library should, by nature, imply the same notions of integrity, trust, and quality, historically associated with libraries. Books might turn into bits, the catalog might become an online database, and shelves might turn into repositories, but the values, structures, and practices of the “institution” should be based on a “familiar model.”

Even as the web continue to grow in importance and scale, the notion that digital libraries were the focus for “serious” information-oriented activities persisted. The web was relegated to a more lowbrow status – an unfiltered mishmash of questionable and frequently objectionable content. For example, a 2001 (U.S.) President's Information Technology Advisory Committee report called the web a “rudimentary” information environment that “only hint[s] at the future of digital libraries” [409].

Even if the funding agencies had decided that the broader implications of the transfer of information to the online environment deserved investigation, it is doubtful whether they were structurally configured to handle such investigations. In a retrospective on the Digital Libraries Initiatives, Griffin [219] takes note of the problem that agencies such as the NSF have with research that is by nature long-term: “The program funding

¹⁶ It is interesting to contrast this quotation with one from the position statement of noted futurist Esther Dyson in a workshop at roughly the same time: ‘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]’

models did not work optimally, particularly for the mid-size, longer-term, interdisciplinary research and test bed projects.”

Perspective of the computer science research community

The computer science research community had every incentive to follow the funding agencies in this selective interpretation of digital library research. The DL initiatives were a new and relatively large stream of funding for extending their pre-existing database and information retrieval research into a new application area [82]. As stated by Paepcke, et al.: “[The computer scientists] could see, or at least imagine, *how current library functions would be moved forward by an injection of computing insight*” (emphasis added).

The computer scientists who dominated DL research had little interest in examining the nature of “current library functions” or in understanding how the “injection of computing insight” might affect the foundations of these functions. The library was really only a convenient platform for technically focused work.

Indeed as Paepcke, et al. note the computer science researchers had little patience for the less technically manageable aspects, and “nagging downsides” of digital library research. For example, issues related to copyright and intellectual property were perceived as an annoyance that interfered with work on more interesting technical problems. Furthermore, the work often required collaboration with librarians who seemed overly focused on metadata “that the computer scientists felt would be replaceable by just another clever search algorithm improvement” [398].

In hindsight, the attraction of the computer science research community to the field of digital libraries was really not based on special allegiance to the library notion, but was just a case of following the funding. When the funding disappeared and it became

obvious that the web was a more attractive, and less restrictive environment for studying and exercising emerging computer science techniques such as machine learning, many of the former prominent members of the digital library community disappeared [398].

Perspective of the library community

The “real” librarians, those with over a century-long tradition collecting, curating, and preserving books and other materials, entered the realm of digital libraries with a considerably more institutionally-focused definition of the library and vision of what the digital library would look like. From their perspective the library, as an institution, had successfully managed previous transitions to new media (the transition of the printed form from scrolls to the codex book to the printed book [117, 389], the inclusion of recordings, etc.) and had a track record of incorporating new technology into established practices, such as the computer-based catalog [78]. The “digital” library would be just another library and through all, the venerable institution would prevail:

The functions of the librarian have always been to select the material that his constituents will require; to catalog it so that those who would use it can know what is available and where it is; and to preserve it so that both contemporary readers and those who will follow will be able to use it...none of these tasks will disappear with the emergence of the electronic library. Somebody will have to perform them: if not the librarian, then his replacement. The anarchy of the Internet may be daunting for the neophyte, but it differs little from the bibliographic chaos that is the result of five and a half centuries of the printing press. [332]

Because of this allegiance to the institutional basis of the library and the belief that it was a necessary component of a useful information environment, librarians vigorously resisted the encroachment of the web on the domain formally dominated by the library. As noted by Paepcke, et al. “For librarians the intrusion of the web into the work on digital libraries was much more difficult to integrate” [398]. Initially, they

were largely dismissive and disdainful of the web as a serious information space, declaring that the “web is not a library” [220] and likening it to a bookstore in which “the entire stock is just piled up in the middle of the floor” [140]. As the amount of valuable content increased on the web, they responded with efforts to fold the web into standard operations, such as cataloging [391]. While these efforts to catalog the web were ultimately abandoned, they demonstrate how persistent traditional practices can be even in the face of a rapidly changing technical landscape.

In summary, the application of the library concept to the uncharted and unruly context of networked information reveals the distinctly narrow and flawed assumptions of the three parties responsible for its origin and use. The funding agencies mistakenly assumed that they could fund (and shape) the development of a new information infrastructure as a mainly technical endeavor. The computer scientists followed suit by framing digital libraries by-and-large as applications of familiar distributed database problems in which “predictable, repeatable ... access and retrieval is a prime value. [398]” In fact, the unpredictability of the web and its seemingly autonomous dynamism has not only affected our perceptions of information use and management, but it has had a far-reaching effect on computer science shifting it from its deterministic, algorithmic foundations to a more probabilistic and socially-oriented focus [268]. Finally, the librarians assumed that they could safely wrap radically new technology and traditional organizational values and structures in the same embrace.

In combination, these flawed assumptions lead to a research area that by and large treads the middle ground. At a fine granular level, it produced a number of interesting research results and applications of those results. But, at the higher level, it failed to explore the more far-reaching questions of how putting information online and giving people power over their information might change the nature of the information and

the way people use it. It is useful in closing this section to quote Agre [17] who, in his essay about “Information and institutional change”, spoke of the dangers of naively mixing historical forms with innovations:

A concept of “library” that is too fully rooted in past historical forms will make innovation impossible, but a superficial concept of “library” that draws out only a few aspects of those past historical forms (for example, a library as a big container of documents) will pass over phenomena whose absence in a newly designed system may be fatal. The middle ground between the maximal and simplistic conceptions of “library” is enormous and is not easily mapped.

Influence of the library on digital library technology

Table 1 - Comparison of digital library and web architectures

	Digital Libraries	Web Architecture
Core Architecture	Repository-centric	Resource-centric
User Model	Portal Searching	Browsing
Content Model	Digital Objects	Resources
Indexing Model	Surrogates	Full-text and links
Identification	Persistent IDs	URIs
Federation Model	Federated Search, Metadata Harvesting	Centralized Indexing

The previous section described the set of assumptions about libraries and networked information that led to the choice of the term “digital libraries.” This section describes the manner in which that term and the presumptions underlying it have affected the nature of the technical artifacts produced by digital library research. This effect is reflected in both the overall architectural framework and on the individual architectural components of that framework. The contents of this section are summarized in Table 1.

Core Architecture: Repository-centric versus Resource-centric

Digital library systems are by and large based on the notion of the institutionally-managed *repository* as the central architectural entity. The repository acts as the container for storage of and access to “digital objects” [255], the content “within” the library. In this manner, the repository is a virtual boundary defining the locus of institutional management, curation, and preservation of the contained digital objects. This virtual boundary is the functional equivalent of the physical boundary in the “bricks and mortar” library in which the physical structure defines the limits of library curation and stewardship of the information resources within it.

In contrast, the *resource* is the central entity in the web architecture [246]. Uniquely identified resources are the nodes in a virtual directed graph, in which the edges are the hyperlinks that connect resources. Notably absent from this graph model is the notion of containment or location. There is no first-class entity that corresponds to the repository in digital architecture. Although repositories are sometimes compared to websites, the comparison is incorrect due to the nature of the latter. A website is an ambiguously defined, second-class technical artifact – it may be all the web pages served within the same DNS domain, or those accessible through a single server. Technically, it has no identity (URI) and therefore it cannot be the target of any protocol requests. Conceptually, it does not imply control or management in the same manner as a repository.

The remainder of this section describes the major components of digital library systems that support this repository-centric architectural core.

Portals

The portal, or the “front door of the digital library”, serves the same purpose as the physical entry to the traditional library. It provides the user with the clear notion of

being “inside” the digital library. Services and content within the portal are thereby blessed with the imprimatur of the library, endowing them with a level of trust and integrity. This is commonly known as “branding”. Correspondingly, most digital library applications clearly indicate to the user when they are “leaving the library”, for example by traversing a hyperlink to a page outside the boundary of the library.

The focus of a portal is usually a search interface, that in most cases is field-based, providing more functionality than the single text box search paradigm employed by most web search engines. This allows users to search on specific bibliographic fields such as title, author, or subject. This search paradigm reflects the influence of the library cataloging tradition [452], which eschews simple keyword searching that is predominant in mainstream crawler-based search engines (e.g., Google) in favor of more targeted search capabilities. Metadata, which is the basis of this field-based searching, is described in the next section.

In contrast to this “front-door” paradigm, the web user metaphorically “surfs” among linked information resources without regard for their location on the network. The informal notion of a “homepage” for a website does exist, but there is no presumption or enforcement of this as the uniform entry point to the collection of pages of that site. The notion of uniform, location-independent sources has proven to be quite powerful. In its simplest form it makes it possible to aggregate information from multiple sources in a single webpage, in the manner that a page may include an image that is stored in some other location on the net. As I will describe later, location independence is leveraged in Web 2.0 in a much more powerful manner in the form of “mash-ups.”

Metadata – cataloging in the digital context

The shaping effect of the library tradition on digital libraries is perhaps most evident in the focus on descriptive metadata. This focus has its roots in cataloging, one of the core functions of the modern library [146, 158, 201, 213, 336, 452].

The traditional catalog developed for a number of reasons. At the simplest level, in a library of physical resources it gave users an easy and compact tool for finding information resources without having to traverse the shelves. However, describing the catalog as merely a compact shelf list trivializes its complexity and intellectual content. Underlying cataloging is the concept of information entities having uniform attributes, such as author, title, or subject classification, and the utility of those attributes for logical organization of those entities. This organization presents multiple *access points* based on those uniform attributes, and allows users to search and browse within those access points [452]. For example, a user may search for information by author name and then traverse the resources associated with that author, or alternatively they may search for information by subject classification and traverse the resources associated within that class. As a result, the organization of the catalog and the manner in which it is made available to the user (e.g., cards in drawers or screens in an electronic catalog) is independent of the manner in which the physical, or digital, resources are organized on shelves, or in repositories. Furthermore, an individual information resource (e.g., a book) may have multiple catalog instances, each accessible through specific access points (e.g., title, author, subject, etc.).

Efforts to extend the practice of cataloging into the context of online information reflects an ongoing belief that in a world where even the books that are part of library collections have been digitized and are available for full-text search [261], structured search over surrogates is more functional and ultimately preferred by users. This is

despite empirical evidence that users seem to prefer the “one text box” search paradigm of Google to the fielded-search paradigm employed in most digital library portals and online catalogs, and decades-old evidence of the frequent superiority in recall and precision of automated full-text search to human-assisted indexing and cataloging [128, 129].

Traditional library cataloging is both complex and expensive, especially when applied to the rapidly expanding and diverse set of digital resources. The notion of metadata emerged as a simpler and less expensive alternative to traditional cataloging records, perhaps making it possible for nonprofessionals to create structured bibliographic information. The predominant digital library metadata effort is the Dublin Core Metadata Initiative¹⁷, which is described in considerable detail in later chapters of this dissertation.

Ironically, the origins of Dublin Core lie in improving search and retrieval on the general web [483]. However, this effort to develop easy-to-use bibliographic standards for networked information objects has been deemed irrelevant, ill-conceived [165], or even counterproductive [168, 222], by the mainstream web community and most notably the search engines that dominate it. As I describe later, the attempts to translate the benefits of cataloging to the online domain via metadata have been compromised by problems with ensuring the quality of the metadata records that are produced by non-professionals and preventing so-called “metadata spamming” by unscrupulous agents trying to falsely lead information consumers to their sites [149].

¹⁷ <http://dublincore.org>

Digital objects – containers for complex data and metadata

The content model of most digital library architectures is based on the notion of a *digital object* [255, 357, 403]; an identified (first-class) information resource that is an aggregation of multiple information units consisting of multiple formats, multiple subsidiary units (chapters of a book, issues of a journal), versions, or document components (e.g., the text, data, images, etc. of a scholarly paper). These are generally known as *compound objects*.

These object models reflect an ongoing effort by the library community to account for the complexity of information in both its abstract form and the physical or digital manifestations of it [103, 104, 336, 340]. These efforts have focused on mechanisms to represent the various relationships among information resources [452]; and to describe those resources at multiple levels of granularity [329], for various purposes, and in various descriptive formats [286].

Access to compound objects and their components is frequently mediated by protocols unique to the particular repository architecture. These protocols are usually embedded in the URLs that carry user requests from the digital library portal to the repository. These protocols allow operations such as access a digital object in a specific form, access a portion of a digital object such as its descriptive metadata, and the like. The proliferation of these architecture-specific access protocols has spawned a virtual cottage industry of repository interoperability initiatives [32, 234, 354, 393-396] in the digital library community.

In contrast, the web architecture [246] includes a quite simple information model based on the atomic resource. “Interoperability” is defined in the simple terms of the web architecture [246] – resources, URIs, and HTTP. There is no architectural notion of a compound object, or aggregation of resources. Ad hoc and de facto aggregations

exist, for example a logical document split into a set of interlinked web pages. However, these aggregations are not first-class objects; they do not have a unique identity and are essentially ephemeral. There is, in fact, an increased awareness in the web community that more complex information models are appropriate in a number of instances; for example, scholarly publishing. Chapter 12 describes our work to define one that is grounded in the principles of the web architecture.

Federation

The issue of federation [330] arises because digital library systems are conceived as discrete institutionally-managed entities with distinct boundaries accessible to the user through branded portals. Sometimes, a user might want to search across multiple digital libraries when a selected resource is not available in their “local” library. In the physical library domain, this problem is solved by union catalogs such as WorldCat¹⁸ and by interlibrary loan. Digital library applications employ two mechanisms to allow users to search for and access information outside the confines of a single digital library.

The first is federated searching or meta-searching, in which a single search query is multicast to several digital library search engines. The query is then individually processed at those search engines; the individual result sets are then returned and integrated at the site from which the original query was multicast. Federated searching was the subject of substantial work in the early years of digital library research [169-171, 194, 214, 393, 397] and Chapter 7 describes our own work in this area. Although instances of federated search still exist, the technique has fallen into some disfavor

¹⁸ <http://www.worldcat.org/>

because of problems with dependence on the reliability of multiple search sites and the problems with the ranking of search results from several sources.

The second is metadata harvesting, in which bibliographic records from several distributed institutional sources are combined at a single indexing site, which provides a search interface across the resulting “union catalog”. The most widely deployed mechanism for metadata harvesting is the Open Archives Protocol for Metadata Harvesting (OAI-PMH) [314], which is described in greater detail in Chapter 10. That same chapter describes complications with metadata harvesting.

Neither of these techniques has achieved widespread deployment in the general web information space. As described in an earlier section, boundaries and the repositories that implement them are not a part of the web architecture. Web search engines such as Google crawl the web via graph traversal, ignoring the notion of the location of a webpage, except as a tool for optimizing graph traversal strategies [84]. In addition, ranking algorithms such as PageRank [95, 96] are designed to operate over a centralized index, and are difficult if not impossible in the context of distributed methods such as federated search,

Persistent identity for network information

A final example of the difference between digital library and web architecture is the notion of “persistent identity” for information stored in digital repositories. The attention to this issue in the digital library contexts reflects concerns about both preservation and control of intellectual property [400]. The Handle System [449] is the best known of this class of technologies. Like many persistent naming systems, the Handle System depends on a hierarchy of identity resolvers, and therefore the notion of a central *root* name server. These efforts have gained little traction in the mainstream web community, which has historically resisted centralization and has

comfortably adapted to the fragility of URLs, deeming identity persistence as a policy problem rather than a technical problem [51].

Coexistence of digital libraries and the web

The previous section described the distinction between the repository-centric digital library architecture and resource-centric web architecture. In addition, it described how these different core architectural principles affected the technical components of each architecture. The digital library applications that have been assembled from these components are indeed quite powerful and include advanced searching capabilities, complex information models, and rich user interfaces. Ironically, the same architectural features that enhance their functionality have often interfered with the interoperability of digital library applications with the mainstream web and thereby mitigated the impact of these applications in the broader web context.

The problem comes from the fact that the specialized, repository-specific access protocols that provide access to these digital library resources often do not follow the conventions of mainstream HTTP access methods. For example, in many cases the URLs used to access objects are conflated with query predicates, the syntax of which is unique to the digital library and is hardcoded into portal/repository interaction. This is not a problem when access to the digital library resources occurs through the “front door” portal and through its respective search user interface that generates these query-based access URLs.

However, mainstream crawler-based search engines, such as Google, do not access objects through the front door, but rely on generalized graph traversal. The nature of the access URLs in digital libraries and their interdependence on the respective digital library search interface often makes these URLs unreachable via these graph traversal techniques. This is because the URLs of the digital objects in the repository are not

explicitly linked to, but are generated by the digital library based on search engine queries. These query-generated URLs are not visible in the web graph traversed by mainstream search engines and, as a result, the digital library resources are not crawled and are subordinated to an information black whole – the so-called “deep web” [49]. They fail to appear in result lists returned by mainstream search engines, which have emerged as the universal tool for discovery of information (much to the chagrin of the library community).

In an effort to increase search engine visibility, digital library providers frequently generate special link pages that expose the individual URLs of repository contents to crawlers as conventional hyperlinks. Digital library resources then appear as search results in Google and similar search engines. As a result, a steadily expanding amount of access to digital library resources occurs through these commercial providers¹⁹. But this reverse engineering to increase visibility of contained resources subverts the role of the digital library as a control zone, and the intention of the portal as a branded entry to that control zone. The “digital library collection” becomes just another set of web resources, with no joint identity or imprimatur. The digital library becomes “invisible infrastructure” [81], barely evident through a web-dominated information paradigm.

Digital libraries and the evolving web

In addition to the technical incompatibilities between the digital library and web architectures that were described in the previous section, there is a widening gap between their underlying information models. The web that Tim Berners-Lee invented in 1989 has undergone an explosive growth in scale, measured in terms of number of

¹⁹ Personal communication, J. Blake (NSDL) and S. Warner (arXiv).

URLs, servers, and traffic. At the same time, it has experienced a radical change in form and impact, referred to as Web 2.0 [390]. In contrast to the relatively passive and transactional search/access paradigm characteristic of the library and Web 1.0 in which the delineation between authors and consumers was relatively distinct, information interactions in Web 2.0 are highly interactive and participatory. Rather than just browsing and reading web pages, web users, acting as both authors and readers are writing reviews on Amazon, annotating and tagging pictures on Flickr, writing and updating articles on Wikipedia, publishing observations and research results in blogs, and mashing up online content into new content. This section examines the evolving web and the coexistence of digital libraries within that context.

The metaphor of versions – Web 1.0, Web 2.0, and the recently coined Web 3.0 – is obviously artificial and overly simplistic. However, it is a useful rhetorical device.

The predominant “features” of these versions are as follows.

Web 1.0 – called the document web, the “web of cognition” [412], or the “read-only” web [39]. The time span of this version roughly extends from the invention of the web until 2000. It primarily consisted of hyperlinked, semi-static, atomic documents (HTML, PDF, GIF or JPEG images). Interaction and collaboration were minimal except for document authoring and querying. Content creation required specialized tools and, as a result, was restricted to a small subset of web users.

Web 2.0 - called the “web of communication”[412] or the “read/write” web [420]. This “version” includes participation-oriented tools such as wikis, blogs, social applications like Flickr, and instant communication tools like Twitter. Another prominent feature is the notion of a “mash-up” whereby new information objects are created via the dynamic combination of existing information resources [432]. These features enable a phenomenon that Engestrom calls “object-centered sociality” [181,

182], in which information objects, people, and social exchanges are linked together in web space. This has effected a phase transition in the web's impact on economics, scholarship, learning, and most recently politics. The effect of this impact on national politics is exemplified by the recent observation that "... Barack Obama's victories in the Democratic primary and in the presidential election would not have been possible without Internet-empowered fund-raising and social networking" [126].

Web 3.0 – called the “web of meaning” or the “contextual web” [130], this currently emerging web functionality incorporates concepts of the semantic web [22, 56, 185, 360], the underpinnings of which Tim Berners-Lee and the W3C have been developing since the late 1990's. The key features of the semantic web include machine readability and interpretation of web data and the ability to reason over that data. The technological foundation of the semantic web is the Resource Description Framework (RDF) [273] a data model for expressing statements about entities (web resources) and their properties (ontologically defined relationships).

Raffl et al. [412] adopt the language of Evolutionary Systems Theory [144] to illustrate how the features of these versions are cumulative: “[E]ach new layer is built upon a preceding one and ... the new stage comprises not only the new layer, but parts of the old one”.

Figure 1 illustrates the changing nature of the web through these versions. As shown, Web 1.0 was primarily a one-way channel from producers to consumers. In Web 2.0, the bifurcation of web participants blurs into consumer/producers who collaboratively author, manage, and annotate content. This is enhanced in the Semantic Web (Web 3.0) in which machines (agents) process and interpret this collaboratively produced content and contribute new content back on the web.

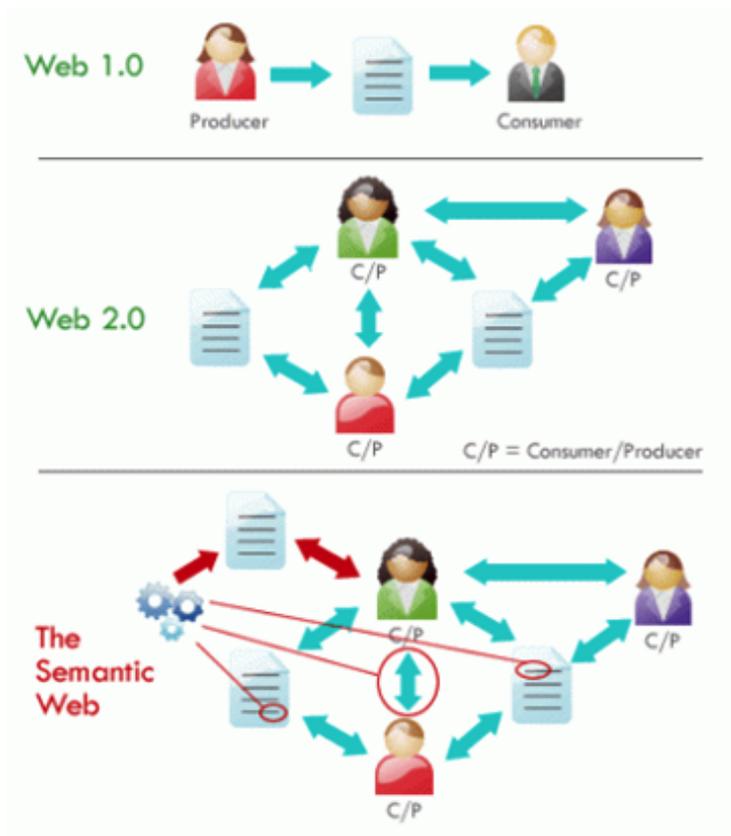


Figure 1 - Expansion of web functionality (from [130])

The participatory nature Web 2.0 should not be dismissed as just a popular phenomenon, manifested in increased use of mainly social sites like FaceBook. According to many experts in the field, we are witnessing a fundamental change in the prevailing information paradigm that is transforming all aspects of our culture. Notably, this impact extends to the nature of scholarly research and communication, one of the backbones of the research library. As pointed out by Paul Ginsparg, who create arXiv and is considered one of the icons of Internet-based scholarship:

... there are ... objective reasons to believe that we are witnessing an essential change in the way information is accessed, the way it is communicated to and from the general public, and among research professionals - fundamental methodological changes that will lead to a terrain 10-20 years from now more different than it was 10-20 years ago than in any comparable time period. [210]

Timo Hannay of Nature, one of the most prestigious scientific journals, makes the following observation of the impact of the current and future of the web on science scholarship:

For all but a very small number of widely read titles, the day of the print journal seems to be almost over. Yet to see this development as the major impact of the web on science would be extremely narrow-minded – equivalent to viewing the web primarily as an efficient PDF distribution network... Though it will take longer to have its full effect, the web’s major impact will be on the way that science itself is practiced. [227]

In addition, there is mounting evidence that, for a number of important information-oriented activities such as education, the participatory information paradigm in Web 2.0 has advantages over the traditional consumption-based model. Fuchs and Raffl et al. [200, 412] argue that Web 2.0 paradigms of collaboration, construction, and participation are more closely aligned with recognized models of human cognition and knowledge development than the more restrictive and controlled library model. Downes [167] and Ullrich et al. [458] describe the utility of the Web 2.0 model for education because of the manner in which it facilitates activities such as group collaboration, exploration, and manipulation that are key to learning according to cognitively-oriented constructivist theories. Gee argues that the “affinity spaces” facilitated by the Web 2.0 environment are powerful tools for learning [206, 207]. Black identifies the notion of “beta-reading” in online fan communities where contributors grow as readers and writers based on mutual feedback [65, 66]. Finally, others see the general benefits of the “wisdom of crowds” [266, 450] that is enabled by the collaborative nature of Web 2.0.

Chapter Wrap-up

This chapter described the origins of digital library research, the manner in which those origins shaped the technology produced by that research, and the compatibility that technology and the assumptions underlying it with the evolving web information

context. As described, digital libraries began with a rather simple assumption: the attributes, information model, and practices of the library could be translated relatively unscathed to the online environment. The prevailing belief was that the library would benefit from and be enhanced by the new technical developments, and the users of those libraries, the public, would in turn benefit from these “libraries without walls or operating hours”.

This early digital library work leveraged the concurrent development of the web. In its initial Web 1.0 form, it was primarily a technical system – a set of protocols and standards that enable browsing over a network of documents. In this form it provided an inert technical foundation for Digital libraries due to the fact that there was reasonable convergence between the web document-centric paradigm and the digital library document collection-centric paradigm.

However, as the web morphed into its 2.0 form it adopted a significantly more complex social nature overlaid on the core technologies introduced in Web 1.0. In this new social guise, the web was no longer inert, but profoundly active, embodying participation-based information interactions incompatible with many of the established concepts of the library. These concepts – institutionally-based boundaries or control zones, the document as a fixed unit of information, the unidirectional flow of information, and intermediation – are described in Chapter 3.

Whether the library as a meme or institution is flexible enough to adapt to these new paradigms and leverage their benefits is a matter of speculation. A set of initiatives known as “Library 2.0”, which intermingle traditional library services with Web 2.0 features, are now gaining in popularity in some elements of the library community [113, 358, 372]. Rather than comment on Library 2.0 in particular, I refer back to Christensen [122] who notes the difficulty of instituting radical transformations within

entrenched corporations and institutions [123]. Too often, these legacy institutions are burdened with continued support of legacy practices and with the demands of pre-existing customer communities. Furthermore, as mentioned earlier, the future viability of the library, digital or physical, rests not only in institutional changes, but also in modifications to public perceptions of it as outdated. Certainly, accomplishing both, especially the latter, is a formidable challenge.

Chapter 3

A Meme-based Analysis of Digital Libraries and the Web

Analyses of the web and libraries and their positions vis-à-vis each other have often degenerated into pedantic definitional disputes about the distinctions between them. The questions “what is a library” [34, 82, 102, 140, 260, 332], “what is a digital library” [29, 75, 76, 82, 192, 211, 299, 333, 337, 341, 349, 350, 354, 478], whether the web is a digital library [220, 274, 321], and whether there is any sense in the notion of a “digital library” [217] have been argued about repeatedly and to little avail. They reduce the web and the library, both complex entities, to enumerated lists of services and technologies that implement them – discovery, selection, cataloging, preservation, reference, and the like. These definition-by-enumeration exercises produce a sort of Theseus’s paradox²⁰, focusing on whether a library is still a library as its functions, services and technologies are replaced or stripped away.

It is more useful to view both digital libraries and the web as *sociotechnical information systems* [57, 318, 469], “networks of technology, information, documents, people, and practices”. This broader view takes into account the multiple facets and contexts of use of each system, making it possible to analyze their origins, historical trajectories, technical artifacts, and perceived value and utility.

²⁰ “The ship wherein Theseus and the youth of Athens returned from Crete had 30 oars, and was preserved by the Athenians down even to the time of Demetrius Phalereus, for they took away the old planks as they decayed, putting in new and stronger timber in their place, in so much that the ship became a standing example among the philosophers, for the logical question of things that grow; one side holding that the ship remained the same, and the other contending that was not the same.” [408]

The notion of a meme, recently adopted by the digerati for emerging web concepts (in particular Web 2.0 [390]), is useful because of the manner in which it endows a word with more meaning than its simple definition²¹. A meme is similar to a semiotic *sign* [375], with three dimensions of meaning:

- Semantic – the *denotata* of the signs, the things they refer to. In the case of the library these are the services (selection, preservation, collection, organization, reference), physical artifacts (impressive buildings, shelving), and internal culture and principles [24] (commitment to privacy, service orientation) that, in aggregate, are the library.
- Syntactic – the use of the sign in relationship to other signs, taking into account the effect on the sign when it is put in the context of another (e.g., combining “digital” with “library”).
- Pragmatic – the external perceptions of the sign by those who use it. In the case of the library these are cultural perceptions such as notions of trust, integrity, traditionalism, conservatism, and professionalism.

For the remainder of this dissertation I will use the terms “library”, “digital library”, and “web” to indicate their respective memes, unless explicitly qualified in some other fashion. Therefore, my use of the terms will connote the goals and missions of their individual communities, the core concepts underlying them, their technical and external manifestations, their positioning relative to other co-existing concepts, and the perception of them by external parties.

²¹ The word “meme” was originally coined by evolutionary biologist Richard Dawkins as “a unit of cultural transmission, or a unit of imitation” [154]. It was later used as the basis of a controversial field of “memetics” that proposed a largely discredited theory of cultural evolution, parallel to biological evolution [69].

This chapter expands the concepts introduced in the historical analysis in the previous chapter by examining in detail both the library and the web (both as version 1.0 and version 2.0) memes²². It uses the notion of a *meme map*, introduced by Tim O’Reilly [390], to illustrate the multidimensional nature of each meme. The four meme maps, shown in Figure 3, Figure 6, Figure 10, and Figure 11, have the structure illustrated in Figure 2.

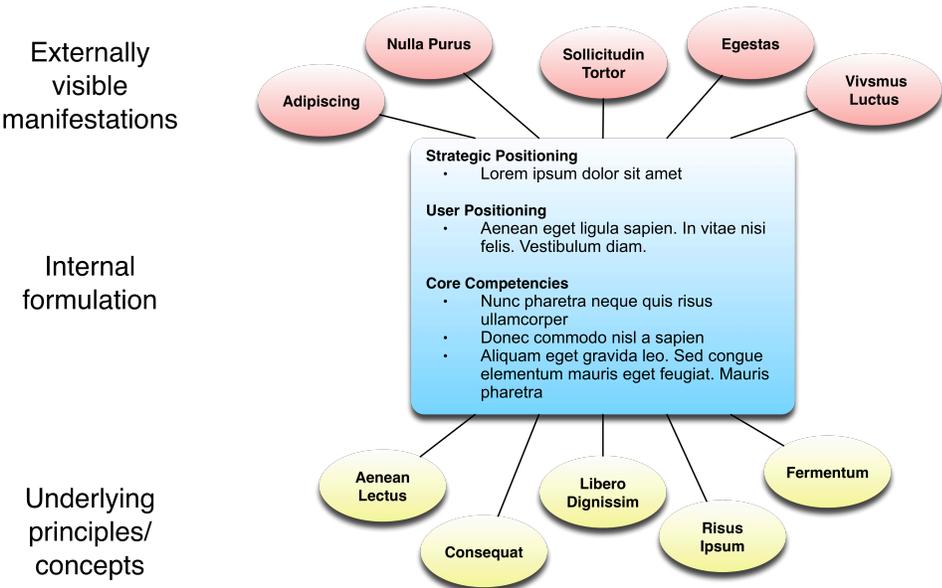


Figure 2 - Meme map

The rectangle in the middle contains information about the internal formulation of the meme. It includes the perceptions of the respective community about its *strategic role*, the *position of the user* relative to that role, and of its *core competencies* or perceived added value. The connected ovals below the rectangle show the underlying principles and concepts for the meme. Finally, the connected ovals above the rectangle show the

²² As with any argument, the characterizations made here about traditional and digital libraries and the web are, by nature, generalizations. Certainly, there are libraries, both physical and digital, and web applications that diverge in some ways from these generalizations.

externally visible manifestations and applications of the meme. Following a description of the four memes, the chapter concludes with an examination of the manner in which the Web 2.0 meme challenges almost all the core concepts underlying the library meme.

The library meme

Figure 3 is a meme map for the library. In this particular case it is the pre-web library that exists as a building housing physical resources (books, maps, serials, etc).

However, as indicated in Figure 6, I argue that, except for the external manifestations in the upper ovals, the other aspects of the meme map have by-and-large persisted in the transition from physical to digital libraries.

The center rectangle contains information about how the library community perceives itself and its utility to users. The notion of being the “first stop” for information was de facto true in the pre-web era. Outside of buying all the books that they needed or subscribing to all of the publications, users of the library in the pre-web era had little choice but to use the library as the first and every stop, especially if they wanted to search for information in any form. The remainder of the internal perception text in the blue rectangle includes the key notions of trust, integrity, organization, professionalism, and attention to the long-term. These are all manifestations of the core principles that are described in the remainder of this section²³. Finally, the ovals

²³ The issue of longevity is not included in these descriptions. Unlike the other concepts the need for long-term preservation of information resources remains constant in the transition from the digital to the online environment. But, despite years of research in this area including investigation of tools, policies, and mechanisms for preservation, many of the questions remain unanswered. In the words of Clifford Lynch, digital preservation is “enormous issue”. Furthermore, it is a “hard area to do compelling research in” and “digital libraries have made some contributions to this area, but limited ones.” Therefore, a comparison of the concept of longevity across these information environments is not relevant.

at the top of the meme map illustrate the external manifestations of the meme specific to the traditional library. The translation of these to the digital library was summarized in Chapter 2 and is described in greater detail later in this chapter.

Core principle 1: The boundary and the control zone

The functioning of the library depends on the definition of a clear boundary, a demarcation of what lies within the library and what is outside. This boundary is an essential foundation for two key library functions. The first key function is *selection*, the definition of the set of resources that are in the library’s collections. The second key function, which follows from the first, is *curation*, the stewardship or management of those selected resources to ensure their consistent availability over the long-term.

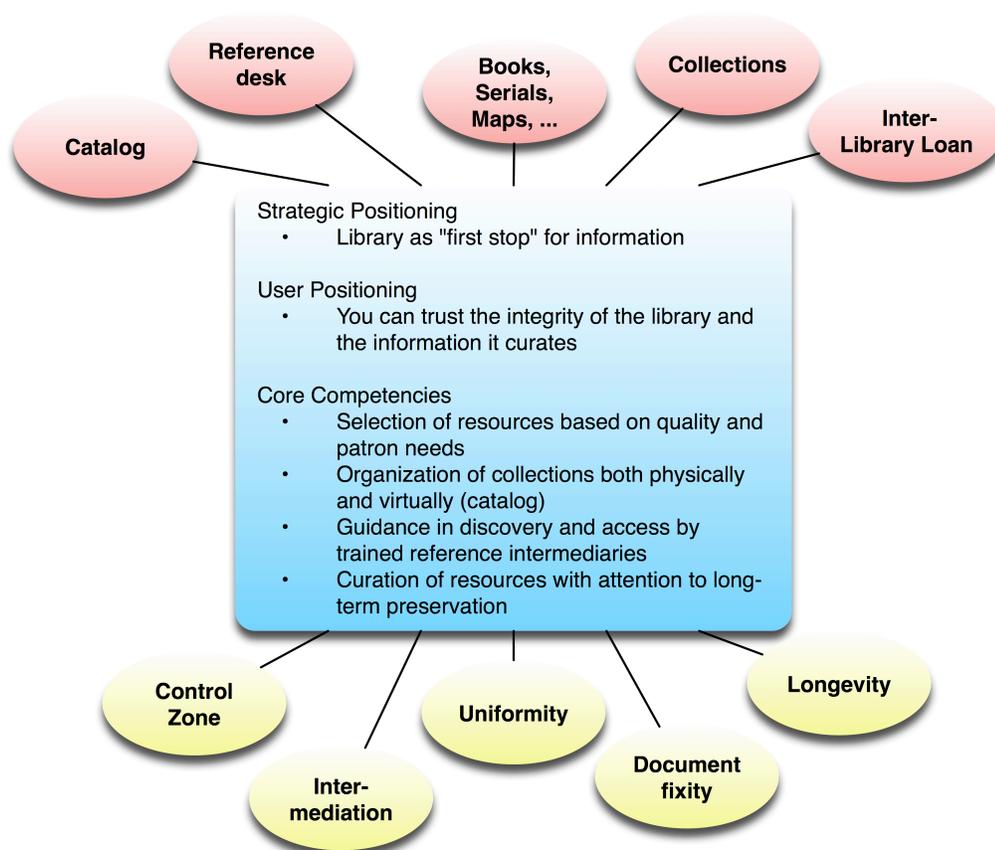


Figure 3 - Library meme map

The boundary of the traditional library was easy to define. It was the “bricks and mortar” structure that contained and protected the selected physical resources over which the library asserted control and curation responsibility. Correspondingly, from the patron’s point of view, the boundary marked what could be called a “trust zone”, an area in which they could presume that the integrity guarantees of the library existed.

The importance of the boundary and its utility as a *control zone* was described perhaps most eloquently by the late Ross Atkinson, Associate University Librarian at Cornell, in his seminal 1996 article “Library Functions, Scholarly Communication, and the Foundation of the Digital Library: Laying Claim to the Control Zone” [34].

Atkinson’s concern in this article was the manner in which the core qualities of the library would be maintained as libraries increasingly moved from physical to online form. As implied by the title and by the following quote, the ability of the library to maintain the boundary and the control zone it establishes lie at the root of this issue:

Some of the *most fundamental aspects of library operations* entail the existence of a border, across which objects of information are transferred and maintained. [34] (emphasis added)

Later in the essay Atkinson expands on these thoughts and clarifies the main purpose of the control zone, the establishment of value for the objects within that zone:

To add value to certain objects of information, therefore, always necessarily entails a reduction in the value of other objects. Therein lies the dynamics of *selection, which is the core operation of all library services*. [34] (emphasis added)

Finally, Atkinson states the manner in which this control zone provides guidance for the patrons at the library:

A library, digital or otherwise, is always a highly selective subset of available information objects, segregated in favor, to which access is enhanced into which the attention of client-users is drawn in opposition to objects excluded. [34]

As I will describe later in this chapter, this control zone, and its functionality, are difficult to maintain when its quite clear physical manifestations (bricks and mortar) are replaced by fuzzy virtual (online) devices.

Core principle 2: The library and intermediation

The notion of the control zone, and the selection service that it supports, are linked with a second core concept in the library meme: the belief that the library should provide trusted *intermediation* between information suppliers (publishers) and consumers (users or patrons). Intermediation is based on three assumptions. The first assumption is that participants in the library information system play fixed roles. The second is that librarians and the library itself are objective and trusted parties. The third is that there is primarily a unidirectional information flow amongst the roles²⁴. The roles and flow of information are illustrated in Figure 4 and are as follows:

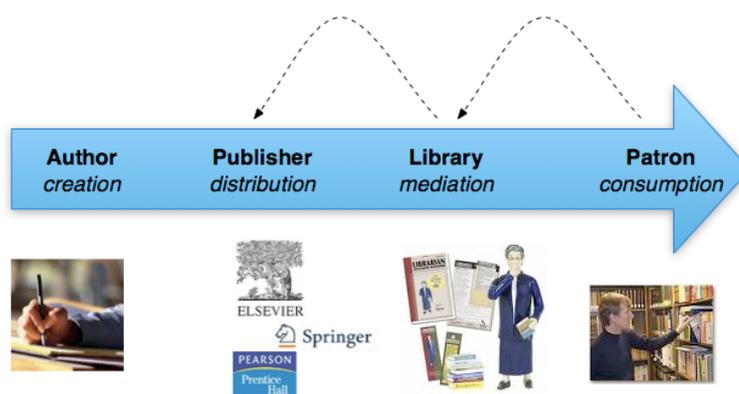


Figure 4 - Library information flow

- The *author*, or the producer of content, sits at the beginning of the information chain, creating content, copyright to which s/he then sells to a publisher.

²⁴ Obviously, there are other contextualized information flows, such as book purchasing, in which retailers and wholesalers play a role. These information flows have also been disrupted by the Web and digital availability of content (eBooks such as Amazon’s Kindle®) [265].

- The *publisher* assumes control of the copyright of content and subsequently controls its physical distribution, selling it to customers including libraries.
- The *library*, within the boundary of its control zone, selects the information that is available to patrons, manages and curates that information, and provides organizational and search tools for the discovery and access to that information.
- The *patron*, the consumer of information, uses the organizational tools provided by the library and sometimes the direct help of a reference librarian to discover appropriate information and gains access to it.

As noted, there is no direct interaction between the information supplier, the publisher, and the consumer, the patron. In addition, there are no established paths for backward flow in the system. The dotted lines pointing backwards in the information chain indicate that there is some feedback from patrons to the library and from the library to the publishers, but this feedback is generally out-of-band and private. For example, while the library may have a suggestion box for patrons, they are not permitted to directly annotate or supplement the information resources or metadata about them in a matter that is visible to other patrons.

As Atkinson [34] points out, intermediation, both direct via reference services and indirect via cataloging and selection, provides the information consumer with guidance through an otherwise confusing information environment.

Mediation is that service intended to assist the client-user in gaining access to objects of information with a specific content needed for a specific purpose.

As Levy points out, the “assist the user” motivation of intermediation takes on an even more active role in the public library context [337]:

In the same way that the teacher in the classroom steers the student to particular knowledge, an information service has the responsibility for guiding the user to certain information.

Core principle 3: The library and documents

Throughout their long history libraries have successfully incorporated new genre of materials in their collections. Any modern library, even most modest public libraries, includes books, serials (magazines and journals), maps, audio recordings, video, microfilm and microfiche, and, recently, electronic texts and databases. The facile nature with which they include these new genres of materials is possible because of the manner in which libraries use the common abstraction of the *document* as a means of hiding genre differences for most basic services. The document abstraction includes two fundamental characteristics:

- *fixity* – the relative stability of documents, which makes it possible to name them, describe them, and trust the nature of their contents (and that those contents will not spontaneously change).
- *provenance* – the ability to determine from where and from whom documents are sourced, which are essential for determining their integrity and legality relative to copyright.

The notion of the document has been the subject of considerable scholarly writing in the fields of library and information science.

Michael Buckland, in two short essays [103, 104], cites Suzanne Breit [94], to argue that “A document is evidence in support of a fact”. This functional, rather than material, view minimizes differences in physical form, thereby opening the notion of a document to media very different than conventional text (Buckland even argues that an animal in a zoo could be considered a document since it is “evidence in support of a fact”!). This expanded definition makes it possible for libraries to incorporate new genre of digital content, applying to them the attributes of “documents”:

If we sustain the functional view of what constitutes a document, we should expect documents to take different forms in the contexts of different technologies and so we should expect the range of what could be

considered a document to be different in a digital and paper environments. The algorithm for generating logarithms, like a mechanical educational toy, can be seen as a dynamic kind of document unlike ordinary paper documents, but still consistent with the etymological origins of "document", a means of teaching - or, in effect, evidence, something from which one learns. [103]

David Levy, who has written extensively about documents and their transition to digital form [338-340], takes a similarly expansive view of documents. He calls documents “talking things” that “we’ve imbued with the ability to speak.” [338]. This is similar to the perspective of Latour [327] who, with the notion of *delegation*, views a document as an object to which the author has delegated her/his communication about some subject. Because talking things can take many forms, Levy expresses confidence about the ability of the library to incorporate new genre of documents, including digital, into their collections.

Core principle 4: The library and uniformity

Reality is complex and the information resources that are part of it and reflect it share that complexity. The previous section described one aspect of that complexity; information (i.e., documents) may be instantiated in several physical forms. However, that complexity extends into several dimensions in addition to the physical. For example, there is the temporal dimension; documents may change over time and be manifested via new editions, derivations, and versions. There is also a linguistic dimension; content may be translated into several languages, and each translation may be published separately. Finally, there is a semantic dimension; different information resources may be interrelated via subject similarity or common authorship. Extracting the details, or bibliographic features, of these multiple dimensions of complexity and the relationships among them is often non-trivial. For example, authors may publish under multiple pseudonyms or may change their name over time. And, determining

the subject of an information resource requires interpretation of its content; a task that until recently was an exclusively human skill.

One of the historically great contributions of libraries is “providing coherence and *uniformity*” [350] to this complexity and thereby providing access to it. The Catalog and cataloging standards that underlie it are “order making” tools [336], permitting libraries to present a veneer of uniformity on the natural disorder of the information resources that they manage. This veneer consists of a set of surrogates [284] that conform to a common information model [452], and thereby represent the underlying heterogeneous resources in a homogeneous manner. The development of order-making standards is an important part of library history and includes the invention of schemes such as Dewey’s [115] and Cutter’s [146] classification systems, the development of the almost universally accepted Anglo-American Cataloging Rules (AACR2) [213], the MARC standards²⁵ for encoding catalog records in machine-readable form, and the more recent introduction of entity-relationship-based Functional Requirements for Bibliographic Records (FRBR) [3] information model. The central role of the catalog in the functioning, and perhaps relevance, of the library is articulated by Calhoun [106]; “[t]he library catalog has long been the keystone supporting the mission of libraries...” Furthermore, loss of the catalog would place “in jeopardy the legacy of the world’s library collections themselves”.

The library made digital

As described in the previous section the library is built on the core concepts of boundary, intermediation, documents, uniformity, and longevity. In the traditional, or

²⁵ <http://lcweb.loc.gov/marc/marc.html>

pre-digital library, these core concepts were instantiated by the set of externally visible technologies and artifacts shown at the top of the meme map in Figure 3.

Figure 5 realigns the concepts and external artifacts from Figure 3 to indicate the concept-to-artifact mapping. As indicated, this is not a one-to-one mapping, with some external artifacts realizing a number of concepts. For example, cataloging, which earlier was described as one of the keystone library functions, embodies the control zone, uniformity, document fixity, and intermediation concepts.

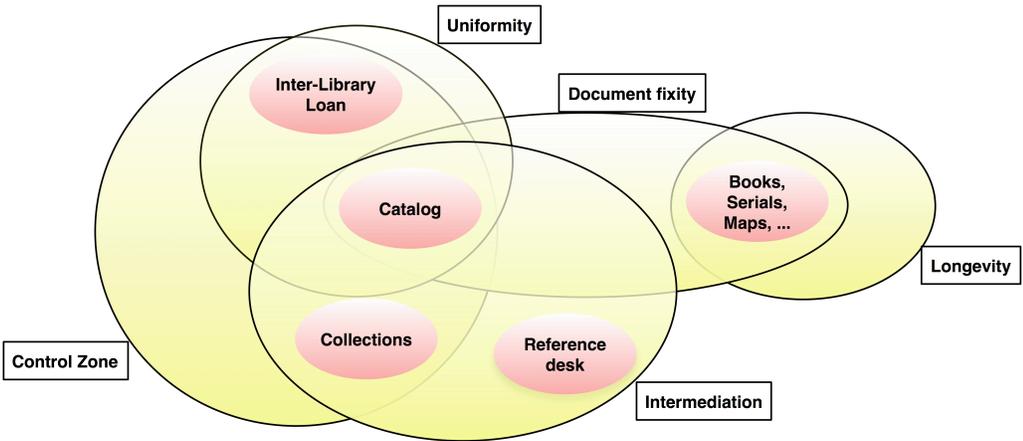


Figure 5 - Mapping of concepts to external artifacts in traditional library

I asserted in Chapter 2 that the nature of the communities involved in the creation of digital library research initiatives and their preconceived notions of what it meant to “make a library digital” led to a projection of traditional library concepts into the technical artifacts of digital library applications. This led to a largely institutionally-based repository-centric architecture with components consisting of metadata, portals, digital objects, repositories, and federation. The mapping of these digital library components to their traditional library predecessors is illustrated in the digital library

meme map shown in Figure 6. Note that this is identical to the library meme map in Figure 3, except for the externally visible technologies displayed at the top of the map. Transitively, these “made digital” library components could be inserted in Figure 5, replacing their traditional predecessors, while realizing the same core underlying concepts. This is illustrated in Figure 7.

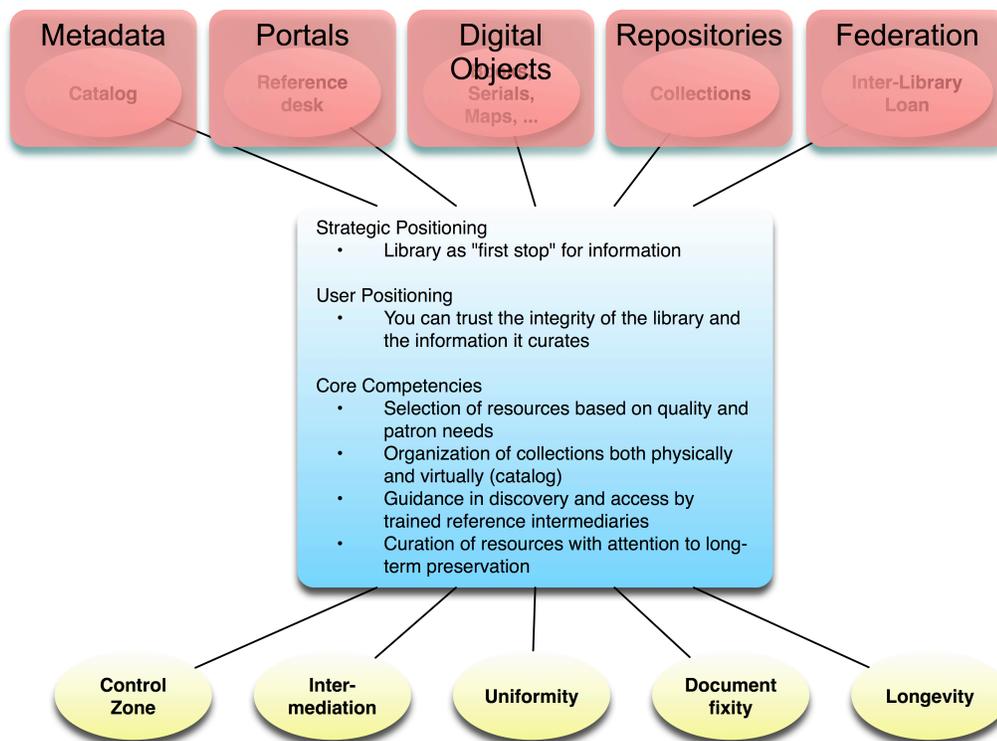


Figure 6 - Digital library meme map²⁶

Admittedly, not all digital library applications exactly conform to this idealized digital library meme map. But aspects of this model are evident in most deployed DL applications. The Dienst system described in Chapter 6, which I classify as a “classic”

²⁶ The manner in which traditional library concepts are visible behind the digital library technologies in this meme map is intentional and corresponds to the notion of pentimenti described in Chapter 1.

digital library architecture, exhibits all of these characteristics. In addition the initial architecture of NSDL, described in Chapter 10, generally conforms to this model.

The final part of this chapter describes the problems and contradictions that arise in this direct mapping from the traditional to digital library. These contradictions arise from the coexistence of the library meme with the increasingly dominant and divergent web meme, which I describe in the next sections.

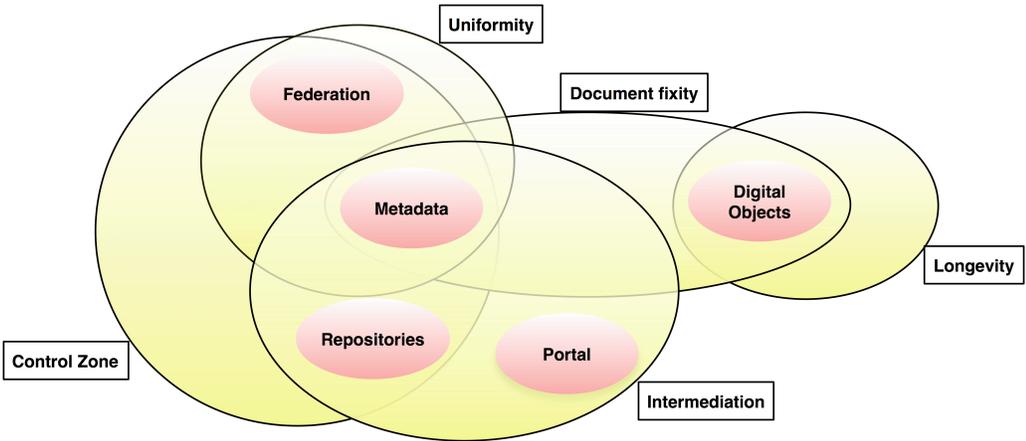


Figure 7 - Mapping of concepts to external artifacts in digital library

The web as a technical artifact

Before examining the meme maps of Web 1.0 and Web 2.0, which frame them as sociotechnical systems, it is useful to examine the web solely as a technical architecture. This architecture has been remarkably stable over the almost two decade history of the web, despite dramatic changes in the nature of the information model and applications that leverage that architecture. This is the motivation underlying Tim O’Reilly’s comment that Web 2.0 is an “attitude not a technology” [390].

A full description of the architecture of the web is contained in [246]. The core concepts and technologies are summarized in the remainder of this section.

- A *Resource* is an item of interest.
- A *URI* is a uniform global identifier for a Resource [54]. The URI specification is syntactic, rather than semantic, specifying the components (and their respective delimiters) of a URI that can be parsed out by implementations. A key component is the *scheme*, which prefixes the URI and is delimited by a colon (“:”) character. The scheme identifies the *namespace* of the URI and the specification (e.g., HTTP) that should be used to parse the remainder of the URI, and possibly the protocol to access a representation of the respective resource (if the URI is resolvable). Note that this was commonly called a uniform resource locator (URL) in the early web, but this name is deprecated because the notion of “location” implies that all URIs can be dereferenced. This is not universally true, especially in the semantic web context where URIs may denote entities in the physical world (e.g., people).
- *HTTP* is The Hypertext Transfer Protocol [186], the most common URI scheme in the current implementation of the Web. HTTP is a stateless, text-based protocol consisting of eight request methods for access and deposit of data streams. The stability of the web as a technical artifact is demonstrated by the fact that the current version of HTTP is 1.1, which was specified in 1999.
- A *Representation* is a data stream transmitted to a user agent (e.g., browser) that corresponds to the state of a Resource at the time of a dereference of a protocol-based URI (such as one with an HTTP or FTP scheme). Note the distinction of this physical entity – it is a stream of bits – from the abstract nature of a Resource. The relationships between a Representation, a URI, and a Resource are shown in Figure 8.

The web Architecture allows for multiple Representations of a Resource, with access mediated by *Content Negotiation*. This is a feature of HTTP that specifies the interaction between a client and server to determine the nature of the representation returned from a client request to dereference a URI. For example, a *user agent* (a browser) may request an HTML Representation of a Resource from a server. The server may return such Representation or may return another available Representation.

- A *Link* is a directed connection between two Resources. In most common usage, a link is expressed via link or anchor tags (a hyperlink) in an HTML Representation [413] of the originating Resource to the URI of another Resource.

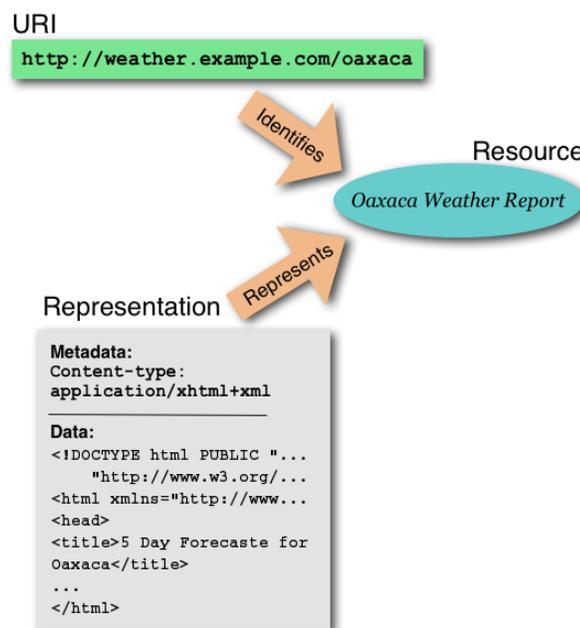


Figure 8 - Relationship of basic web architecture components [246]

The combination of these concepts forms what is commonly referred to as the web Graph [98], with nodes that are Resources (identified by URIs), some of which provide access to Representations, and edges that are Links. An example of a web

The importance of XML and its use as the basis for recent versions of HTML [497] for Web 2.0 is twofold. First, by formalizing the structure of HTML, and other markup languages for which it provides a syntactic foundation, it enhances the ability of machines to parse web document contents. This makes it possible to then process these contents algorithmically, enabling a new generation of rich web applications. Second, it is the foundation upon which syndication formats such as Atom and RSS are built. These syndication formats are used in feeds, which are the primary means of aggregating content for Web 2.0 applications and mash-ups.

A number of other technologies have emerged in Web 2.0 that leverage the core web architecture and enhance the user experience. These include Ajax [203], JavaScript [188], and JavaScript Object Notation (JSON) [141]. Further description of these is out of scope.

The Web 1.0 meme

In early Web 1.0, the nodes in the web graph shown in Figure 9, the Resources, were almost exclusively HTML pages or other static formats (e.g., PDF). This produced the so-called “document web”, a hypermedia (text and images) network that extended the earlier notion of hypertext developed by Nelson [383].

The essence of the Web 1.0 document web is illustrated in the meme map in Figure 10. At the bottom of the map are the core concepts underlying the meme. Note that these core concepts have a one-to-one mapping to the fundamental web technologies described earlier. Indeed, Web 1.0 was largely a technical phenomenon.

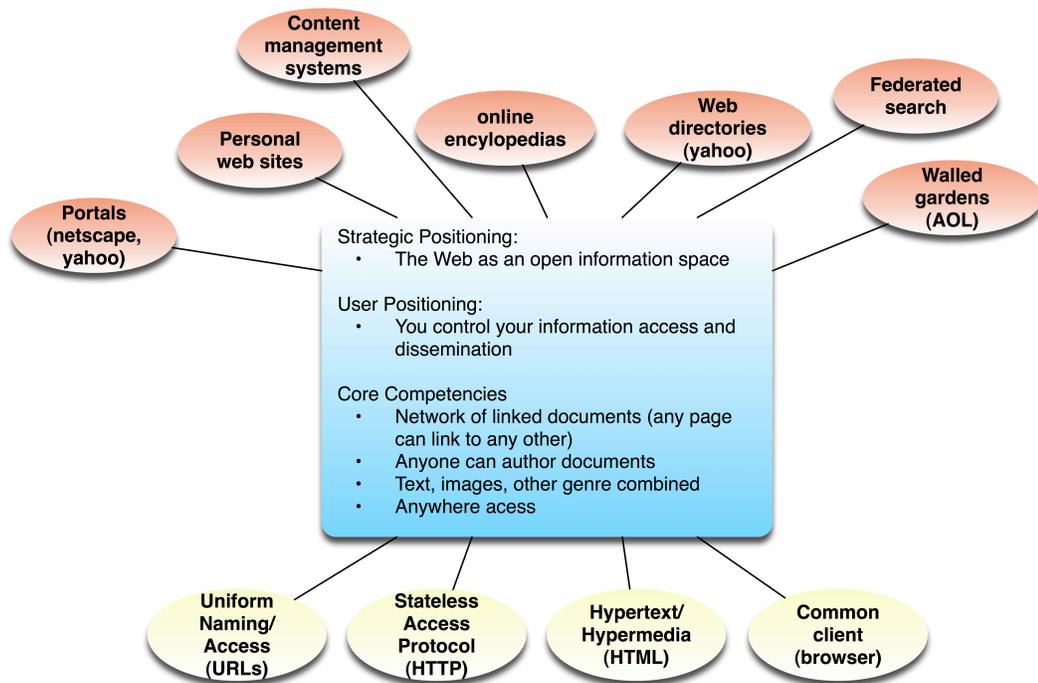


Figure 10 - Web 1.0 meme map

The box in the center shows the strategic positioning of Web 1.0: a hyperlinked open information space in which users controlled their access to information and anyone could author documents. However, the authoring process for Web 1.0 was out-of-band from access, requiring specialized non-browser-based tools. This is the basis for calling Web 1.0 the “read only” web, because these authoring tools were distinct from the browser, the ubiquitous tool for web interaction²⁷.

An examination of several of the externally-visible application shown in the ovals at the top of the meme map indicates the manner in which the information model of Web 1.0 and its applications were influenced by the traditional library model. A clear break from these traditional information metaphors had yet to occur.

²⁷ Note that this was not the intent of Berners-Lee, who envisioned the web client as both an authoring and reading interface [55].

Portals – The notion of an “entry point” to the web was a popular theme, with Netscape and Yahoo as primary examples. These portals were designed to provide users with an ordered view of the unruly web by linking to a *selected* set of Resources. This reflected the traditional notion of *intermediation*, the belief that users of the web needed guidance as they entered it.

Walled Gardens – These are also called “gated communities”, and are stronger examples of the intermediation and user guidance principle. Services such as America Online (AOL) provided their subscribers with a highly controlled segment of the web, promoting it as higher quality, “kid safe”, and easier to navigate [35, 320, 486].

Online Encyclopedia – The most prominent example of this was, of course, the Encyclopedia Britannica, which was originally introduced on the web as a cost-based service. This exemplifies the transfer of pre-web professional publisher dynamics and information flow to the Web.

Federated Search – Early web search engines were quite imprecise. Methods to exploit link information to improve the precision of results had yet to be invented, and the traditional information retrieval algorithms used by these search applications were poorly matched to the scale and diversity of the web. Federated search provided an alternative. Rather than searching the entire web, it made it possible to limit the search to a set of “trusted” institutional-based providers, thereby improving the precision of search results (while simultaneously limiting recall relative to a search of the entire web). In fact, Z39.50, a non-web-based protocol for federated searching, had already been proposed by the library community, creating essentially an online equivalent of already established multi-library union catalogs [5].

Directories – In another effort to improve navigation and organization of information, some services, in particular the early Yahoo!, manually organized web pages into

hierarchical directories based on category taxonomies, which are conceptually similar to traditional library catalog classification schemes [115]. In an article critical of metadata and ontologies (and promoting Web 2.0 tagging), Shirkey said the following sarcastic comment about this effort by Yahoo!: “Yahoo, faced with the possibility that they could organize things with no physical constraints, *added the shelf back*” [434].

Collectively, these applications exemplify the infiltration of library information archetypes into the early document-centric Web. The next section describes how the nature of web applications changed as a new web meme emerged.

The Web 2.0 meme

An examination of Tim Berners-Lee’s original design of the web reveals that Web 1.0 was actually a limited instantiation of his initial vision. The web was supposed to be read/write from the beginning, with browsers incorporating both access and authoring functionality [55]. Evidence of this exists in the initial design of HTTP, which included, and still includes, protocol requests for access to and deposit of content. Furthermore, the “Information Management Proposal” [52], which Berners-Lee wrote in 1989 while at CERN illustrates that he envisioned a network of intermixed documents, data, people, and organizations. In other words, the “social web” was part of the original vision.

In architectural terms, Web 2.0 is really just a full realization of the web graph abstraction illustrated in Figure 9. Note the purposefully abstract definition of a resource as an “item of interest”. In Web 1.0, these items were almost exclusively HTML documents. But, this definition is sufficiently ambiguous to denote any entity in addition to the simple HTML documents. A resource may be browser-renderable digital genre such as text, images, movies, and audio. It may also be non-renderable

digital artifacts such as computational services and scientific data (experimental results, sensor feeds, etc.). Finally, a resource may be a concrete, physical entity such as a person or an organization.

In addition to fully realizing this powerful heterogeneous graph model, Web 2.0 applications such as blogs, wikis, social network sites, and the like integrate authoring into the user experience, making participation the norm rather than the exception. The combination of these features has produced a veritable mapping of the richness and dynamism of the physical, intellectual, cultural, social, and political world into a common network model. Kleinberg [268] calls this "The Convergence of Social and Technological Networks": "a coming together of the technological networks that connect computers on the Internet and the social networks that have linked humans for millennia". Phenomena such as "six degrees of separation", "small world phenomena" [371], and group formation [37] can be studied at scales unimaginable in earlier limited sociology experiments. In addition because of the intermixing of social networks and the document/data web, it is possible to study what Engestrom calls object-centered sociality [182]; the interaction of people with information, the flow of ideas, and the development of new concepts amongst people and the artifacts in which they imprint those ideas [334].

The essence of this Web 2.0 phenomenon is illustrated in Figure 11, details of which are described in the remainder of this section. The text in bold in the figure corresponds to aspects of the meme map that are described in this section.

The core competency "*architecture of participation*" is the foundation for the read/write nature of Web 2.0. This builds on the underlying concept of "*trust your user's*", and represents a fundamental breakdown in the de facto assignment of roles among participants in the web information system. As shown in the meme map, the

notion of *user as contributor* that extends from the ranking of search engine results in Google to the annotating of products on Amazon to the creation of blogs and Wikipedia eliminates the distinction between readers and authors, which was strictly enforced in the pre-web world and was partially mandated by the nature of the tools in the Web 1.0 world.

Although the quality of information on the web remains an issue, the core competency of “*harnessing collective intelligence*” has proven to be a useful quality-assurance mechanism in a number of cases. The well-known example is Wikipedia (“*radical trust*”), which has been shown to have quality that is comparable or better in some cases the Encyclopedia Britannica [209, 447] (the online presence of which was ironically a significant Web 1.0 achievement).

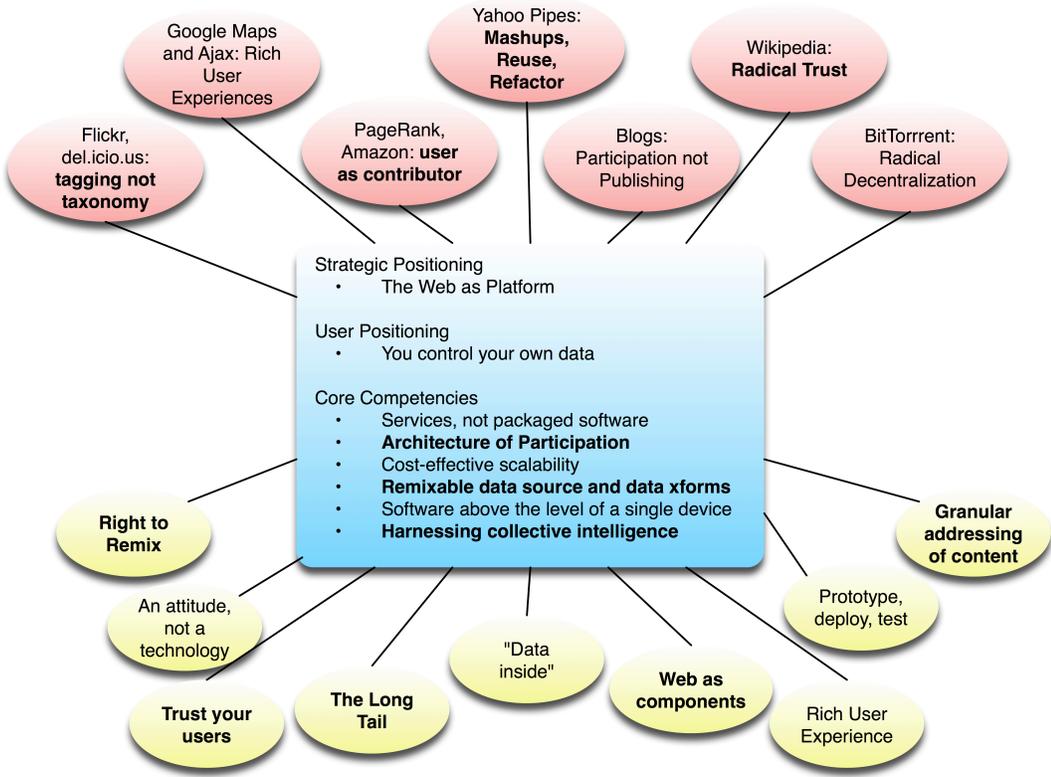


Figure 11 - Web 2.0 Meme Map

The core competency “*remixable data sources and data transformations*” builds on the concepts of “*right to remix*”, “*web as components*”, and “*granular addressing of content*” and is realized in architectures such as Yahoo Pipes²⁸ (“*Mashups, Reuse, Refactor*”). These principles are the foundations for mashups, Web 2.0 “documents” that can be created on the fly from portions of pre-existing documents and/or transformations via web services of those documents or portions thereof. This is a significant change from the atomic and relatively stable of web pages in Web 1.0, and a considerable distance from materials in the traditional library.

“*The long tail*” is based on the notion introduced by Anderson [21]. The value of access to a massive amount of web content, without pre-filtering, is widely recognized. The guardian principles of Web 1.0, evident in the popularity of portals and walled gardens, have disappeared in favor of “*harnessing collective intelligence*”, upon which situated machine-learning based filtering and personalized selection of content can occur.

“*Tagging not taxonomy*”, a principle underlying Flickr and many other Web 2.0 applications, distinguishes Web 2.0 organizational techniques from the taxonomy-based directory applications in Web 1.0. Tag clouds, or Folksonomies [221, 365], are another way of “*harnessing collective intelligence*”. Furthermore, recent work [89, 460] on machine learning and advanced clustering techniques has made it possible to automatically refine these human-based informal tagging methods into more formal ontologies and thereby match some of the “intelligence” of traditional professional cataloging. This paradigm of combining volunteer, non-professional human effort with advanced computational techniques (articulated by Arms [28]) is increasingly

²⁸ <http://pipes.yahoo.com/pipes/>

widespread in Web 2.0, especially with the increased availability of massive compute power due to the commoditization of cloud computing [25].

In summary, the Web 1.0 to 2.0 transition fundamentally changed a number of the core assumptions in the early web, especially those modeled upon aspects of the traditional library meme. The “user as reader” paradigm of Web 1.0 morphed into the powerful notion of “users as participants at all levels”. Established notions of guidance, selection, and taxonomic organization were put aside in favor of more flexible and informal filtering and organizational techniques based on the principle of harnessing collective intelligence. Finally, the largely static notion of a webpage as a digital document was deconstructed into the remix and reuse mashup culture of Web 2.0.

Conflict among the memes

This section returns to the library meme, compares it to the two web memes, and illustrates the widening gap between the fundamental assumptions and principles of the library and web information systems. The content of this section is summarized in Table 2, where each row corresponds to one of the core library meme concepts, described earlier, and each column corresponds to a meme. A cell, therefore, describes the manner in which a meme supports, or does not support, the corresponding concept.

Table 2 - Essential library elements compared

	Library	Digital Library	Web 1.0	Web 2.0
Control Zone	<ul style="list-style-type: none"> • Selection and collection by “bricks and mortar” • Branding by entry 	<ul style="list-style-type: none"> • Repository • Federated Search 	<ul style="list-style-type: none"> • Selection and collection by web sites • Branding by Domain names and portals 	<ul style="list-style-type: none"> • Long tail • Personalization
Intermediation	<ul style="list-style-type: none"> • Fixed roles one-way information flow 	<ul style="list-style-type: none"> • Portal 	<ul style="list-style-type: none"> • Separation of reading and authoring • Publisher web sites • Content management 	<ul style="list-style-type: none"> • Boundary between reading and authoring is undefined
Documents	<ul style="list-style-type: none"> • Definition by physical binding • Fixity and provenance certain 	<ul style="list-style-type: none"> • Digital Object 	<ul style="list-style-type: none"> • Web pages • Relative fixity • Provenance based on domain name/site origin 	<ul style="list-style-type: none"> • Wikis/Blogs/Mashups • Fixity and provenance uncertain or non-existent
Uniformity	<ul style="list-style-type: none"> • Cataloging standards • Physical organization 	<ul style="list-style-type: none"> • Metadata 	<ul style="list-style-type: none"> • Metadata standards • Directories/Taxonomies 	<ul style="list-style-type: none"> • Tags/Folksonomies • Google

From control zone to crossing boundaries

As noted by Van House, “digital information crosses boundaries easily... [469]” In other words, online content, even within the document-centric paradigm of Web 1.0, is by nature problematic for the control zone principle of the library. However, the notion of a boundary is particularly at odds with the underlying principles of Web 2.0.

I am not asserting that the notion of a boundary and the management of the objects therein are irrelevant. As Marshall points out, boundaries are useful: “... the practical everyday reality of workplaces and public institutions... is rife with boundaries that shape human interaction and any associated engagement with technology and documents.” [362]. However, in an era of networked distributed scholarship, the utility of an institutionally-mandated boundary, in the manner of the library meme, is debatable. There are two reasons for this.

The first reason is the decreasing relevance of selection, due to the combination of the recent mass digitization projects, which retrospectively are putting unprecedented amounts of legacy content online, and the near ubiquity of various forms of digital publishing, which are ensuring that the vast majority of new content is “born digital” and available online. Although not all content is available online, the amount that is online is sufficient to meet almost all the needs of an increasing number of information consumers. As commercial endeavors such as Netflix and research efforts such as movielens²⁹ have shown for movies, making the full extent of a collection available to users (i.e., starting with the premise of no selection) seems more sensible than pre-determined selection based on the supposedly common interest of a community (i.e., a library’s patron community). In this manner, it is then possible to exploit the long tail with subsequent application of individual or community-focused, algorithmic selection techniques (e.g., recommender systems, collaborative filtering) [36, 369]. These automated techniques are especially relevant in the participatory environment of Web 2.0 in which information users actively annotate the utility and quality of content.

Second, it has become increasingly difficult for an institution such as a library to define exactly what is the community that should guide their boundary-setting activities due to the inherently location independent nature of digital information delivery. Take, for example, the “community of scholars” that an academic research library supposedly serves. As global communication has become easier and cheaper, the locality of scholarship has broken down and collaborations frequently involve scholars spread across several institutions and continents. As Wheeler notes [485],

²⁹ <http://movielens.umn.edu>

“... scholarly work products are increasingly multi-authored, and often with scholars from two or more institutions and multiple disciplines.” In addition, “... the Internet has hastened the informal and formal communication of scholarly results, and the research process is increasingly a process of contributing to community data repositories and conducting further research from community data.” Lastly, “... these connectionist endeavors are not only within a research community, but many advances in human knowledge rely on deep interdependencies between disciplines...” As described then, scholarship is increasingly global and collaborative with a highly diffuse and situation-specific community structure. This is problematic for libraries that choose to define an institutional control zone based on a defined community’s needs and interests.

Some members of the library community have called for libraries to embrace the notion of the long tail and, in fact, recognize their contribution to it. In a recent article, Dempsey [156] notes how “libraries collectively manage a long tail of research, learning and cultural materials.” He further describes how libraries “... need new services that operate at the network level, above the level of individual libraries.” Proposals such as this are interesting and may point out a path forward for libraries in the Web 2.0 context. However, if as Atkinson suggests the control zone principle is essential to the library, contradictions between exploitation of the long tail and the control zone must be resolved.

From intermediation to disintermediation

Like other components of the library meme, mediation should not be totally dismissed. Navigating an information universe often requires expertise that students and even researchers lack, and all of us could at times benefit from the help of an experienced reference librarian.

But the notion of the benevolent guide or gatekeeper standing between well-defined producers and consumers of information began to break down in Web 1.0 and is clearly contrary to the participatory culture of Web 2.0. Direct searching of the global information space using commercial search engines and subsequently bypassing the library for information access is part of every day life. This will only become more prevalent as increasingly powerful mobile devices and ubiquitous conductivity infiltrate further into the culture.

Intermediation is also subverted in the Web 2.0 environment due to the breakdown in fixed role assignments. Participants in the Web 2.0 information system rapidly move between authoring, consuming, publishing, annotating, and a host of other information roles. This cyclic and dynamic information flow is illustrated in Figure 12 (contrast with Figure 4), which shows how users of the web continually cycle through multiple roles, all of which contribute to an information space that is shared amongst all participants.

The “user” or “patron” has been replaced by the “participant” in Web 2.0. As Van House [468] points out the notion of the user has all along been artificially defined and imposed:

... the user is constructed, and configured, not a natural object with characteristics to be described and information needs to be “discovered.”... [the] representations of users (even the term “users”) are culturally and historically situated, intended to help in the design of services and systems, but not likely to reflect the participants’ (information users and producers, knowledge workers) own views of their situation.

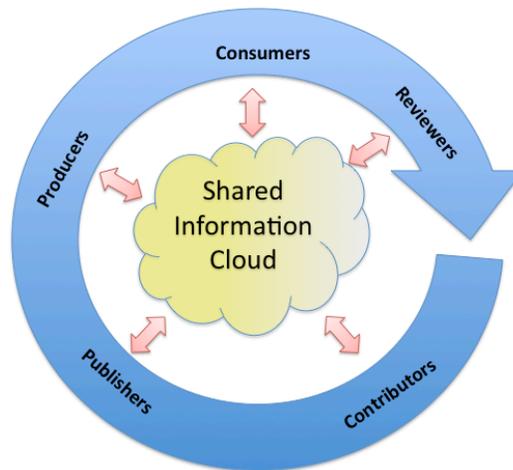


Figure 12 - Web 2.0 information flow

Benkler [48] similarly states that the pure notion of the information consumer, or user, is really just a side-effect of a constrained physical information environment and of an information economy in which the channels of distribution were expensive and therefore controlled by large institutions (e.g., publishers, libraries, media companies). The complete democratization of the distribution system in Web 2.0 has eliminated the pre-existing role constraints, and we are seeing the progressive breakdown of fixed roles in all aspects of information flow, and elimination of the need for intermediaries to broker among them.

From documents to mashups – the decomposable and composable document

As described earlier, both Buckland [103, 104] and Levy [338, 340] adopt an evolutionary perspective on digital documents. In their view, digital content is just another genre of materials, to which the important notions of fixity and provenance can be ascribed.

Others take a more revolutionary perspective and are less sanguine about the fit of increasing complex and dynamic digital documents with the standard library

formulation of the document. Lyman [349] observes the manner in which digital documents clash with the information flow and fixed roles mentioned above:

The physical, phenomenological, and epistemological differences between print and digital media have fundamental consequences for the nature of publishing. In print publishing, value is created by the author and the publisher through the reproduction of the mass-produced book in a standard form and format. The computer user, however, controls not only the creation of the text but also its reproduction and distribution through the network.

Bauman [40] concurs with the notion that digital genre are fundamentally different than their predecessors:

Internet genres represent a revolution in text because they signify vastly different organizations of work-in-the-world; they allow for new relationships between reader and writer—between actor and agent, to borrow from genre theory again—that heretofore could not exist in the world.

At first glance, the apparent difference between physical and digital documents lies in their respective mutability. However, a deeper examination reveals that this may not be *the* important factor. Indeed, as Levy claims [338]: “Rather than think of paper as fixed and digital as fluid, we would do better to realize that all documents are fixed and fluid.” For example, although an individual physical book is relatively immutable (even it can be marked up and destroyed), the information content of the book itself may vary across several editions, each containing corrections, revisions, and updates. Thus, a physical manifestation of some work is not fundamentally different from the digital (e.g., PDF) manifestation of that same content that is published online and updated with some frequency.

I argue, instead, that the real inflection point is the transition of the document from its physical and “traditional” digital forms (e.g., pdf, HTML, MS Word) that dominated Web 1.0 to the multi-source mashup “documents” that are increasingly prevalent in Web 2.0. Berman, et al. [50] argue that mashups are evidence of two fundamental

shifts in the nature of media (i.e. documents); the move toward open distribution platforms or “Content hyper-syndication” (through RSS, ATOM, etc.), and the shift to user-generated content or “new platform aggregation” (using tools like Yahoo Pipes³⁰ or the Google Mashup Editor³¹). Examples of mashups range from the relatively mundane single-sourced, periodically-changing Google Maps presentation of property listings³² to highly dynamic, mixed source mashups. One example of the latter is a mashup that presents a changing display of current issues of global interest by selecting photos from Flickr that best match the content of recent “tweets” on Twitter³³.

Is the latter example still a document, even in the more expansive definitions presented earlier? By all accounts, there is almost a complete lack of fixity in this “document” – the periodicity of tweets is completely indeterminate - and is even more uncertain due to the fact that document content is also dependent on the change cycle of Flickr. The nature of its provenance is similarly complex, given the multi-source nature of its two intermediate sources, Flickr and Twitter. The increasing presence such “documents” in our information space will demand new forms of curation, especially to ensure that valuable information persists over the long term. However, the nature of this curation will need to assume forms quite different from traditional methods that assume some level of fixity and some notion of provenance.

³⁰ <http://pipes.yahoo.com/pipes/>

³¹ <http://code.google.com/gme/tour/tour1.html>

³² <http://propertylistingmaps.com/site>

³³ <http://portwiture.com/>

From enforcing uniformity to celebrating diversity

Earlier in this chapter I described the manner in which the library employs its keystone service, the Catalog, to impose “order on chaos”. Cataloging records are surrogates that overlay a uniform model on heterogeneous information. The modern library catalog has proven its worth for more than a century as an outstanding intellectual and technical achievement and as an invaluable tool for discovering information and traversing its semantic relationships [452]. Yet, even in the library community, it is widely accepted that “[t]he catalog is in decline, its processes and structures are unsustainable...” [106]. The obsolescence of the catalog is due to both practical problems with implementing it in the online context and, more fundamentally, with its conceptual incompatibilities with the information model that has emerged in Web 2.0.

The practical problems arose early in Web 1.0 as an increasing number of valuable information resources appeared on the Web. The role of the Catalog as the single entry point to information was challenged as a growing number of people bypassed it for information discovery (this was described earlier in the context of disintermediation).

In response, many libraries attempted to “catalog the Internet” [391], at least the “valuable” resources, effectively trying to “order” this new genre using their traditional rules and tools for imposing uniformity. This effort was ill-conceived and was generally abandoned, except for a very select group of resources, because of the prohibitive cost of conventional cataloging [134] and because of the ephemeral nature of many web resources.

Apropos of the latter, including an information resource in the catalog de facto brings it into the control zone of the library, endowing it with some measure of integrity and assuming a modicum of curation of it. These are quite problematic for ephemeral

digital resources, which may disappear or morph into content far outside the library's integrity standards.

Although the general attempts to catalogue the Internet were relatively short-lived, initiatives to develop methods for describing *selected* digital resources with bibliographic records flourished in the Web 1.0 era of the mid- and late 1990s. These efforts were based on the assumption that some sort of surrogate record was a necessary component of resource discovery. Acknowledging the expense and complicated nature of traditional library cataloging, these efforts developed the alternative of *metadata*, a simplified form of structured record.

The most prominent metadata effort was, and still is, the Dublin Core Metadata Initiative³⁴ (DCMI), which was originally conceived in 1994 to address the shortcomings of early crawler-based web search engines that matched queries to documents using word vectors alone. These methods were not yet supplemented by link-based searching and result ranking techniques [96, 267], and were not well-adapted to the scale and diversity of the web. These technical problems and the enduring influence of the role of the Catalog in history led many assume to that metadata “records, created by content experts, were necessary to improve search and retrieval” [480]. The success of these efforts as a tool to improve resource discovery, especially relative to their expense, is examined at some depth in Chapter 8 and Chapter 10.

The radical transformations in the information model in Web 2.0 have added to the practical problems with cataloging – for example, as noted mashup documents are an order of magnitude more ephemeral than standard HTML web documents. But, more

³⁴ <http://dublincore.org>

fundamentally, the web, and Web 2.0 in particular, reveals and perhaps provides an opportunity to ameliorate basic flaws in the assumptions and implementation of the uniformity principle underlying cataloging.

As Bade describes [38], cataloging has been historically driven by the assumption of the “platonic ideal” of a “perfect record”, containing *intellectually complete* and *objectively assigned* information. But the creation of the perfect record is not the simple result of transcribing “natural” and universal properties of information resources [336]. Sometimes basic cataloging fields such as title and author information have to be inferred according to a set of professional standards [213].

Furthermore, the scholarship on subject classification, arguably one of the most important features of cataloging [452], indicates that this ideal of objective uniformity is a chimera. As Borges [74] stated: “there is no classification of the universe that is not arbitrary and conjectural”, and indeed the uniformity presented by the library catalog is constructed according to dominant cultural norms and biases, that have been projected globally in the sometimes benign and sometimes pernicious effort to create that uniformity over heterogeneous information. As Bowker and Star describe in their landmark book on classification [85], the classification schemes underlying this effort have at times been far from objective and have, at times, been objectionable. These same points are raised by Weinberger in an article specific to Dewey and his subject classification scheme [484] that is widely deployed in libraries. For example, as noted by Kim [264], “In the main class for religion (200–299), divisions from 230 through 289 are dedicated to Christianity while other religions are compacted in 290 to 299.”

In addition to these problematic intellectual assumptions of cataloging, its traditional implementation reflects a set of technical restrictions no longer germane to the online web context. In the physical library context cataloging surrogates were created at

time of accession, bound to a physical artifact, and were relatively immutable (modulo occasional corrections). The uniformity imposed by them, independent of its objectivity or bias, was institutionally determined, and pre-imposed, regardless of the immediate needs of the user or her actual community of context. Notably, in the physical library, these communities of context were at least easy to determine; they were the neighborhood of the local library, or the campus of the academic library.

These technical limitations are no longer relevant in Web 2.0 and they are inconsistent with user information behavior and the nature of online communities. The Web 2.0 space is a collection of self-organizing communities [45, 412] that form and dissolve at a constantly accelerating speed. Instead of having to pre-impose order making on this dynamic environment, increasingly powerful technologies for personalization and customization [107, 455] make it possible to create situated, individualized, and community-specific uniformity and “make it theoretically possible for people to think and act globally and, simultaneously, to act locally and individually” [46]. The library meme-centric activity of teams of professional catalogers institutionally creating “order from chaos” may indeed be part of the past.

Chapter Wrap-up

This chapter explored the full sociotechnical dimensions of the library in both its physical and digital manifestations and of the web as it has evolved from its early form as a largely document network to its current Web 2.0 manifestation. The goal of this analysis was to understand the library and the web as more than institutions, technological phenomena, or physical instances. As described, these more visible aspects of each information system are really evidence of sets of motivating concepts and principles about the organization, management, and availability of information independent of its physical or online form.

I used the notion of a meme map to illustrate the sociotechnical aspects of each. Memes and the maps that illustrate them are multi-dimensional, revealing the internal perception, the underlying concepts and principles, and externally visible artifacts of the respective entity.

Using this tool I asserted that there are five key concepts underlying the library, independent of its physical or digital form. These are the control zone, which is an internally and externally relevant demarcation of the boundary defining the library's scope of management and curation; intermediation, which situates the library as an objective broker and guiding agent between producers and consumers in a unidirectional flow of information; the document abstraction, with its assumptions of fixity and provenance upon which the library's handling of multiple genre of information resources is based; uniformity, which is projected upon a complex and heterogeneous information universe making it possible to assert a common information model to describe the information universe through cataloging; and longevity, which establishes a commitment to the availability and management of information content over the long-term.

As described in the previous chapter, these principles persisted through the transition from the physical to digital library because of the manner in which digital library research was formulated and funded and because of the nature of the communities involved in that process. As a result, many of the key technical artifacts of digital library systems, which are the externally visible realizations of the underlying concepts, are translations of components of the physical library into the online environment.

In contrast, the creation of the web and its evolution over the last two decades has not been constrained in this manner. The original design of the web by Tim Berners-Lee

envisioned a rich, participatory, social environment. The phenomenon that we now refer to as Web 1.0, the read-only document web, was actually a limited instantiation of that vision. In fact, many of the initial web applications and design concepts, such as portals, directories, and gated communities, display traces of the library legacy themselves. The initial document-centric web and the traces of the library paradigm within it facilitated the integration of digital library systems with its technology.

This comfortable coexistence has undergone a profound change in the transition to Web 2.0. As I described, the Web 2.0 meme contradicts virtually all aspects of the foundation of the library (except for longevity, the solution to which in the digital context remains unresolved). Web 2.0 embodies a new information model based on participation, role fluidity, dynamic content, and situated organization and selection. As this information model has spread into all aspects of our information society including politics, business, and scholarship, the viability of digital library applications that manifest a more restricted, institutionally managed library information model has become questionable.

Certainly, these digital library systems themselves could embrace this new information model in the form of Library 2.0. But the questions of Theseus's paradox return. Are these then digital libraries? Or are they just collections of content and services that have been assimilated into the web?

Chapter 4

An Network-Centric Approach for Examining Disruption

The notion of the library and the web as *sociotechnical systems* has been used throughout this dissertation to describe both as more than simply technological phenomena. The origin of this term lies in the fields of Science and Technology Studies (STS) and Workplace Studies, which have developed a number of theories and frameworks for examining technology in its broader societal context.

STS is a broad and diverse scholarly field, the full reach of which is beyond the scope of this dissertation. The reader looking for more a complete summary of STS, especially as it applies to information studies and information systems, is referred to Van House's excellent publication [468]. The general characteristic of STS work that is applicable here is that "technology does not exist in a vacuum, [it] is mutually implicated with ... *ensembles of technical, social, political, economic elements*" [468] (emphasis added). This notion of the *ensemble*, or "heterogeneous network," that contextualizes technological development and deployment was formulated as an alternative to the notion of technological determinism, which presents "technological development as following an autonomous process of change; and technology as an independent force for social change, which causes changes in society, the 'social effects of technology' approach" [468]. This broader contextualized view is the focus of two well-known conceptual frameworks developed under the umbrella of STS: Social Construction of Technology (SCOT) [58] and Social Shaping of Technology (SST) [490].

As noted by Van House there is a naturally close connection between STS and the study of information systems because of the manner in which these systems support the deeply social and cultural act of “knowledge work”, an analysis of which “requires an understanding of peoples knowledge processes, practices, and artifacts ... as well as a reflexive, sociotechnical approach to technology.” As a result, STS and workplace study frameworks have been used by a number of researchers to investigate factors related to digital library design, deployment, and evaluation [63, 262, 263, 470].

The notion of *disruption*, also called instability, disturbance, or resistance, is a common theme in these STS frameworks. Because technology develops and is deployed in the context of a heterogeneous network of other factors such as social role, political context, and economic realities, it is natural for contradictions to arise in this network as components change. These contradictions not only affect the developing technology, but also affect other parts of the network as a whole, changing the role and importance of some of the network components, and possibly leading to their elimination from the network. This perspective is particularly useful for the subject of his dissertation; the changing positions of the library, digital library, Web 1.0, and Web 2.0 information systems as they co-exist and compete within the overall global information context in which creation of information is profoundly affected by digital technology.

This chapter connects the analysis developed thus far to a number of these STS frameworks. Needless to say, an STS-focused dissertation might be dedicated entirely to the application of a single one of these frameworks to an analysis of digital libraries and the web. Being that this is not an STS dissertation, the purpose here is to highlight aspects of these frameworks that are particularly relevant.

The chapter begins with two brief sections that highlight relevant features of Actor-Network Theory and Information Ecology respectively. It then continues with a lengthier section that analyzes library and web disruption using the illustrative tools of activity theory.

Actor-Network Theory

Actor-Network Theory [108, 325, 326] (ANT) developed in the 1980s primarily from the work of Latour, Callon, and Law. It is a conceptual framework for explaining sociotechnical systems. ANT is a complex, nuanced, and controversial topic that combines elements of semiotics, dialectics, Marxism, philosophy, sociology, and critical theory. The curious reader should consult Law's and Hassard's edited collection of essays from various scholars [328] or the many excellent summaries and bibliographies available on the web³⁵. This section will only highlight features of ANT relevant to this dissertation.

The notion of a heterogeneous network of "actants" is core to ANT. The neologism *actant* is an intentional replacement for "actor". It indicates that in addition to the fact that the network is heterogeneous, consisting of humans, organizations, technical artifacts, documents, concepts, meanings, etc.; all of the participants in the network, human and inanimate, may exert influence or have agency in the network.

This notion of human and material (non-human) actors is known as the *principle of generalized symmetry*. As described by Van House [468], "Not only is the distinction between the social and technical artificial, but humans and non-humans are to be

³⁵ Law, one of the creators of ANT, maintains a set of web pages, available at <http://www.lancs.ac.uk/fass/centres/css/ant/antres.htm>, which includes an exhaustive bibliography. The online essay from Garson is also an excellent starting point.

analyzed in the same terms.” The interactions among actants are described by Hanseth and Monteiro [228] as follows: “Stability and social order, according to actor network theory, are continually negotiated as a social process of aligning interests [among the actants in the network].” The result is that these networks are highly unstable, even prone to dissolution, as actant relationships and the nature of actants change [204].

An example of an inanimate, but active, actant is a law or code. The ruling in the case of *Sony vs. Universal Studios* [8], which upheld the rights of consumers to use digital video recorders for timeshifting (i.e., recording broadcast shows for later viewing), changed not only the nature and influence of the recording technologies, but caused substantial realignment of relationships across the entertainment industry.

Generalized symmetry is admittedly a radical and controversial aspect of ANT [468]. However, its relevance to the analysis in this dissertation is intriguing. Consider, for example, the web as an actant that has changed form and caused instability in the broader information activity network that includes the research funders, the mainstream library community, learned societies, publishers, students, scholars, and many others. In its Web 1.0 form as a primarily read-only document network it was mainly “inanimate” with limited capacity for agency. Nevertheless, by providing information seekers direct, rather than library-mediated, access to information it generated some imbalance in the larger network. In the transition to Web 2.0, the web transcending technology and assumed a level of influence and an ability to “act” that is at times indistinguishable and at times more powerful than animate objects. This influence and agency comes from the “collective intelligence” [342] and “cooperative knowledge” embodied in Web 2.0 artifacts such as wikis and more recently in Twitter. The web has become an enabler of and perhaps surrogate for a class of activities that are usually associated with individual or social activity – cognition,

communication, and co-operation – exerting considerable agency on society at large, including the outcome of national elections [126].

Instability caused by such actant transformation forces realignments in the relationships among the elements of an activity system as they try to re-establish balance. This rebalancing is a threat to established components of the network, in this case libraries, whose position of centrality in the information infrastructure has changed radically in the digital context. As noted by Christensen [122] and others who have examined the evolution of infrastructure, established entities often face uncertain futures in the face of such realignments. For example, Edwards, et al. [174] state: “Across virtually every type and class of emergent infrastructure we can identify provisional ‘winners’ ... and ‘losers’.” Furthermore, “Emergent infrastructures function as redistribution mechanisms, reorganizing resource flows across scales ranging from the local workplace or research laboratory to the global economy. Few if any come free of distributional consequences altogether.”

Another relevant concept from ANT is *punctualization* or *black-boxing* [324]. In order to simplify a complex network, it is useful sometimes to treat a subnetwork as a single element in the larger network. For example, the library is internally a complex activity network with many actants – librarians, suites of services, patrons, etc. However, when looking at the higher-level information activity network of the pre-digital era, it is possible to treat the library as a single actant, because of its fundamentally exclusive provision of information service such as preservation, selection, discovery, and the like.

However, such punctualizations are particularly sensitive to disruption as the role of actants in the network change. This can result in *depunctualization*, in which the former atomic entity loses its coherence on the activity network and engages in the

network as a set of deconstructed actants [174]. Latour compares this “to the opening up Pandora’s box” [326], because of the vulnerability it presents to the previously inherently intact actant.

The library in the current web context is a classic case of depunctualization. Services such as information discovery, which were formally uniquely contained to the library and expressed in the catalog, are now in the digital era depunctualized and competing on the open market against commercial search engines [157]. In addition, other services such as reference and access to books are also being exposed to competition, with the appearance of alternatives such as Google Books³⁶. As Christensen notes [122], this deconstructed form of competition is indeed a “Pandora’s box” for institutions like the library for whom nimble reaction at this individual service level is hampered by the constraints imposed by the overall institutional structure.

Competitors such as Google are not bound by either legacy or institutional fabric. As formerly bound and implicit services such as cataloging and access are ceded by the library to more nimble commercial competitors, the viability of the institution as a whole is threatened. The position of the library in this deconstructed state begins to look like Monty Python’s Black Knight, all four limbs chopped off, saying to King Arthur “okay, let’s call it a draw” [6].

Information ecologies

Nardi and O'Day [378, 388] use ecological metaphors to describe the network of entities engaged in an activity system. They state: "we define an *information ecology* to be a system of people, practices, values, and technologies in a particular local environment. In information ecologies, the spotlight is not on technology, but on

³⁶ <http://books.google.com/>

human activities that are served by technology" [378]. They then define relationships among the entities in the ecology using the notion of a *habitation* as follows; "the habitation of a technology is its set of family ties in the local information ecology" [378]. Finally, they use the ecological metaphor to characterize how change in the network impacts the ecology as a whole. "Change in an ecology is systemic. When one element is changed, effects can be felt throughout the whole system." [378].

Gay and Hembrooke [205] also use the ecological metaphor and closely link it to activity theory, which is described in greater detail in the next section. They describe how activity networks change and evolve in the same manner as complex environments, and how this change affects components of the network:

Component systems within ecological systems are characterized by progressive, mutual accommodation and extinction throughout the life of the system... Ecological systems are not harmonious and functioning but have constant tensions, discontinuities, and breakdowns that are necessary for survival and adaptability. [205]

Nardi and O'Day make extensive use of the ecological metaphor for evaluation of digital libraries and for comparing them to their pre-digital form [388]. Extending this ecology analysis to describe the position of libraries and digital libraries vis-à-vis the web is straightforward. Libraries in their traditional form were the "dominant species" in the pre-digital information ecology. As described in the previous chapter, the library meme was congruent with the technological restrictions of the physical environment. The development of the web and its manifestation in the form of Web 1.0 added a competing species to this habitation. Aspects of the library meme were challenged in the same manner that competing species might compete for some food source. For example, users who discovered information exclusively through the library catalog could now use web search engines for that same task. But because of the still relatively primitive nature of the web these competing species still lived in a state of

“peaceful coexistence”. In addition, because, as was described earlier, many of the metaphors in Web 1.0 such as the document were relatively congruent with library metaphors, it was possible to build digital libraries that leverage web technology without significant clash between their respective memes.

The transition of the web to Web 2.0 has destroyed the basis of this coexistence and produced a “survival of the fittest” situation. The web not only consumes the “food source of libraries” – its users, its content, and its services – but it constantly adapts itself and the nature of its information model in a nimble manner that is almost impossible for the library to follow, which is bound by tradition and legacy and the need to maintain support for traditional customers and services. The incompatibility between the library meme and the Web 2.0, 3.0, and beyond memes continues to grow. And as Nardi and O’Day warn, “...parts [of an ecology] can disappear without a trace if they are incompatible with the rest of the system” [378].

Activity Theory

Activity theory [179, 180, 377] is a component of workplace studies, which has its roots in Suchman’s seminal work [448] in the late 1980s. In the words of Heath, et al. [231] the core of Suchman’s analysis was that “we can only understand technologies, and the various formalisms which may be involved, by considering how they feature within practical action and with regard to circumstances in which mundane activities are produced.” This context-aware (situated) perspective on technical change is necessary to understand the “convergence between technological innovation and organizational change” and “direct attention towards the social, the interactional, and the contingent” [231].

Activity theory uses the notion of the “activity” as the core unit of analysis for understanding the contextual, mediated nature of technological innovation and organizational change. According to Kuuti [281] activities have the following properties:

- They are motivated by the desire to transform an object toward some desired state.
- They have a subject who is the agent of and who understands the purpose of the activity. The subject may be individual or collective.
- They involve other participants who may not all understand the motivation behind the activity, or recognize even the existence of it.
- They exist in a broader material context, transform that context, and are transformed by it.
- They exist along a time dimension, due to the fact that they are affected by historical circumstances and produce future consequences.
- They are subject to mediation by tools, by cultural norms, and by relationships with other participants.
- They encounter instability due to contradictions with mediating factors that ultimately change the nature of the activity.

The notion of mediation is illustrated in its simplest form in Figure 13. As shown, a *subject*, for example a researcher³⁷, wants to act on some *object*, for example access some information resource like a journal article, to accomplish some *outcome*, for example write a research paper. That access is *mediated* by technologies and tools current at that time. In the case of the traditional library these are the physical journal,

³⁷ I will use scholarly research as the exemplary activity throughout the remainder of this section.

the library cataloging system, and the shelving system at the library. In the case of the web these mediating technologies are the browser, a search engine like Google scholar that allows the scholar to find the article, and the repository or website in which the article is stored.

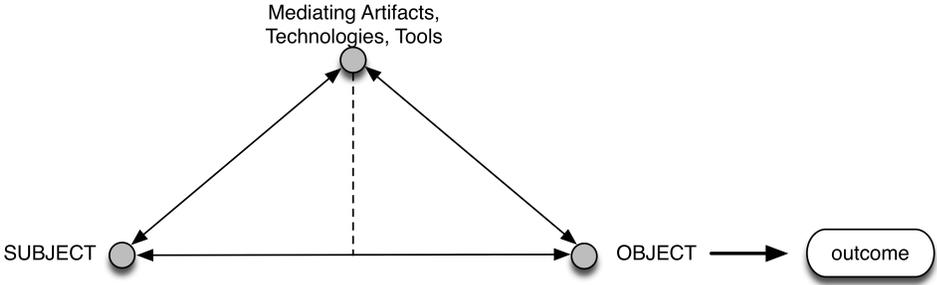


Figure 13 - Simple mediation in an activity system

Mediation is actually more complex than this simple example; activity theory posits that activities are *triply-mediated* [441]. This is shown in Figure 14. A subject's effort to manipulate some object towards a selected outcome occurs in the context of the *broader community* in which they act. Their interactions with that community are mediated by the *rules and standards of community behavior*. For example, the scholar trying to write a research paper in the earlier simple example performs the task in the context of his or her broader scholarly community according to the rules and standardized behaviors of that community; rules including notions of integrity, proper citation, acknowledgment, reward systems, and the like. Furthermore, the interactions of the community (which includes the subject of the activity) vis-à-vis the object of the activity are mediated by the *standardized divisions of labor* within the respective community. For example, in the traditional library environment access to a journal might be mediated by a gatekeeper librarian and dissemination mediated by commercial and society publishers.

The remainder of this chapter uses the template illustrated in Figure 14 to examine information activity systems in the traditional library, the Web 1.0 context, and the Web 2.0 context, and to demonstrate the locations of disruption to the library meme as the web has changed. Libraries house a broad variety of activities ranging from preschoolers learning about books, to casual browsing for popular fiction, to public school students writing research papers, and finally to PhD scholars conducting original research. In this section I will focus on the latter, the process of scholarly research and publication. Support of these activities is one of the primary missions of the academic research library. The wholesale movement of scholarly publications to the online environment, and the increasing reliance by scholars on the web for other aspects of research such as collaboration and data sharing, is currently the focus of considerable research attention [80].

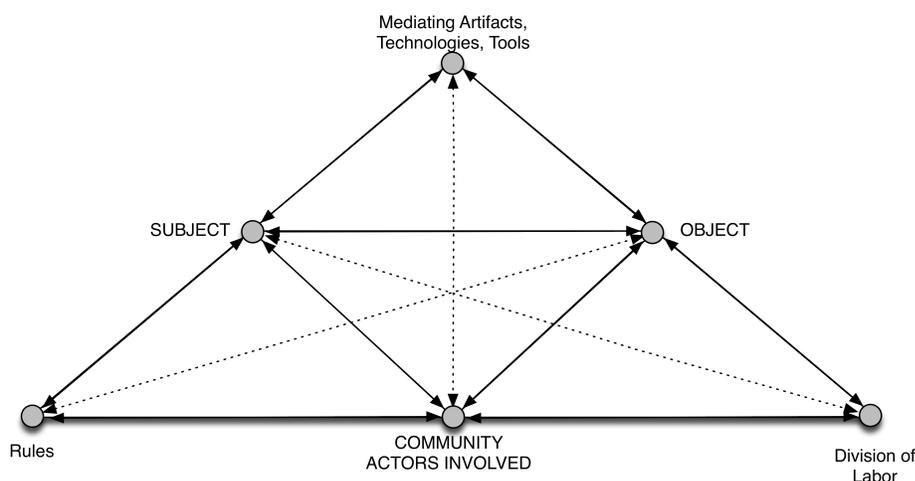


Figure 14 - Triple mediation in an activity system

Pre-web scholarly communication activity system

Figure 15 shows the components of the activity system for scholarly research in the pre-web, physical library context. This drawing and subsequent drawings are

simplified by collapsing the object and outcome into a single box. The components of interest in the illustration are as follows:

- The *subject* of the activity is the scholar, or student scholar, trying to discover and access resources for research.
- The *mediating tools and artifacts* for doing this are the physical information resources in the library, that are discovered using the library catalog or indexing and abstracting publications, and are accessed either from the local library or through inter-library loan.
- Scholarship takes place in the context of a *community of actors* including local colleagues and professional societies, and external communities such as librarians and publishers.
- Their actions vis-à-vis these communities are mediated by the standard *rules* of scholarly communication including academic integrity, and peer review.
- In the pre-web era, there was a strict *division of labor*: scholars performed the research and authoring task, but publishers controlled the actual distribution and editorial task, and librarians mediated information access.

Figure 16 modifies the previous figure slightly to illustrate the publication (production) activity, once the research (consumption) activity is complete. The depiction of these two activities, research and publication, as distinct is intentional, depicting the reality of their separation in the pre-web era. Changes from Figure 15 are as follows:

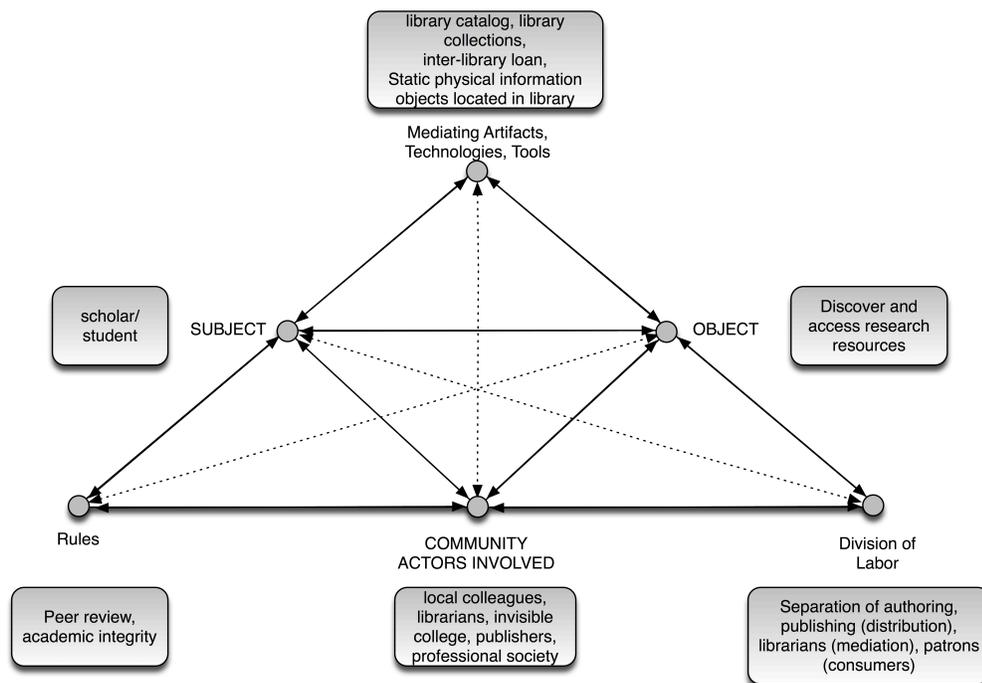


Figure 15 - Library-centered research

- The *objects/outcomes* are the set of functions of scholarly communication – registration, certification, awareness, reward, and archiving – as defined by Roosendaal and Guerts [421].
- The *tools* for doing this in the pre-web information system were the physical journals, conference proceedings, and monographs, available by subscription or in the library.
- Copyright has been added as one of the mediating community *rules*. Without any serious competition from alternative publication venues, journals and other publications imposed copyright transfer rules in which scholarly authors generally signed over rights to publishers.

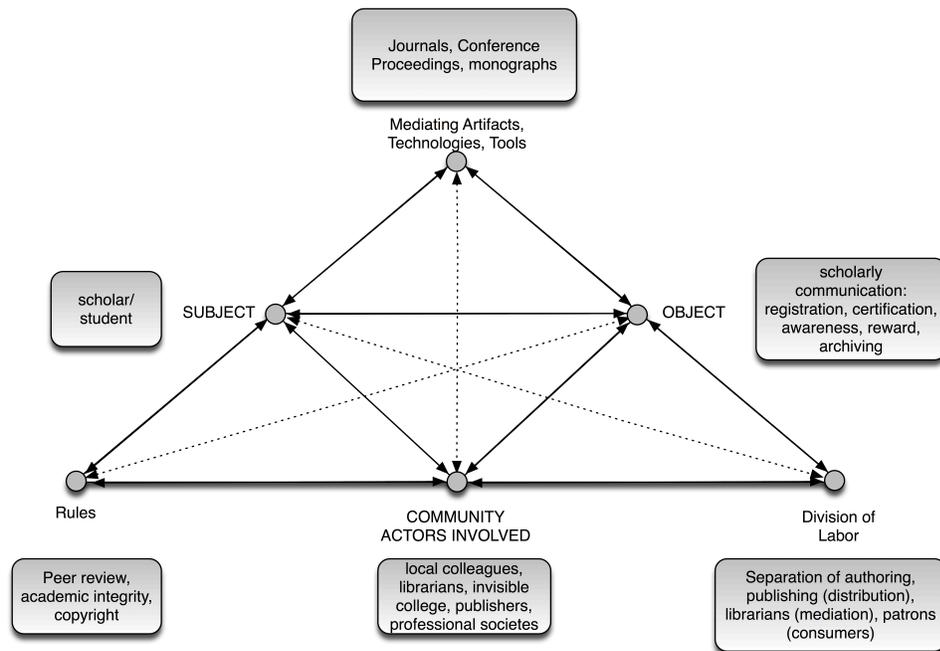


Figure 16 - Pre-web publication

Web 1.0 scholarly communication activity system

The emergence of the web in the early 1990s had two important effects on the scholarly communication activity system, the result of which was a hybridized system. First, although most “formal” publication (that which played a role in tenure decisions for example) was still controlled by publishers, these publishers increasingly moved their journals and conference proceedings to the online environment, mainly in the form of digital libraries (for example, the ACM digital library³⁸). Second, a separate all-digital “ePrints” or “gray literature” publication channel developed. Led by the arXiv³⁹ in the physics community, these “ePrints repositories” were part of a nascent *open access movement* [247] that leveraged web technologies as a means of disrupting

³⁸ <http://portal.acm.org/dl.cfm>

³⁹ <http://arxiv.org>

the perceived stranglehold that publishers had over the scholarly communication process [229].

Open access advocates resisted publisher-imposed copyright transfer requirements on moral and scholarly grounds, and saw these new open access repositories as not only a way of reasserting scholars' ownership of the products of their work, but also as a way of providing rapid turnaround to latest research results. The open access movement has had substantial disruptive influence in a number of key fields, such as physics, in which the online repositories have become the dominant publication media. Other effects of the open access movement include changes in the nature of copyright itself as reflected by the work of the Creative Commons⁴⁰. The sociotechnical aspects of the open access movement, its enablers, and results are active research areas [87, 150, 473].

Figure 17 illustrates the Web 1.0 scholarly research process, an evolution from that illustrated in Figure 15. The changes from this earlier activity system are as follows:

- The set of *mediating artifacts, technologies, and tools* have been enhanced with a set of digital technologies and applications including web search engines, publisher digital libraries, and ePrint repositories. Most significant, however, is the addition of “anywhere access”, eliminating the restriction that research could only occur in the physical library.
- As a result, there is a fundamental change in the *division of labor* with the addition of direct scholar searching and access without mediation, bypassing the gatekeeper role of the library.

⁴⁰ <http://creativecommons.org/>

- Finally, there is the addition of the open access culture to the set of *rules* that mediate community behavior.

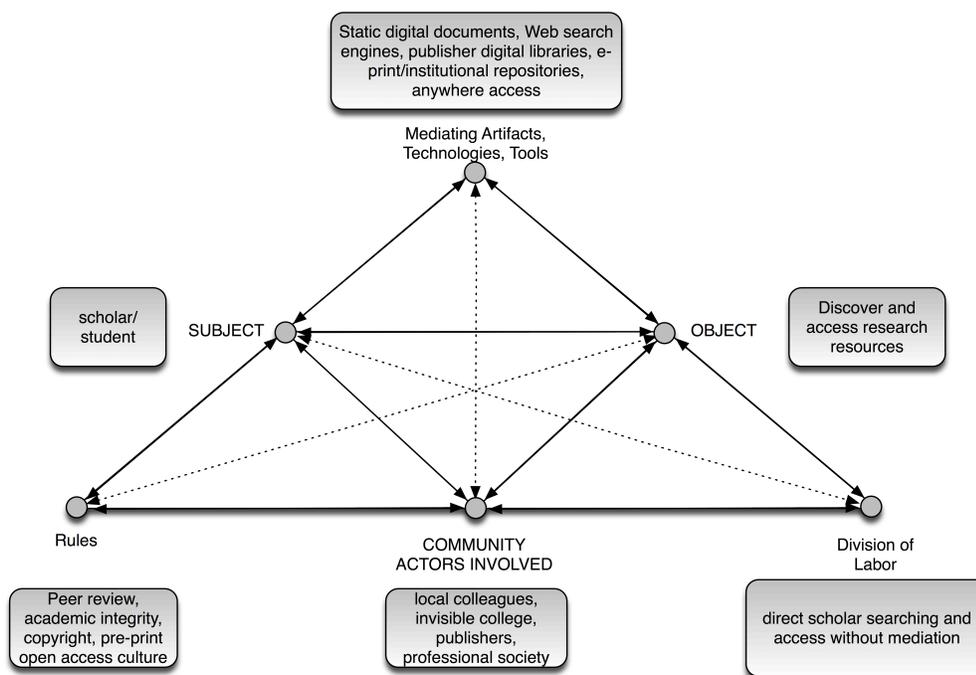


Figure 17 - Web 1.0 research

Figure 18 is a minor modification of the previous figure, illustrating the Web 1.0 publishing activity. The major change is the modification of the *division of labor* area in which the publisher is bypassed for ePrints and gray literature.

Disruption from Web 1.0

Disruptions or contradictions are usually illustrated in Activity Theory diagrams among the internal components of activity systems. Figure 19, adapted from [179], shows an example of such a disruption, in which the red lightning-shaped arrows indicate contradictions between the mediating tools and the object, and between the object and the division of labor that may occur due to changes in these components of the activity system. However, some researchers such as Spasser [441], Gay and

Hembrooke [205], and Boer, et al. [70] have extended the notion of an individual activity system into an *activity network*, a group of related activity systems. According to Gay and Hembrooke [205], this makes it possible to indicate how individual activity systems are *situated* in time and space alongside coexisting activity systems. I will use this notion of an activity network here to illustrate the disruption between the different, and co-existing, activity systems described above; the library, Web 1.0, and Web 2.0.

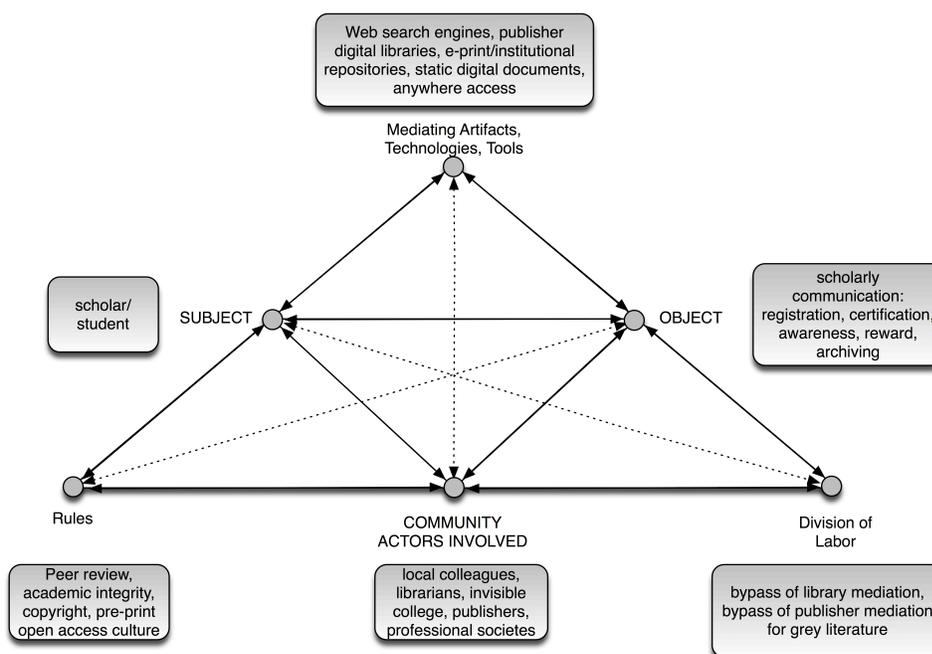


Figure 18 - Web 1.0 publication

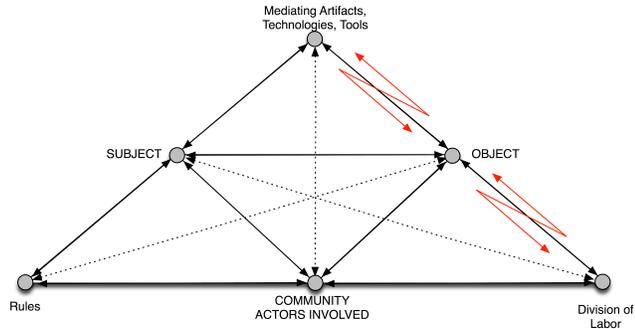


Figure 19 - Internal disruption to an activity system

Figure 20 uses this technique to show the disruption or contradictions between the library based research and publication systems illustrated in Figure 15 and Figure 16 and the corresponding Web 1.0 systems illustrated in Figure 17 and Figure 18. For simplicity's sake the research and publication activities have been collapsed into one activity. As indicated, there is disruption between the two systems in two of the core library meme concepts described in Chapter 3. The notion of the “control zone” has been disrupted via the removal of the restriction that resources were only located in the physical library, and the emergence of “anywhere access”. In addition, the notion of “intermediation” was disrupted by the changes in the division of labor, whereby scholars can directly access information via web search engines and web access, and by direct publishing in ePrints repositories.

While the impact of this disruption is notable, it can not be characterized as revolutionary. In fact, as we have noted elsewhere “The current [digital] scholarly communication system is nothing but a scanned copy of the paper-based system” [466]. By this we mean that the medium may have changed from paper and ink to bits, but the fundamental structure and institutional model was preserved.

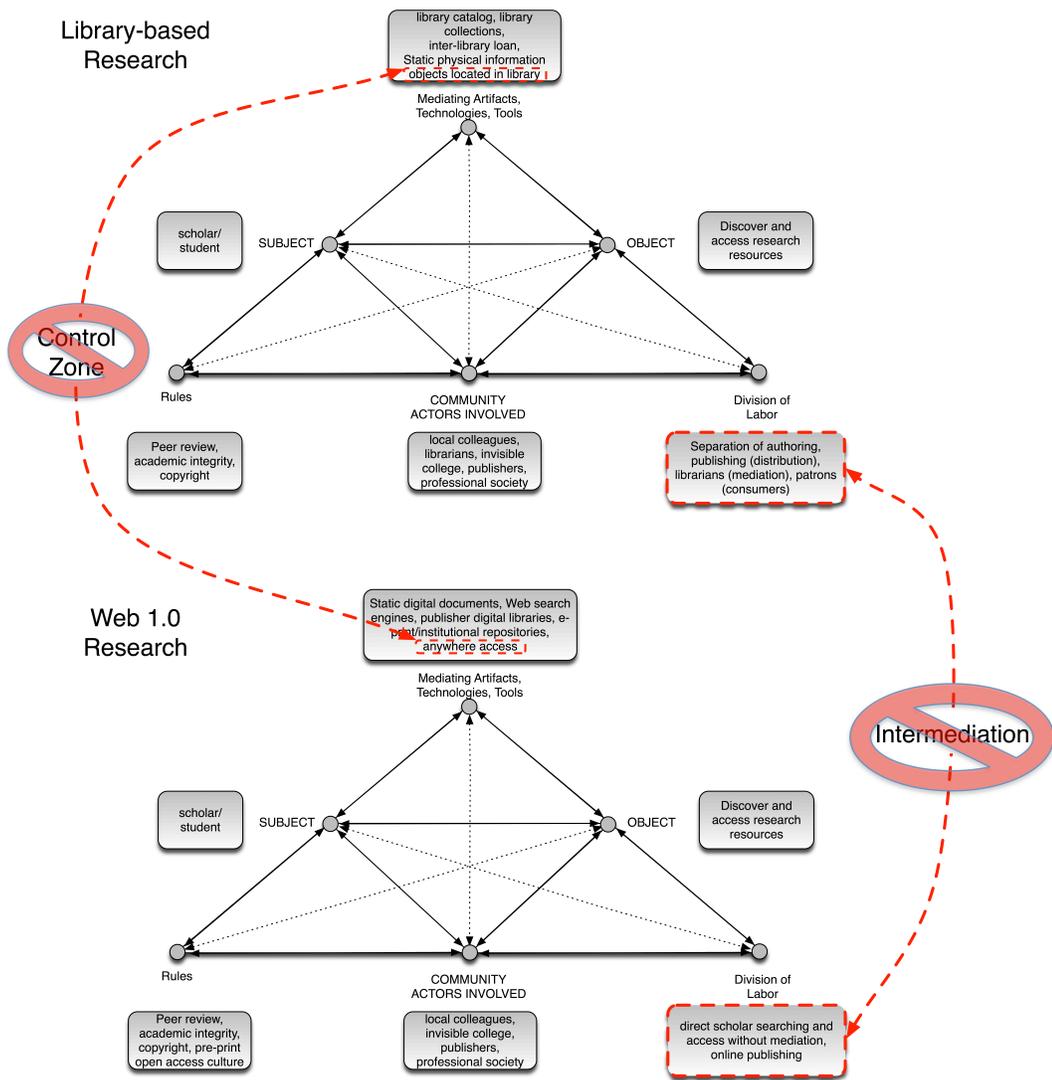


Figure 20 - Web 1.0 disruption

Web 2.0 scholarly communication activity system

Figure 21 illustrates the scholarly communication activity system in the Web 2.0 environment. The combination of research (consumption) and publication (production) in one diagram is intentional to indicate the essential merging of these two activities in the ideal Web 2.0 environment. Clearly, this ideal environment exists only in exemplary cases, and instances of intermediation, traditional scholar copyright transfer, and strict separation of roles are still prevalent. However, the potential for this ideal exists,

and there are an increasing number of realizations of it such as eScience blogs and wikis [88] and innovative publishing systems such as the Public Library of Science⁴¹.

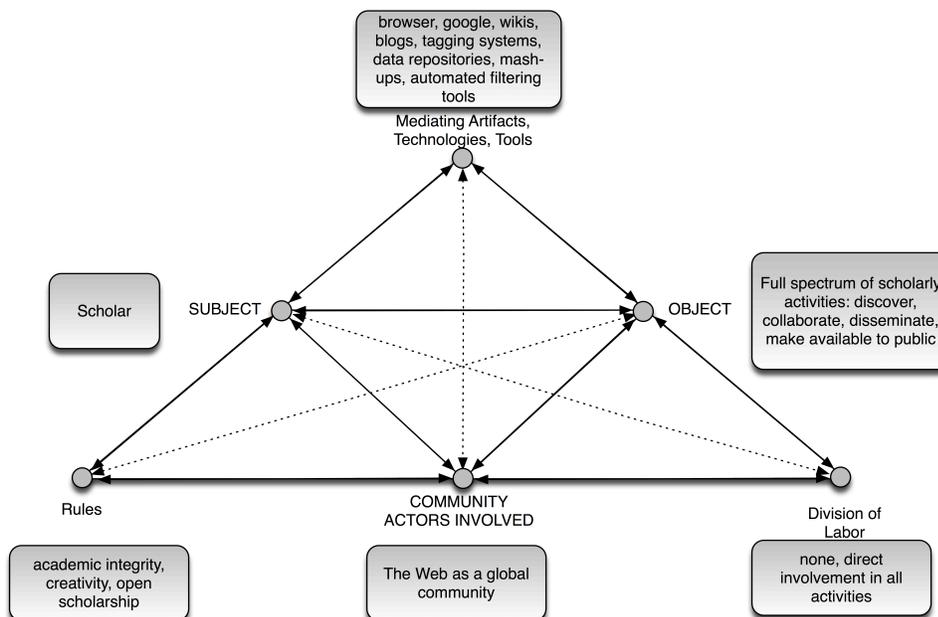


Figure 21 - Web 2.0 scholarly communication

The key features of this Web 2.0 activity system are as follows:

- The *mediating artifacts and tools* include the full suite of Web 2.0 applications such as wikis, blogs, tagging systems, etc.
- No *division of labor* is enforced (however it may de facto exist in some situations), and scholars are able to directly participate in all relevant activities such as discovery, collaboration, and dissemination.
- The scope of *community actors involved* has opened up from the limited scholarly community to the entire web community, who can take advantage of

⁴¹ <http://www.plos.org/>

and use the products of scholarly activity, and even contribute to scholarship through innovative *citizen science* activities such as SETI@home⁴² and Project FeederWatch⁴³.

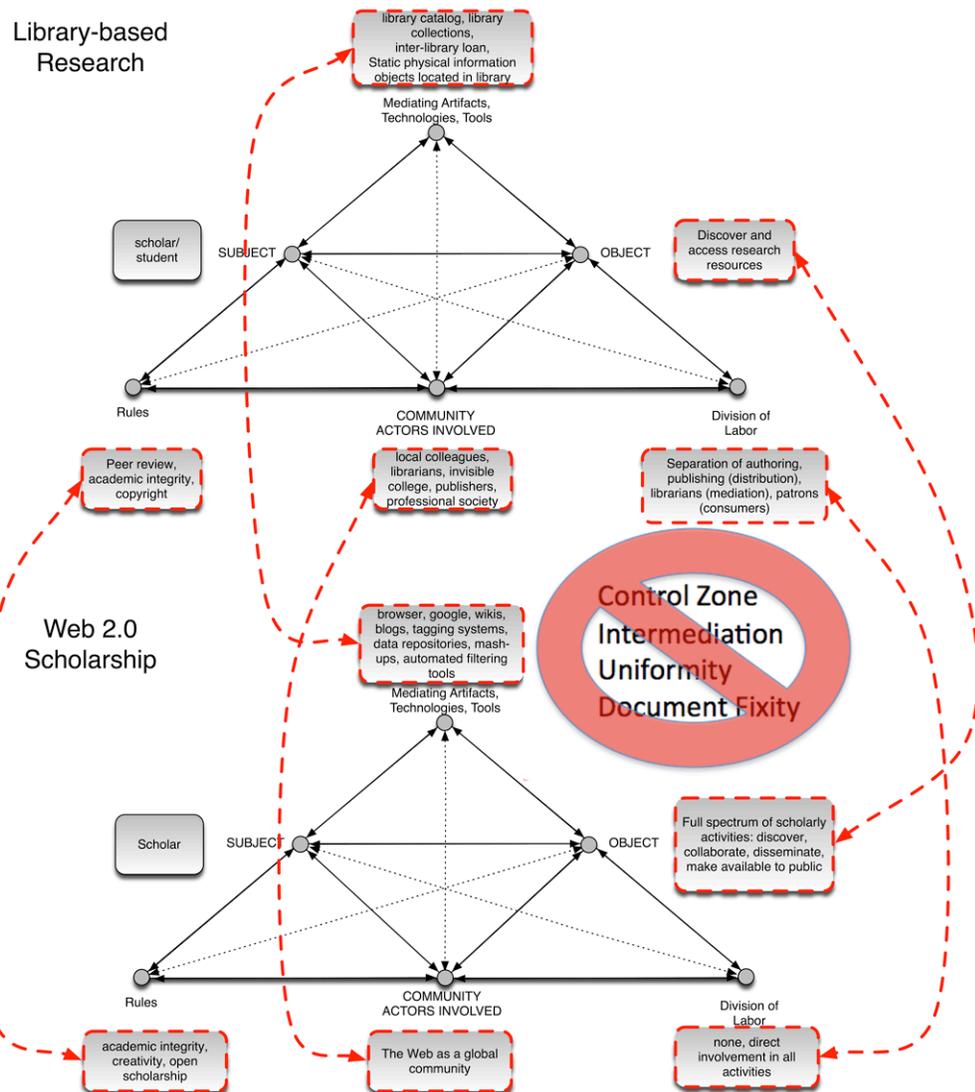


Figure 22 - Web 2.0 disruption

⁴² <http://setiathome.ssl.berkeley.edu/>

⁴³ <http://www.birds.cornell.edu/pfw/>

Disruption due to Web 2.0

Figure 22 shows how the Web 2.0 paradigm disrupts all aspects of the library based scholarly communication activity system. As indicated, all core concepts of the library meme, except for longevity, have been disrupted.

Chapter Wrap-up

The theories and frameworks developed within the field of Science and Technology Studies (STS) have been applied by several scholars to understand the evolution of and use of information systems. In this chapter, I have described the manner in which the analysis developed in earlier chapters next with three particular STS frameworks; Actor-Network Theory, Information Ecologies, and Activity Theory. As described, the latter is particularly useful for illustrating the nature of and degree of disruption to the process of scholarly research and publication that was formerly exclusive to the academic research library and formal publishers.

Chapter 5

Review of Related Work

This chapter reviews work in various areas related to the material covered in this dissertation, some of which is cited in other chapters. The work is divided into four sections: technologies for interoperability in networked information systems, historical overviews of digital library research, evaluations of the nature and the impact of the Web 1.0 to Web 2.0 transition, and digital libraries as sociotechnical systems.

Technologies for interoperability in networked information systems

The notion of “networked” information – information units inter-connected by various organizational paradigms – precedes the invention of the computer, Internet and the later World Wide Web. Raymond [474] provides a relatively complete history of the pre-digital historical origins of underlying concepts.

A fundamental underlying notion is *hypertext*, “connect[ing] text across more than one document boundary” [136], thereby breaking down the traditional atomistic notion of the document (e.g., book, monograph, etc.). The origins of this notion can be traced to the visionary, and often forgotten, early 20th century work of Paul Otlet [392]. More frequently cited as the origin of hypertext is the microfilm-based *memex* envisioned by Vannevar Bush in his famous post-World War II article [105] about harnessing wartime science for peacetime challenges.

The invention of the computer and the spread of desktop workstations interconnected by a common network – the Internet – provided the foundation for the realization of Otlet’s and Bush’s prescient ideas. The two most notable early pioneers and creators

of demonstrable examples of the power of inter-linked digital information are Ted Nelson, with his Xanadu system [384], and Douglas Engelbart [177, 178], who is perhaps best-known for his invention of the computer mouse.

Coincident to this, the idea of extending the notion of the library to computers and networks began to take form in the 1960's. Early manifestations include the visionary designs of Licklider [346], the development of library automation systems and the machine-readable catalog [78], and the invention of modern information retrieval [428]. These were followed by a number of pre-web or non-web-based digital or electronic library concepts, experiments, and applications. These include the rather detailed design of a digital library system (and a plan for a digital library research program) by Kahn and Cerf [256] and client-server-based systems such as Schatz's Telesophy [431], the System 33 Document Service from Xerox [411], the RightPages system [237] from AT&T Bell Laboratories, and the CORE project [183] from BellCore, Cornell University, OCLC, and others.

The introduction to this dissertation describes the reports and workshops that led to the contemporary (post 1992) digital library initiatives. As described the notion of *interoperability* – providing the user with a seamless experience as they use heterogeneous, distributed information services (discovery, access, browse, etc.) – has been a central aspect of digital library research. Paepcke, et al. [394] describe the issues and historical lineage of interoperability in the digital library community. As described there and in Lynch, et al.[354], interoperability exists across a spectrum. At the lowest level it provides minimal interfaces and tools with which humans can navigate and infer coherence across multiple systems. The more common intermediate form is *syntactic* interoperability whereby common protocols, metadata formats, and digital object exchange standards provide a modicum of coherence across

systems. Considerably more complex and still the subject of research is the notion of *semantic* interoperability, which according to Paepcke, et al. [354]

... deals with the ability of a user to access, consistently and coherently, similar (though autonomously defined and managed) classes of digital objects and services, distributed across heterogeneous repositories, with federating or mediating software compensating for site-by-site variations. It also extends beyond passive digital objects to actual services offered by specific digital library systems.

This section focuses on related work on syntactic interoperability in the digital library community, the primary locus of the work described in this dissertation, while touching on tangential results in the web community. This work is divided into five categories: federated interoperability infrastructures, modeling of compound digital objects, metadata standards and ontologies, repository architectures, and semantic modeling.

Interoperability infrastructures for federated digital libraries

As described by Leiner [330] a federated (or confederated, the term used by Leiner) infrastructure seamlessly links distributed library services and content. The Dienst architecture and protocol [151, 289] and its instantiation in the NCSTRL global digital library [153], described in Chapter 7, is an example of an infrastructure for federated digital libraries. The reader is directed to that chapter for more details.

A distinguishing aspect of Dienst is its foundation in the web architecture – HTTP, HTML, and URIs – thus allowing accessibility to its functionality through standard web browsers. Two systems contemporary to Dienst, also built on web technologies and focused on computer science content, were the Wide Area Technical Reports Service (WATERS) [356] and The Unified Computer Science Technical Report Index (UCSTRI) [471]. Both of these systems had lower functionality than Dienst. They were by and large interfaces to central indexing sites (no federated search, no

collection model) and provided simple document access through HTTP and/or FTP URLs (no document model).

A considerably more complete, and complex, interoperability digital library infrastructure, which arose from Phase I of the Digital Library Initiative (DLI-1), was the Stanford InfoBus [422]. The goal of the InfoBus was to “extend the current Internet protocols with a suite of higher-level information management protocols” [422]. The use of the root “bus” in the name InfoBus connotes the same meaning as with hardware “buses”, a pluggable infrastructure. InfoBus was implemented over a Java-based CORBA [396] foundation and was used as a vehicle for tying together a protocol layer for managing items and metadata, and services such as search, payment, and rights and obligations. While these experiments produced demonstrable results, the InfoBus was not used for any widespread production environments, perhaps due to the decreasing use of CORBA in favor of more web-based paradigms or due to the complexity of the interoperability paradigm.

Perhaps as a reaction to the complexity of their previous full-functionality interoperability effort, the follow-on Stanford Digital Library project under Phase II of the Digital Library Initiative (DLI-2) provided a more limited federation mechanism, the Simple Digital Library Interoperability Protocol (SDLIP) [393]. SDLIP limits functionality to “search middleware” – providing protocols and software for mediating search interactions among several information sources. It was proposed as a mid-point between the widely-used (in the library community) but quite heavyweight Z39.50 standard [352] and extremely light-weight single text box, web-crawler based search paradigm. SDLIP was the basis for experimentation among several DLI-2 projects. However, as we noted in [170, 171] federated searching is fraught with performance, reliability, and interface problems, making it considerably less practical for

widespread use than centralized indexing (such as those used by crawler-based search engines) or harvesting (using mechanisms like those defined by the Open Archives Initiative Protocol for Metadata Harvesting [303]). Thus, while interesting technically, SDLIP has not found widespread use.

Another interoperability result of DLI-1 was the distributed agent architecture from the University of Michigan [60]. This architecture relied on cooperating system modules (or agents) to mediate actions across heterogeneous process and tasks. The intention was complete flexibility since a new function could be added to a digital library system via the addition of a new agent that understood basic interaction protocols. Only a few prototype implementations were successfully constructed.

The relative simplicity, flexibility, and web integration of the Dienst architecture led to its extension and adaptation in a number of follow-on systems and architectures that extend its protocol and services. OpenDlib⁴⁴ [112], a product of considerable EU funding, is an architecture for linking digital library services. A even more direct descendent of Dienst is the DPubS⁴⁵ (Digital Publishing System) at the Cornell University Library, which is an open-source software system designed to enable the organization, presentation, and delivery of scholarly journals, monographs, conference proceedings, and other common and evolving means of academic discourse.”⁴⁶

A final piece of related work in the area of federated digital library infrastructure is the current effort to develop a “digital library reference model” [109]. This work grows out the substantial work on federated digital libraries in the U.S., Europe, and

⁴⁴ <http://www.opendlib.com/area1/project.html>

⁴⁵ <http://dpubs.org>

⁴⁶ <http://dpubs.org/about.html>

elsewhere (rooted in part in the Dienst work) and is an attempt to formalize the structure and characteristics of such systems and provide the theoretical basis for future interoperability efforts.

Modeling of Compound Digital Objects

The notion of a document model has been a focus of digital library research from the beginning. The work is motivated by the opportunities for an expanded notion of “content” or “document” once the constraints of physical media are eliminated. In general, the work has focused on the notion of a compound digital object, a container for aggregating data, metadata, rights information, administrative (e.g. logging) data, and other related data streams.

This dissertation describes the various aspects of my work that fall into this area. Chapter 7 describes the Dienst architecture, which includes a compound document model, visible through the protocol. Chapter 8 describes work in the Dublin Core Metadata Initiative, which included the so-called Warwick Framework, a packaging abstraction for multiple metadata formats. Chapter 9 describes Fedora, a repository system with a powerful document model that not only includes compound aggregations, but also allows dynamic disseminations (i.e., linkages of static data streams and distributed web services), and expresses semantic relationships among digital objects. Chapter 12 describes work in Open Archives Initiative – Object Reuse and Exchange, which defines a standard for identifying and describing compound objects in terms of web architecture fundamentals.

The roots of this work and related efforts lie in what is known as the Kahn-Wilensky Framework (KWF) [255], which was a result of the Computer Science Technical

Reports Project⁴⁷ (CSTR) . KWF defines at an abstract level the notion of digital objects, identified via uniform naming system known as Handles [449], that contains key metadata, multiple other packages of data and metadata, and possibly recursively containing other digital objects. The KWF does not cover implementation details. However, in an early paper [290, 449] we outlined these issues, and the KWF has had a considerable influence on our follow-on work described in the remainder of this dissertation.

Nelson, in his PhD work [382], created a related compound object architecture inspired by both the KWF and Dienst work. His notion of “buckets” exist in a “Smart Objects and Dumb Archives” model that pushes intelligence or functionality usually found in repositories or archives down into the object. The motivation is to enhance the long-term survivability and portability of the compound objects. The buckets architecture was leveraged in a number of other digital library experiments including one where bucket functionality was expanded to dynamically change inter-bucket relations according to user retrieval patterns [73].

Over the past decade a number of compound object formats have emerged in the digital library community. One of these is the Metadata Encoding and Transmission Standard (METS) [367] that arose from experimentation with the KWF in the Library of Congress National Digital Library Program (NDLP)⁴⁸. METS is XML-based and accommodates the encoding of various forms of metadata for a digital object including bibliographic, administrative, rights, and structural. METS is used as the default storage format for the popular DSpace [439] institutional repository software. One of

⁴⁷ <http://www.cnri.reston.va.us/cstr.html>

⁴⁸ <http://lcweb2.loc.gov/ammem/dli2/html/lcndlp.html>

the common criticisms of METS is that it requires the classification of object components into pre-specified metadata categories, limiting its flexibility. Also, for some applications its hierarchical rather than graph-based model is too restricted.

Other communities have created or adopted their own object formats: examples include IMS-LOM [4], from the Learning Objects community, and MPEG-21 DIDL [245], originally from the consumer electronics community and adapted to the DL environment by Los Alamos National Laboratory [47]. Although the syntax and application domain for these formats differ, they all have goal of combining descriptive, structural and administrative metadata to represent digital manifestations of “intellectual works”.

Despite their utility, these formats all share a common problem, which motivated our OAI-ORE [316] work described in Chapter 12. There is no clear mapping of these compound objects into the web architecture. The result is that agents and services, such as crawlers for search engines, are unable to interpret the contents of these compound object representations without special provisions, which are generally not implemented and are deemed undesirable (each special case interferes with the efficiency and scalability of the crawler). As a result objects represented in these formats are frequently invisible to widespread web search techniques.

Metadata standards and ontologies

The term “metadata” is used differently across a number of contexts, among them scientific data, software engineering, databases, and digital libraries. The use of the term here is restricted to the (digital) library context. A NISO report defines DL metadata [13] as “... structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource”. In addition, it falls into three main categories: descriptive (sometimes also called

bibliographic), structural, and administrative (including rights metadata and preservation metadata).

Historically metadata has been positioned as a successor and supposedly simpler alternative to traditional library cataloging. Library cataloging standards include both the rule sets, the most widely used of which is the Anglo-American Cataloging Rules Version 2 (AACR2) [213], and machine encoding rules, including the widely used MACHine-Readable Cataloging (MARC) format⁴⁹.

As stated elsewhere in this dissertation, metadata has been both a popular and controversial subject in the digital library research community. Outside the immediate digital library research community (e.g., in the general Internet community) its utility as a general tool for resource discovery has been profoundly criticized [165, 434] and it has been generally rejected by virtually all popular crawler-based web search engines because of quality and integrity problems in favor of content-based and link-based indexing.

Nevertheless, the work of the Dublin Core Metadata Initiative⁵⁰ stands out as one of the most well known results and ongoing efforts of the Digital Library research efforts. Chapter 8 provides more details on Dublin Core, describes my personal work in that initiative including work on the so-called Warwick Framework [286], and includes criticisms of the Initiative and the use of Dublin Core metadata in general. The chapter also includes a description of an alternative more descriptive metadata ontology that is the result of my work on the Harmony/ABC research effort.

⁴⁹ <http://lcweb.loc.gov/marc/marc.html>

⁵⁰ <http://dublincore.org>

A complete review of related work in the area of digital library metadata would be inappropriately lengthy and out of scope for this dissertation. The interested reader is referred to the aforementioned NISO report [13] for more details. As an alternative, I will enumerate a few of the more popular and active metadata efforts and standards that coexist with Dublin Core:

- Text Encoding Initiative (TEI)⁵¹ is a standard for marking up electronic text with a focus on humanities scholarship. It includes the ability to incorporate bibliographic information into the header of the marked up text to assist in the discovery and use of the included text.
- CIDOC/CRM [142] is a formal ontology for describing the structure and relationships among objects, which developed in the cultural heritage community. It is a complex description format that encompasses temporal, part/whole, and epistemological relations.
- Metadata Object Description Schema⁵² (MODS) was developed by the library community as a simpler derivative of MARC that is more expressive than Dublin Core for the description of electronic resources⁵³. It is encoded in XML.
- <indec> [425] is a format developed with support by the European Commission to support the encoding of rights information for electronic commerce.

⁵¹ <http://www-tei.uic.edu/orgs/tei/>

⁵² <http://www.loc.gov/standards/mods/>

⁵³ The creation of the MODS initiative is ironic since the original motivation of Dublin Core was the description of electronic resources, and the qualification effort within DCMI was motivated by a desire for more expressivity than the core elements.

- ONIX (Online Information Exchange)⁵⁴ is a metadata format developed by publishers to distribute electronic information about their books to booksellers, other publishers, and other organizations involved in both industry transactions. ONYX is widely deployed and is the basis of cataloging in Amazon.
- MPEG Multimedia Metadata⁵⁵ is a format for describing the structure, content, and relationships in multimedia audiovisual objects such as pictures, music, audio, speech, or video.
- Functional Requirements for Bibliographic Records (FRBR) [3] is a model developed by the International Federation of Library Associations and Institutions (IFLA) in an effort to modernize the model underlying the catalog. It is based on an entity-relationship framework. One of its major contributions is the distinction between the abstract notions of works and expressions – intellectual concepts and their realization in different genres such as musical scores or screenplays – and the concrete notions of manifestations and items – the translations and material instances of these abstractions such as a book or DVD. FRBR has been used in a number of experiments with new cataloging tools [454]. However, with decreasing interest and commitment to the notion of the catalog in general (in contrast to the increasing influence of web-based search engines), there is some question about the viability of any cataloging model including FRBR.

⁵⁴ <http://www.bisg.org/documents/onix.html>

⁵⁵ <http://www.multimedia-metadata.info/>

Repository architectures

Chapter 9 describes our work on Fedora, an architecture that is both a digital object model and a repository architecture. The notion of a repository has been central to most digital library architecture and, in fact, as described earlier, digital library architecture can be described in general as “repository-centric”.

Despite its wide use, there is some uncertainty in the digital library community about what a repository actually is. According to the KWF [255] “a repository is a network-accessible storage system in which digital objects may be stored for subsequent access and retrieval.” This definition, taken at face value, implies containment. However, arguably, a repository, in the manner of Fedora, may either contain (i.e., store on its own discs) objects, or may reference surrogates or references to those objects or to parts thereof [405]. By this logic then, it makes more sense to think of the repository of a service interface (a set of APIs) for the deposit, access, and management of digital objects regardless of their location of storage.

Digital repositories have proliferated in the library community with the introduction of the notion of an *institutional repository*[143, 253, 351]. These are part of an effort by University and research organization libraries to capture intellectual output of resident faculty and researchers “upstream” and make it accessible in a manner independent of its publication in more formal (e.g. Journal) publication. The institutional repository movement can be seen as one part of a broader *open access* movement [247, 491] that promotes free and open availability to the results of scholarship so that those results can be mechanically harvested [314], reused, and become the inputs for new scholarly work.

Fedora coexists with a number of other institutional repository architectures including:

- DSpace [439] developed by MIT and Hewlett-Packard. Although DSpace lacks a number of advanced features in Fedora, such as a flexible object model and open API, its integrated packaging of a professional user interface with a repository system has made it the most popular architecture for institutional repository applications.
- ePrints⁵⁶ was developed at the University of Southampton specifically to promote open access to scholarly publication. Like DSpace it is a software package that includes the underlying repository, user visible interface, and workflow layer.
- AdoRE [461] developed at the research Library at Los Alamos National Laboratory is a digital repository focused mainly on archiving and preservation. It takes a “write-once/read-many” storage approach and makes generous use of OAI-PMH [465] for protocol-based access to complex digital objects.
- Greenstone [494] developed at the University of Waikato in New Zealand is a repository application and suite of software for building complete digital library systems. It has been specifically constructed for ease-of-use, ease-of-installation, and modest hardware requirements to make it useful for developing countries and empower them to cross the “digital divide”. Because of this focus, Greenstone is supported by UNESCO.

Semantic Models for Digital Libraries

Chapter 11 of this dissertation describes the use of Fedora and Semantic web technologies to build a digital library based on an “information network overlay”.

⁵⁶ <http://www.eprints.org/>

Chapter 12 also covers the use of Semantic Web technologies as a means of integrating digital library compound objects into the web architecture.

This work builds on and is related to the overall Semantic Web Activity⁵⁷ of the World Wide Web Consortium (W3C). The efforts of this activity are wide-ranging and documented completely on their web page and elsewhere [18, 22, 56, 185]. The aspects of this work most closely related to this dissertation are:

- The Resource Description Framework (RDF) [273] a data model for expressing triples, assertions of typed relationships between named subjects and either literals (e.g., strings, numeric values) or named objects.
- The RDF Vocabulary Description Language (RDFs) [91] a mechanism for using RDF to define vocabularies (entities and relationships) for use in other RDF descriptions.
- The OWL Web Ontology Language [43] an RDF-based language for publishing and sharing ontologies on the World Wide Web.

In addition to these core Semantic Web technologies, there are two additional outputs of the Semantic Web Activity of special relevance to the OAI-ORE work described in Chapter 12:

- Named Graphs [110, 111] are a mechanism for instantiating a set of RDF triples (a connected or unconnected sub-graph) as a first class Resource, with a URI.
- POWDER (Protocol for Web Description Resources) [23] is the focus of a W3C Working Group⁵⁸ charged with developing a standard for associating structured metadata with groups of web Resources.

⁵⁷ <http://www.w3.org/2001/sw/>

⁵⁸ <http://www.w3.org/2007/powder/>

Finally, there is considerable interest in leveraging the entire semantic web technology stack as a substrate for building digital libraries and knowledge domains [278]. One notable example in this area is the JeromeDL [277] a self-declared “Social Semantic Digital Library” that allows rich bibliographic description of library content and social activities (e.g., bookmarking, semantic annotations, knowledge sharing) over library content.

Historical overviews of digital library research

A number of other members of the digital library research community have written about the past and future of digital library research, albeit with a different focus than that used in this dissertation.

Lesk [333] and Arms [29] provide textbook-like overviews of the origins and contemporary (at the time of publication) state of the field. Both books were published before the emergence of Web 2.0 (although Arms issued a revised online edition in 2005 [27]), and therefore do not describe the effect of Web 2.0 on digital libraries.

Borgman, who has been cited numerous times throughout this publication, has reviewed and critiqued the process and products of digital library research in a number of publications and presentations. Her 1999 paper “What are digital libraries? Competing Visions” [82] contains an excellent overview of the different communities involved in the shaping of the field. She expands this somewhat in her 2000 book “From Gutenberg to the global information infrastructure: access to information in the networked world” [79], and in a more recent book [80] she places the digital library efforts in the context of the broader cyberinfrastructure initiatives. Her recent talk [77]

at the ACM Joint Conference on Digital Libraries covered subject matter closely related to the analysis of this dissertation⁵⁹.

In his 2004 paper [119] Y.T. Chien, one of the original program managers of digital library funding at the NSF recounted the successes of digital library research but argued for understanding of “disruptive technologies” that DL research needed to pay attention to if it was to remain relevant. These technologies included mobile communications, broadband, distributed storage and retrieval (now called “the cloud”), and E-payment. His analysis is similar to that presented here in that he warns of fundamental disruption of core assumptions of digital libraries in a changing information landscape and changing expectations and demands by users.

The July/August 2005 issue of D-Lib Magazine celebrating the 10th anniversary of that online publication [492], included a number of interesting analyses of digital libraries.

A paper by Stephen Griffin [219], the program manager at the NSF mainly responsible for funding both DLI-1 and DLI-2, enumerates the success of both initiatives but notes the problematic nature of short-term funding for an inherently long-term endeavor such as the curation of information. He argues for attention to content as part of future cyberinfrastructure efforts.

A paper by Clifford Lynch [353], whose work has been cited numerous times through this dissertation, argues that the era of digital libraries as a definable and useful research area may be over. He points out the importance of work in targeted areas, especially digital curation and preservation, with a particular focus on the products of eScience. Another theme of his argument, tangentially related to the material in this

⁵⁹ And, in fact, the talk was partially informed by personal communication with me.

dissertation, is the need for research at the “crossroads of technology and social science” to examine information issues quite outside the traditional library context. These include the issues of personal information management, the role of digital information in teaching and learning, and the integration of collaboration (social activity) and information. Regarding the final point, he states, “...at least some sectors of the digital library community have always found active work environments to be an uncomfortable fit with the rather passive tradition of libraries”. Notably, these “active work environments” are characteristic of Web 2.0.

A paper by Paepcke, et al. [398], with the intriguing name “Dewey Meets Turing”, covers territory that is quite similar in some aspects to the argument presented in this dissertation. In this paper they state:

The coalition between the computing and library communities had been anchored in a tacit understanding that even in the 'new' world there would be coherent collections that one would operate on to search, organize, and browse. The collections would include multiple media; they would be larger than current holdings; and access methods would change. But the scene would still include information consumers, producers, and collections. [398]

They continue to describe what they call “the cuckoo’s egg surprise”, the emergence of the Web, which disrupted these assumptions and

... not only blurred the distinction between consumers and producers of information, but it dispersed most items that in the aggregate should have been collections across the world and under diverse ownership. This change undermined the common ground that had brought the two disciplines [computer science and libraries] together. [398]

The authors assert in their conclusion that despite this “the core function of librarianship remains. The information must be organized, collated, and presented.”

While I support their assertion of a role for librarians (i.e., information experts) in the emerging information paradigm, their conclusion seems to validate that this role will

be quite outside the traditional library meme and therefore may be something quite different than librarianship.

Finally, a paper by Arms [26] in the same issue uses viewpoint analysis, a technique from software engineering, to critically analyze digital library research. In a manner similar in ways to the content of this dissertation, he asserts that too much of the development of and evaluation of digital libraries has been done from an organizational or institutional viewpoint – in particular from the library perspective, which evaluates success in terms of the organization’s prevailing values or memes. He argues for the work to assume a user perspective and states that digital libraries should be integrated into, rather than distinct from the “single unified Internet that we take for granted today.”

Impact of the Web 1.0 to Web 2.0 transition

I have asserted throughout this dissertation that the web has gone through a fundamental transformation known commonly as Web 2.0. Although there is an abundance of work investigating the nature, structure, and evolution of many Web 2.0 applications such as Wikipedia, Twitter, Flickr, and blog usage, there is surprisingly little scholarly work on the phenomenon as a whole and its broad impact.

A large proportion of the literature falls into the class of popular information technology or business literature. The first widespread use of the term was by Tim O’Reilly in [390]. Another O’Reilly published popular business-oriented text is by Nickull, et al., who enumerate the strategies that entrepreneurs should take in order to leverage Web 2.0 technologies for their businesses. A similar argument is presented by Li and Bernoff [343]. Another book by Tapscott and Williams [453] takes a similar approach, describing how the traditional rules of competition and information

secrecy are invalidated by the Web 2.0 notions of benefitting from the “wisdom of crowds.”

In the realm of more scholarly approaches, Beer and Burrows [45] examine the Web 2.0 phenomenon from a sociologist’s prospective. They point out three societal impacts of Web 2.0, the first and third of which are especially relevant to the issues of intermediation and control zone that are described in this dissertation: “the changing relations between the *production and consumption* of content; the mainstreaming of *private information posted to the public domain*; and ... the emergence of a *new rhetoric of 'democratization'*.” [45] (italics in original)”

This notion of democratization of information in the digital age and its affect on the political process is the focus of a growing area of scholarship. Two recent books that cover this area are by Winograd and Hais [493] and a set of essays by Bolter [72].

Another scholarly examination on the social effects of Web 2.0 and the *participatory culture* that it has enabled is a white paper “Confronting the Challenges of Participatory Culture: Media Education for the 21st Century” written for the MacArthur foundation by Jenkins and others [251]. Jenkins has also written two books on this subject [249, 250]. Although the white paper primarily focuses on education, it also describes the profound changes in the flow of information in modern culture because of the *interactivity* [249] of new media and information technology. According to Jenkins et al.: “Participatory culture is emerging as the culture absorbs and responds to the explosion of new media technologies that make it possible for average consumers to archive, annotate, appropriate, and recirculate media content in powerful new ways” [251]. The white paper argues that this participatory culture cuts across education, creative practices, community interactions, and politics and requires

a new way of training young people in their opportunities and responsibilities in this culture.

Katz's excellent volume of collected papers [258] is perhaps the best current source of expert thinking on the effect of the changing web on higher education institutions including libraries. The chapter by Katz himself [259] and another one by Wheeler [485] examine the new model of information sharing in the Web 2.0 meme and describe how university information technology and information (i.e. libraries) will need to make profound readjustments in the face of this phenomenon.

As described earlier, the library community has responded to Web 2.0 with the notion of "Library 2.0". This term was originally coined by Michael Casey, who maintains a blog dedicated to the idea [145] and was further popularized by Chad and Miller [113, 373]. Perhaps the best attempt to dissect the meaning of Library 2.0 is a paper by Maness [358]. He defines Library 2.0 as "the application of interactive, collaborative, and multi-media web-based technologies to web-based library services and collections." The major impact of this, in his opinion is that:

Library 2.0 is completely user-centered and user-driven. It is a mashup of traditional library services and innovative Web 2.0 services. It is a library for the 21st century, rich in content, interactivity, and social activity.

I do not criticize the motivation or underlying premises of Library 2.0. I share with many others the hope that libraries and their collections and librarians as information specialists will be integrated into the Web 2.0 framework. I do find it ironic, however, that much of the Library 2.0 concept seems to imply the disappearance of the library as a recognizable and distinct meme, and projects a reality in which it and its assets are just other participants in a world of "crowd sourcing". This may be good, but seems to have little to do with the traditional mission of the library as a "first choice" for information and scholarly activities.

Finally, in a paper in First Monday, Cormode and Krishnamurthy [138] describe the effects of Web 2.0 on network and server loads. They do not cover any social or institutional effects.

Digital libraries as sociotechnical systems

Chapter 4 of this dissertation focuses on some of the techniques and theories of Science and Technology Studies (STS) to examine the historical context of libraries and digital libraries as information infrastructures. I asserted earlier that the digital library research community by and large viewed DLs from a technical perspective, assuming that technical changes would take place without disruption to established institutional models and roles. Because of this, they failed to anticipate the “genie in the bottle” that would emerge as networked information technologies, especially those manifested in Web 2.0, were adopted and adapted by all levels of society. This “genie” has revealed a number of profound changes in the social and personal relationships to information, knowledge, and the institutions interweaved with them.

This argument falls into a class of critical analyses, rooted in STS, that view infrastructures and technologies as *sociotechnical* systems – “networks of technology, information, documents, people, and practices” [469]. Van House’s excellent review article [468] provides a thorough overview of this type of analysis covering theories including Actor-Network Theory, Social Construction of Technology, Symbolic Interactionism, Epistemic Cultures, and Activity Systems. The interested reader is referred to that review. This short section will focus on the application of this approach to libraries and information systems.

David Levy, formerly at Xerox PARC now at the University of Washington, has written several papers and articles that take a broad perspective of libraries in general and digital libraries in particular. In a 1995 paper [341] with Cathy Marshall, another

colleague at PARC, he looks at libraries through the perspectives of documents, technologies, and work practices. He argues that Digital Library research is at times too narrow-minded and states that “library developments ought to be grounded in a solid understanding of past and present practices. Without this, we risk losing still relevant structures and practices while maintaining an allegiance to mythical and irrelevant features of an unrealized past or an idealized present.”

In a later paper from 2000 [337] Levy argues for the importance of the library as an institution that has “come to symbolize and to exemplify the values we impute to the entire [information] circuit” and states that the development of digital libraries must account for this and not assume that the simple availability of information in digital form is de facto good or better.

Finally, in a 2003 chapter [338] he focuses on the nature of the document as a fundamental constituent of the library and states that any work in the area of digital libraries must account for the historical continuity of the document paradigm in its movement across technologies.

Levy’s colleague Cathy Marshall, now at Microsoft, has also written about the need to recognize both the notions of continuity and change in the transformation of libraries from physical to the digital [361, 362]. Her 2003 article is particularly relevant to the discussion in this dissertation about “control zones” because it describes the nature of boundaries that exist in libraries whether they are surrounded by walls are absent of walls on the Internet. She makes some astute comments about the information loss that occurs when boundaries are broken down, stating that “document disaggregation... has an effect on a user’s ability to interpret the content. A document is more than the sum of its parts.”

There are a number of instances of the use of activity theory as a vehicle for examining and especially evaluating digital libraries. All of the applications of this methodology cite the advantages of understanding the interactions among tools, users, objectives, and community rule constraints for full understanding of complex information systems.

In work described in both a chapter [442] and journal article [441] Spasser uses activity theory to evaluate the effectiveness of the Flora of North America Digital Library Project. He argues for the advantages of this technique because of the manner in which it reveals organizational issues and contradictions in the construction and use of digital library technology. Collins, et al. [132], Blacker, et al. [67, 68], and Boer, et al. [70] also employ activity theory as a tool for evaluation of digital libraries and information environments.

Another theory from STS, actor-network theory, has also been employed as an evaluative mechanism for understanding the effect of digital libraries. In a book chapter on the subject [467], Van House claims that actor-network theory is a useful evaluation tool because of the way it accounts for the complexity of digital libraries as a tool, a boundary object, a locus of multiple translations, and an “active participant in the creation and circulation of documents, images, and other kinds of inscriptions.”

The Social Construction of Technology (SCOT) is another approach for looking beyond technology and understanding its cultural and social context. As described by Van House [468]:

Different groups have different arrays of problems; each problem has an array of possible solutions. The SCOT descriptive model proceeds with the “sociological deconstruction” of the object of interest, showing the different meanings the artifact has for different groups, focusing on the problems and associated solutions that each group sees with respect to the artifact. SCOT contends that a technological artifact possesses

“interpretive flexibility,” revealed through the different meanings attributed to it by the different relevant social groups.

This notion of “interpretive flexibility” is a useful framework for understanding the historical overview presented in Chapter 2: that is, the differing manners in which the involved communities (funders, computing and information scientists, librarians) interpreted the application of network technologies to the library context, and, in fact, the vastly divergent manner in which the DL and web communities applied the same online technologies.

Other applications of SCOT in the context of digital libraries include Kilker and Gay [263], who used SCOT as a framework for evaluating the “Making of America” project, and for understanding users’ perceptions of the performance of digital library technologies.

O’Day and Nardi [388] introduce another holistic mechanism, information ecology, as a mechanism for evaluating the design of digital libraries. They state that “we believe that looking at the broad picture is more important than focusing only on the details of particular technology innovations. Even when a new technology is meant to serve a general purpose, exposure to the richness of users environments’ is a viable resource for design input and creativity.”

The work of Kling, whose name and research are virtually synonymous with *social informatics*, deserves mention in this description of related work. In [270, 271] Kling defines social informatics as the “interdisciplinary study of the design, uses, and consequences of information technologies that take into account their interaction with institutional and cultural contexts”. Kling applies this contextualizing framework to libraries and in particular to digital libraries in [269, 271]. In 1996, Bishop and Star [61] wrote about the utility of social informatics as a tool for the design and evaluation of digital libraries. Agre in [17] employs social informatics and states that “Every

technology is embedded in the social world in complicated ways, and this is particularly true for digital libraries, which are intertwined with the cognitive processes of a complex society. Unless our conceptualization of society stands on an equal footing with our conceptualization of the technology it uses, our analysis will inevitably be overwhelmed by myths”

Finally, Fuchs, et al. [199, 200, 412], in works cited earlier, describe the compound social effects of the web and in particular in its “version” transition. They ground their work as an outgrowth of social informatics, but take a notably Marxist perspective, describing the effect of web technologies and the democratization of information creation and distribution on the structure of capitalist society.

In conclusion, while there exists a body of work examining digital libraries as sociotechnical systems, none of that work examines the nature of those systems and their library-based information model in the context of the increasingly dominant sociotechnical information system, the web.

Chapter 6

Introduction to Chapters 7-12

The second part of this dissertation, consisting of the next six chapters, describes selected results of my research work over the past fifteen years. These results are mainly technical and cover various areas of digital library interoperability.

Collectively, the work described in these chapters demonstrates the influence of the library meme on digital library technology, especially on my early work, and the movement away from the constraints of that meme over the years. Figure 23 illustrates the chronology of these research projects and their association with digital library research areas. The grey bars in the figure indicate research projects displayed on an annual time scale, which is shown on the bottom. The number(s) displayed in each grey bar is the chapter(s) in which the respective project is described. Finally, the multi-colored, bordered rectangles that are connected by lines to the project bars indicate the correspondence of the project to one or more digital library research areas.

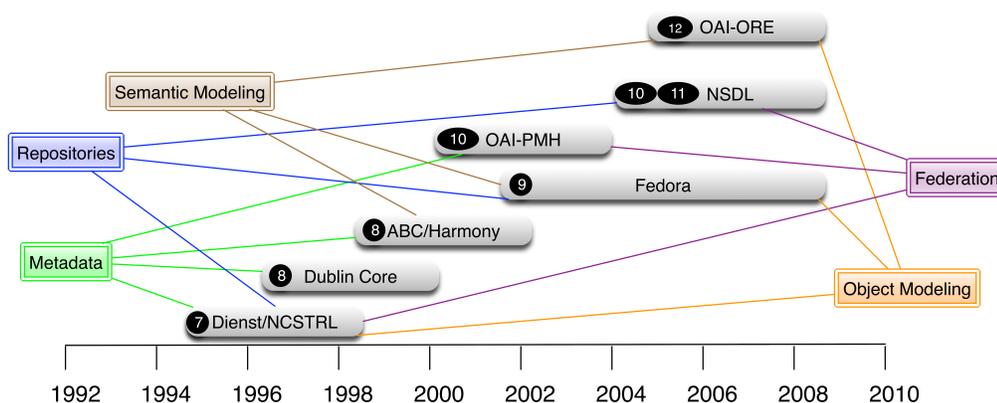


Figure 23 - Research project timeline

A summary of these research projects and their connection with digital library architecture and other concepts described in this dissertation is as follows:

- *Dienst/NCSTRL* (Chapter 7) – the Distributed Interactive Extensible Networked System for Technical Reports and the Networked Computer Science Technical Research Library are a digital library architecture and implementation that exemplifies many of the technical features typical of the early work in this area that was strongly influenced by the library meme. It includes notions of digital objects contained in institutional repositories, a digital library-specific protocol for access to those objects, structured metadata, federated search, and portals for search and access to the contents of the library.
- *Dublin Core* (Chapter 8, Chapter 10) – is the most visible and widely disseminated metadata vocabulary for digital information. The roots of the Dublin Core Metadata Initiative lay in the belief that effective discovery of web information could only occur if “simple” metadata records, authored by everyday content creators, were attached to digital content. Perhaps more than any other aspect of digital library research work, the focus on Dublin Core and on bibliographic metadata in general demonstrates the impact of the library legacy on digital library work.
- *ABC/Harmony* (Chapter 8) – is a metadata vocabulary and ontology that incorporates semantic web concepts and moves beyond the constraints of the simple Dublin Core model. Although it still demonstrates the influence of metadata-centric thinking on digital library research, its emphasis on events as a central ontological feature demonstrates an increased awareness of the fundamental difference between the nature of digital information

(e.g., its dynamic nature, its complex provenance, etc.) and the physical information resources in traditional libraries.

- *Fedora* (Chapter 9) – the Flexible Extensible Digital Object Repository is a widely deployed system that combines digital library content management with semantic web concepts and web services. In this manner, it effectively bridges between the traditional library realm and many of the concepts that underlie Web 2.0.
- *OAI-PMH* (Chapter 10) – the Open Archives Initiative-Protocol for Metadata Harvesting is a widely deployed infrastructure for harvesting structured metadata from digital library and institutional repositories. Experience with OAI-PMH has demonstrated some of the serious shortcomings in the assumptions that underlie metadata-centric thinking in the context of the web information environment.
- *NSDL* (Chapter 10, Chapter 11) – the National Science Digital Library is a multiyear, NSF-funded project that has produced a digital library implementation built on the work in Dublin Core and OAI-PMH. Recently, in recognition of the need to move beyond the rather limited “search and access” principles of traditional digital libraries, work on NSDL has been focused on Web 2.0 principles such as exploiting the knowledge of users and incorporating context with content.
- *OAI-ORE* (Chapter 12) – the Open Archives Initiative - Object Reuse and Exchange project defines semantically-based data models and encoding formats for the identification and description of compound objects (i.e. aggregates) on the web. It demonstrates awareness that the representation of information objects in digital libraries, eScience, and cyberinfrastructure

must be fully integrated into the web architecture and compatible with Web 2.0 applications.

Each chapter is based on a published, reviewed paper contemporary to the time of the respective research project. The citation to the paper is included at the beginning of the chapter.

A newly written preface precedes the paper text in each chapter. This preface explains the context of the paper within the overall theme of this dissertation - the evolution of digital library research over its 16-year history, its basis in and divergence from the library meme, and the influence of the evolving web context. In addition, the preface includes an acknowledgements section updated from the original paper acknowledgements, indicating collaborations in and funding for the work therein. A horizontal line separates the present-day preface and acknowledgments text from the original paper content.

The content of each paper is by-and-large identical to its original published form, thereby preserving its historical integrity and demonstrating perspectives on digital library and web issues contemporary to the time of publication. Therefore, the text should be read with a temporal frame of reference corresponding to the publication date of the paper (as stated in the citation). For example, a phrase such as “We are currently exploring” in a 2001 paper indicates an activity in my research at that publication date.

I have made some modifications to the original papers. Phrases with relative time spans such as “for the past three years” have been replaced with the exact years. References to planned future work have mostly been omitted. Finally, most of the illustrations have been updated for consistency and to bring them up to present-day quality standards.

Making Global Digital Libraries Work

Preface

This chapter is based on:

Lagoze, C., Fielding, D. and Payette, S., Making Global Digital Libraries Work: Collection Service, Connectivity Regions, and Collection Views, in *ACM Digital Libraries '98*, (Pittsburgh, 1998) [291].

This chapter describes work on the Distributed Interactive Networked System For Technical Reports (Dienst), which began in 1992 under the auspices of the DARPA-funded Computer Science Technical Reports Project (CSTR)⁶⁰. This project was a collaboration between the “five leading US computer science departments” – Berkeley, Cornell, Carnegie Mellon, MIT, and Stanford – and was administered by the Corporation for Network Research Initiatives (CNRI). The goal of this early digital library project was to develop the technology and understand the intellectual property issues for putting computer science technical reports online.

As described in Chapter 5 the results of the work had a substantial impact on later digital library architectures and implementations. Dienst was at one point proposed as the basis for interoperability among all projects funded in the Digital Libraries Initiative [330]. In addition, it was widely deployed worldwide in the Networked Computer Science Technical Research Library (NCSTRL) [153], which at its height included over 160 institutions. The work also led to a collaboration with the ePrint

⁶⁰ <http://www.cnri.reston.va.us/cstr.html>

arXiv⁶¹, then at Los Alamos now at Cornell, to develop the Computing Research Repository (CORR) and incorporate that into NCSTRL [226].

The Dienst architecture and protocol was notable for a variety of reasons. The development of Dienst occurred shortly after the introduction of the Mosaic browser, commonly acknowledged as a main factor that caused the initial popularization of the web. Before the existence of web technologies and the notion of a common browser client, client/server systems had to confront the difficult task of creating cross-platform clients and architecting client/server interactions at a lower level in the Internet stack (e.g., at the socket level). Dienst was the first digital library architecture to fully leverage these newly introduced web technologies, including URLs, HTML, HTTP, and the increasingly ubiquitous web browser.

In addition to this innovative use of web technology, Dienst incorporated a number of interesting concepts that had an important effect on later digital library developments. These include a digital object model that included the notion of compound, multiple-data stream resources and protocol-based accessibility to this compound objects; simple bibliographic metadata as an alternative to heavyweight library cataloging records; federated search, which is a special focus of this particular paper; and an extensible distributed service model. All of these technical developments are described in greater detail in the body of this chapter.

The chapter, based on a 1998 paper, reveals a number of the assumptions about digital libraries and the web, which were common in that early time, as described in the first part of this dissertation. It emphatically makes the point of distinguishing the web from a digital library, stating “a distinguishing aspect of a library (digital or otherwise)

⁶¹ <http://arxiv.org>

is management of collections. Given this understanding of a library the World Wide web is NOT a digital library. It represents a set of objects joined together technically (by the common protocol HTTP), but not by any collection management actions.”

Dienst can be characterized as a “classic” digital library architecture, revealing many artifacts of the traditional library meme. Search is based on catalog records, formatted with a simple metadata vocabulary [322], a precursor to the later Dublin Core vocabulary [482]. The entry point to search is through a portal, a “user interface server”, and mechanisms to expose the collection to (then primitive) web search technologies is not considered. Digital objects are organized into collections, and stored in and accessed from distributed repositories.

The focus of this paper, federated or distributed indexing and searching, reflects a particular artifact of the time and mindset of the digital library community, and is particularly demonstrative of these early assumptions, which now seem ill founded.

The following quote from the paper represents the core of these assumptions:

In the present World Wide Web, virtually all tools for resource discovery are based on a centralized model. Typically, a central service creates and deploys a master index, and sometimes creates one or more replicas of the index. Although this model is currently prevalent, we argue that distributed searching will become increasingly necessary to overcome the constraints inherent in the centralized model.

The paper goes on to justify this statement on the grounds of scalability, customizability, and intellectual property. Written before the explosive growth of Google and other crawler-based search indexing, these statements clearly failed to anticipate the complications of distributed searching both in terms of rank merging and reliability and the real scalability of centralized indexing. We described some of the performance issues with distributed searching in other papers [170, 171].

Another relevant concept touched on in the paper is the problematic and ambiguous notion of the control zone [34] in the digital library. Quoting from the body of the paper:

... in the traditional library model, some librarians argue that physical containment of objects (e.g., in stacks) is the primary criterion for inclusion in the collection. This notion of physical control breaks down in the networked environment of digital libraries where both overt and implicit linkages can be made between objects that reside in different physical locations.”

Note that the response to this problem in Dienst was to define the control zone in terms of the digital library catalog: “An object is ‘in’ a digital library's collection if it can be directly discovered using the resource discovery tools defined and implemented by the respective digital library”. Given the present reality where virtually all resource discovery takes place in mainstream search engines, rather than digital library catalogs, this definition certainly seems anachronistic.

Acknowledgments

The work described in this chapter was funded by the Defense Advanced Research Project Agency under Grant No. MDA 972-96-1-006 with the Corporation for National Research Initiatives. The Dienst architecture was originally formulated in 1992 by James R. Davis, then at the Xerox Design Research Institute at Cornell. He continued to make substantial contributions to this work through 2000. Other contributors to this work, all members of the Cornell Digital Library Research Group, were David Fielding, who collaborated on the design of the collection service and wrote the initial implementation of it, Naomi Dushay, and Sandy Payette, both of whom participated in the collection service design and worked on other parts of the Dienst architecture. Acknowledgments also go to Dean Krafft whose thoughtful observations inspired a good deal of the work at the Cornell Digital Library Research

Group. Finally, thanks to Bill arms at CNRI for his helpful feedback and support on this work.

Introduction

Since 1993 the Cornell Digital Library Research Group has been investigating the architecture of globally-distributed, federated digital libraries. In contrast to centralized or replicated stand-alone systems, these federated systems are composed of semi-autonomous services, distributed across the global Internet, that interoperate through an open protocol.

From the point of view of flexibility, extensibility, and scalability this federated model is preferable to self-contained, centralized systems (such as the current generation of library management systems that form the technical basis of modern libraries). Among the benefits of the federated model are:

- Stakeholders can maintain control of digital objects (documents) in their own repositories.
- Customized collections can be created by aggregating digital objects in these distributed repositories.
- New value-added services can be created as the need arises.
- The functionality of existing services can be enhanced in a modular fashion.
- Services can be replicated to enhance global accessibility.
- Customized user interfaces (digital library gateways) can be created to provide community-tailored access to other distributed digital library services.

These advantages gained by modularity, interoperability, and distribution should not come at the cost of decreased usability or performance. As much as possible, users should be insulated from the physical distribution of the system and should be able to

view the digital library as a single collection with uniform tools for search, retrieval, and display of information within. At the same time, the performance of the system should match user expectations. This "illusion of uniformity" should be maintained, whenever possible, in the face of poor and inconsistent network connectivity, variability in server load, inconsistent server administration, and other problems characteristic of distributed, decentralized systems.

Maintaining usability in the presence of such distribution is one of the key challenges for designing digital library architecture. Some of the aspects of this challenge have been extensively covered in the distributed systems literature [59]. However, issues of global scale and a high degree of component autonomy change the flavor of the digital library problem sufficiently to call for some new solutions.

In this paper we examine one aspect of the distributed digital library problem - distributed searching. In the present World Wide Web, virtually all tools for resource discovery are based on a centralized model. Typically, a central service creates and deploys a master index, and sometimes creates one or more replicas of the index. Although this model is currently prevalent, we argue that distributed searching will become increasingly necessary to overcome the constraints inherent in the centralized model. In particular, effective architectures for distributed searching must be developed to address:

- *Issues of Scalability* - As the global information space explodes, it has become increasingly difficult to collect indexing information and keep centralized indexes up-to-date. Commercial web search providers are beginning to recognize this fact and it has even lead to a recent commercial patent for distributed searching technology [244].

- *Issues of Specificity* - While interoperability among search or indexing sites is important, it is also vital that the information infrastructure accommodates the unique needs of specific communities. Such accommodation is best accomplished via separate service providers that can both cater to individual community needs (through custom metadata, specialized data formats, query languages, user interfaces, etc.) and interoperate on a global scale through open protocols.
- *Issues of Intellectual Property* - The current resource discovery infrastructure depends on the fact that almost all the items in the global information space are not encumbered by access restrictions. Certainly this will change as improved technology for digital object rights management evolves [30, 446] and, as a result, more objects with more restricted access proliferate on the net. (In fact, one could argue that in the future the objects with the most value will be those that are not freely available.) In this case, it will become more difficult if not impossible for centralized search providers to collect indexing information by simply walking the global information space (in the fashion of current "web spiders"). As a result, resource discovery will depend on distributed indexing sites that are physically, logically, or legally linked (through licensing agreements) with sites of content providers.

Other researchers have investigated a variety of issues relevant to distributed searching. The distributed database community has a long history of investigating the optimal distribution of indexing information across LANs and controlled WANs [125]. Researchers in the digital library community have examined query translation issues [116], content summarization for query routing [216], and protocols for meta-searching and metadata collection [215].

This paper describes an architecture, and experience with that architecture, for distributing index servers⁶² on a global scale and disseminating meta-information on the location of those servers among participating servers. The architecture has three logical components. The first is a distributed collection service that identifies the index servers of a distributed digital library collection and manages meta-information about those servers. The second is a connectivity region, which is a set of nodes on the Internet with relatively good network connectivity (e.g., low latencies, infrequent partitioning). The last is a collection view, which is a perspective on a collection specific to a connectivity region.

The architecture described here was developed out of our experiences at Cornell building a globally distributed digital library, NCSTRL⁶³ (Networked Computer Science Technical Reports Library). The structure of this paper reflects the development path of the Cornell work. First, we briefly summarize NCSTRL, our global test bed, and Dienst, the technology on which that test bed is based. This section includes a description of the initial Dienst collection service, which forms the basis for the expanded collection service described later in the paper. We then describe our early efforts, or mistakes (depending on your perspective), to deal with distributed resource discovery in NCSTRL. Following this we describe the current evolution of our distributed searching architecture, with an explanation of connectivity regions and how they are implemented using an enhanced collection service. We conclude by describing some future work and opportunities for research.

⁶² Throughout this paper an *index server* is a server that collects meta-information about objects in a digital library collection and returns results (hit lists) in response to queries on that meta-information.

⁶³ Pronounced "ancestral".

NCSTRL – The test bed for a globally distributed digital library

The global digital library architecture described in this paper is the result of our work with Dienst [152], a protocol and a reference implementation for distributed digital object libraries. The initial Dienst system was designed and developed as part of the DARPA-sponsored CS-TR project⁶⁴, which investigated general digital library issues and, in particular, the technology for making technical reports digitally available from the participating institutions⁶⁵.

At the conclusion of CS-TR funding, participants in WATERS [356], one of several other efforts to create a digital library of computer science technical reports, joined with developers of Dienst and other members of CS-TR to form NCSTRL. By June 1998, NCSTRL had grown to include collections from over 100 institutions with over 60 servers worldwide. The globally distributed nature of NCSTRL and the federated, open architecture of Dienst on which it is based, represents a unique test bed for ongoing digital library experiments. Those experiments are both of a technical nature, such as those described in this paper, and of a social nature, exploring the organizational aspects of loosely federated information systems.

Dienst architecture

The remainder of this section summarizes the Dienst architecture that underlies NCSTRL. A more detailed description of Dienst can be found in the Implementation Reference Manual [302]. The fundamental features of the architecture are a logical document model, distributed digital library services, and an open protocol for interoperation among those services.

⁶⁴ <http://www.cnri.reston.va.us/cstr.html>

⁶⁵ U.C. Berkeley, Carnegie-Mellon, Cornell, M.I.T., and Stanford.

Logical Document Model

At the core of the Dienst architecture is the notion of a document, a logical abstraction⁶⁶ that incorporates a number of concepts.

- Each Document has a *globally unique name* that is defined using the handle service⁶⁷.
- A document consists of a number of *components*. The two components currently in use in NCSTRL are the bibliographic description and the "body" of the document.
- Each component is available in one or more *formats*. For example, the body of the document may be available in PostScript, HTML, and as a sequence of TIFF images.
- A component in a format may be divided into a number of *decompositions*. For example, the "body" available in PostScript format may be divided into "pages" or "chapters".

Digital Library Services

The functionality of the Dienst architecture is logically divided among a set of distinct services. Although the services are modular in nature, they are currently implemented as a single physical server. This grouping of services was merely a matter of expediency, and our current research and development efforts are motivated by our belief that the digital library service structure should be physically, as well as logically, modular.

⁶⁶ By *logical* we mean that the abstraction is distinct from the "physical" one-to-one mapping of document to file that exists in file systems (local or distributed), FTP, or HTTP (sans CGI).

⁶⁷ <http://www.handle.net/>

There are three core Dienst services, in addition to the collection management service that we describe in the next section. The core services are:

- The *Repository Service* that stores and provides access to documents identified using the global naming service and structured according to the document model described earlier,
- The *Index Service* that stores indexing (meta) information about documents in the collection and responds to queries on this indexed information, and
- The *User Interface Service* that provides a human front-end to the other services.

Open Protocol.

Dienst services and servers interoperate using a well-defined protocol [151]. The structure of this protocol corresponds to the logical services described above. Each protocol request is framed as a verb to a service.

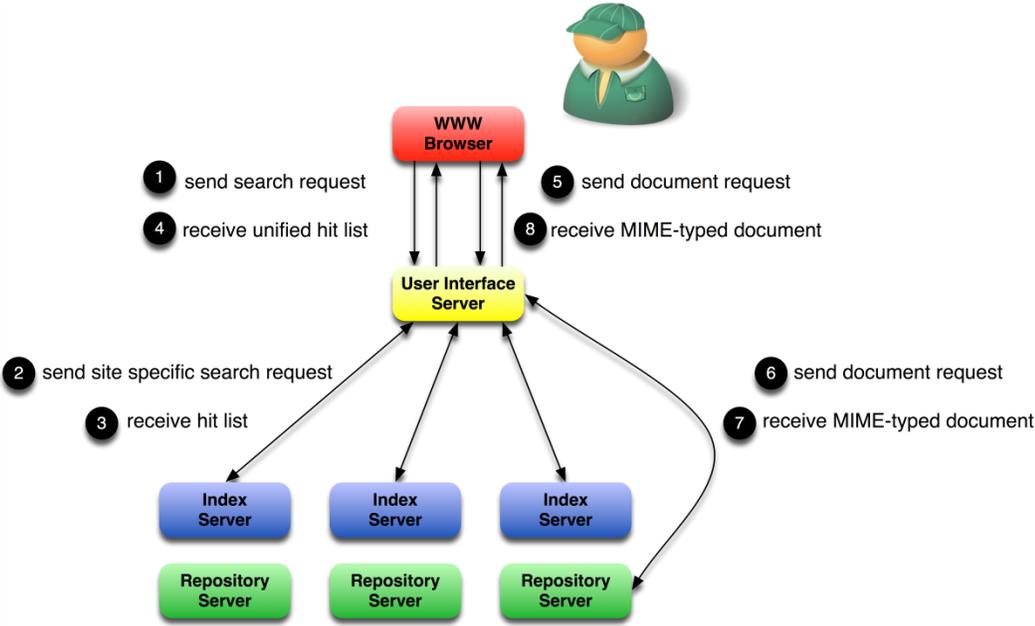


Figure 24 - Dienst Services

Figure 24 illustrates the protocol-based interactions between the core Dienst services for search and retrieval of a document. The user interface service acts as the mediator between the user's browser and the multiple index and repository services in a distributed collection. Requests are made through verbs addressed to one of the three core services. For example:

- The *Search* verb of the index service returns a citation list of documents meeting specified search criteria in the respective index
- The *Fetch* verb of the repository service returns a *dissemination* of a specified document identified by its unique identifier (or handle). Arguments to the *Fetch* verb conform to the concepts in the Dienst document model. For example, a *Fetch* request may specify that page one of the document body should be returned in GIF format.

The use of an open protocol has two key advantages. First, it permits the construction of other value-added services that interact with existing Dienst servers. Second, it allows individual services and, in fact, the entire Dienst implementation to be replaced as other alternative implementations are developed.

The open protocol is a defining feature of the collection service we describe next. This service is critical to the management of a distributed digital library collection in the Dienst architecture.

The Collection Service: Defining the Contents of the Digital Library

A distinguishing aspect of a library (digital or otherwise) is management of collections. Management of the collection begins with selection of the objects to be included in the collection. Objects are selected from a global information space (e.g., the set of all published books, or the set of all objects on the Internet), and become constituents of library collections based on criteria applied by selectors or collection

managers. Depending on the sophistication of the library, there may be other collection management functions such as preservation, archiving, and the like. Given this understanding of a library, the World Wide Web, by itself, is NOT a digital library. It represents a set of objects joined together technically (by the common protocol HTTP), but not by any collection management actions. Similarly, sets of documents residing on servers communicating via the Dienst protocol do not comprise a digital library.

Thus, digital libraries cannot be defined by the mere existence or application of enabling technologies. Digital libraries are distinguished from the more ubiquitous networked information landscape through their incorporation of collection management services, which may involve human intervention.

Even with collection management, the definition of what is actually “contained” in a digital library can become ambiguous. For instance, in the traditional library model, some librarians argue that physical containment of objects (e.g., in stacks) is the primary criterion for inclusion in the collection. This notion of physical control breaks down in the networked environment of digital libraries where both overt and implicit linkages can be made between objects that reside in different physical locations. For example, if object A is included in a collection, are objects B, C, and D that are linked to object A also included in the collection? If so, are all objects transitively linked to object A via other objects also included? The answer to these questions has important implications in the areas such as legal responsibility and public service.

While there are, undoubtedly, multiple perspectives on the definition of digital library collections, in this paper we will adopt the following working definition. An object is

"in" a digital library's collection if it can be directly discovered using the resource discovery tools defined and implemented by the respective digital library⁶⁸.

We emphasize the distinction between discovery and retrieval in this definition. First, discovery of the object may mean that a surrogate of the actual object may be indexed, and the actual object (which may be a physical artifact) must be retrieved through other means. Second, assuming a global name space, any object in the global information space may be retrievable (using its URN) without necessarily being in the library from which it is being fetched. One can think of this type of retrieval as a type of digital "inter-library loan".

Another interesting aspect of collection building is the level at which "inclusion" is evaluated. At the lowest level, individual digital objects are aggregated to form a (sub)-collection. At a higher level, multiple (sub)-collections of items are federated to form larger collections.

NCSTRL is a working example of multiple levels of collection management. The Dienst architecture provides for institutional autonomy in item-level collection building, and the capability for institutions to federate into the larger NCSTRL collection. The Dienst collection service is the mechanism for managing the federation level of collection definition. The data for managing the collection is obtained via protocol requests to this service that return the following information:

- *The list of organizations that are part of the collection.* In NCSTRL the granularity of an organization corresponds to the computer science departments and research institutions that are members of NCSTRL (*e.g.*,

⁶⁸ This brings up the interesting question whether the set of objects discoverable through one of the web search services is part of the digital library defined by that service. The nature of digital libraries and their collections provokes many interesting questions.

Cornell Computer Science Department, Georgia Institute of Technology
College of Computing).

- *The network location.* The service provides the address and port of the Dienst index servers that store indexing information for each organization. For example, indexing information for Cornell Computer Science may be stored at foo.ncstrl.org port 80 and bar.ncstrl.org port 8083.
- *Meta-information about each of the index servers.* At present this meta-information indicates whether the index server should be considered primary or secondary. However, our intention is to expand this meta-information to include data about last update of the index, performance information, content summaries, and the like.

From the administrative perspective, the collection service allows easy management of the NCSTRL collection. Organizations join NCSTRL by submitting an application to our collection librarian⁶⁹ via the Web. Following our confirmation that the organization conforms to the collection profile (the institution should be a Ph.D. granting institution in computer science) and has a working Dienst-protocol-conformant server, the NCSTRL administrator at Cornell adds the institutional information to the collection service tables. This new institution then becomes visible to each NCSTRL user interface server after its next collection service request.

We originally implemented the collection service on a single Dienst server. In this configuration, the address and port number of the collection server is stored in the configuration file of each Dienst server. Periodically (every hour) each Dienst server issues a collection service protocol request to obtain the collection information, which

⁶⁹ Rebecca Wesley at Stanford University.

we described earlier. The requesting Dienst server then stores the collection information internally in a table. At this point the user interface services have access to the current list of participating organizations (provided by the collection server). Figure 25 illustrates the interaction between the collection service, user interface servers, and index servers in Dienst. As shown, each user interface server queries the collection server for collection information. For a specific query, an individual user interface (labeled UI₁ in the figure) uses this collection information to determine which index servers should process the query.

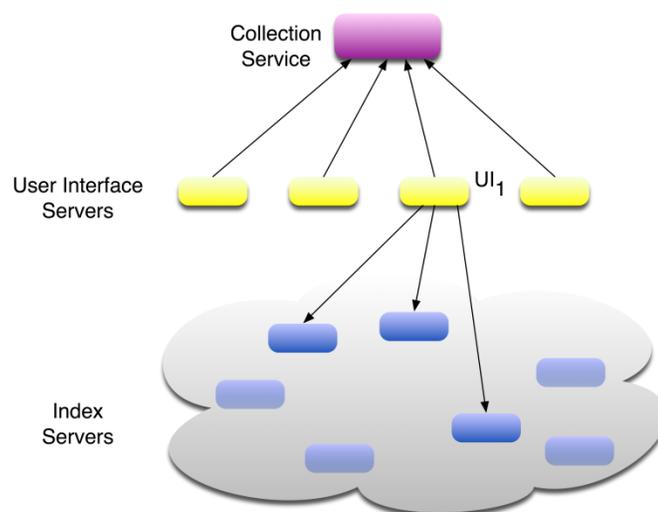


Figure 25 - Dienst service interactions.

From a user perspective, the latest organizations appear on the search form provided by the user interface service. When composing queries to the NCSTRL collection, users choose which organizations should be included in the search results. The respective user interface service can determine where queries should be dispatched by using the network location and contents data provided by the collection service. Once this information is obtained, the user interface service submits the actual query to the target index servers using a Dienst index server protocol request. When responses are

returned from the target index servers, the user interface service merges the responses into a single result set.

The next section of this paper describes the use of the collection server to implement two initial distributed search topologies in NCSTRL. The section that follows then describes a distributed version of the collection service based on connectivity regions, enabling globally distributed search.

The evolution of a distributed digital library: early experience

The flexibility of the collection service and its interaction with the user interface services allowed us to rapidly expand the NCSTRL collection from five sites in 1995 to over 100 sites in 1998. In the course of this rapid expansion, we implemented two initial distributed searching topologies: simple distributed searching and distributed searching with backup. In this section we briefly describe those topologies, and the lessons learned from deploying them in NCSTRL.

Some of the architectural solutions described in this section may, in hindsight, seem rather naïve and the results predictable. While that may be true, these solutions were developed and retrofitted onto a rapidly growing production distributed system. In addition, the experience gained from this incremental approach proved valuable and helped contribute to the architectural solutions described later in this paper. Finally, some have argued that given the present scale of the NCSTRL collection, centralized replicated searching is the more preferable and predictable model⁷⁰. This may also be true. However, as we argued earlier in this paper, the distributed searching problem will have to be investigated for future digital library infrastructure to operate, and NCSTRL has been, and still is, a unique test bed for researching those issues.

⁷⁰ Ed Fox, personal communication.

Simple Distributed Searching

The five institutions that participated in the CS-TR project and that participated in the initial Dienst-based collection shared two characteristics having implications for distributed searching reliability:

1. **Connectivity.** Among the five institutions connectivity was good; network down-time was minimal and latencies were fairly low.
2. **Commitment.** Due to joint funding within the CS-TR project, these five institutions shared a common interest and commitment to the success of the test bed technical report collection. As a result, the five servers and their contained collections were well administered.⁷¹

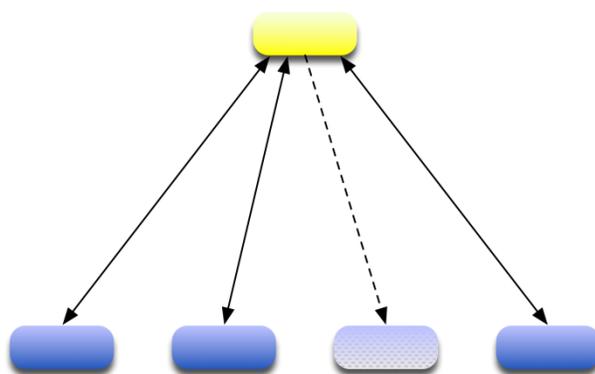


Figure 26 - Simple distributed search with server failure

Based on the high technical and administrative reliability, we made the initial decision to implement a simple distributed searching topology. In this topology only one index server existed for each organization in the collection. In fact, the index server was

⁷¹ We strongly emphasize that the factors that contribute to the success of a federated library are not restricted to the technical domain. We have found throughout the existence of NCSTRL that poor management of a few individual servers in a federated system can seriously degrade the reliability and integrity of the entire system. Poor management can take a variety of forms including a server that is periodically unavailable, descriptive metadata that is incomplete or incorrect, a collection that is not kept up to date, or any number of other factors.

resident in the same Dienst server as the document repository that it indexed. A search query from any of the user interface servers in the collection was, regardless of the origin of the search, dispatched to the same set of indexing servers. If an individual indexing server was unavailable (due to network failure or server failure) or overloaded (resulting in a time-out) the user was alerted that results could not be returned for the organization stored on that index server.

This simple topology is illustrated in Figure 26, with a connection failure to one of the index servers. The loss of access to information resulting from unavailable or slowly responding servers motivated the introduction of backup servers into the distributed searching scenario.

Distributed Searching with Backup

Even with a controlled set of servers, as was the case in the original CS-TR project, server failures occurred too often. As the size of the collection grew beyond the original five institutions, the number of failures increased dramatically. In fact, most search result sets were incomplete, showing one or more "unavailable organizations".

In response to this situation, we soon introduced replicated index servers, with a ranking of which server was primary, secondary, etc. This was done by extending the collection service protocol so that it could indicate the priority order of a specific index server for a specific organization. For example, the protocol response might indicate that `foo.ncstrl.org` port 80 is the primary index server for the Cornell CS collection, but `bar.ncstrl.org` port 8083 is the secondary index server for that same collection. Using this information, an individual user interface server could then first distribute the search request to the appropriate set of primary index servers. In case of failure or time-outs, the user interface could then distribute the same to the secondary

index servers corresponding to the "unavailable organizations" in the primary phase of the search.

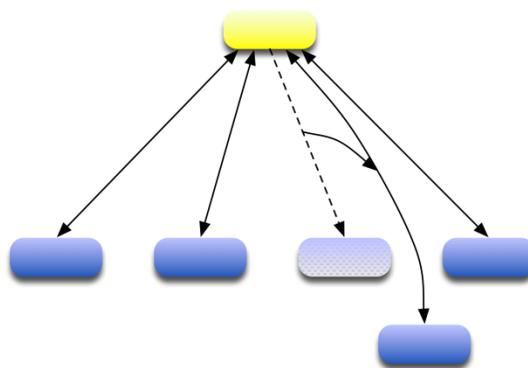


Figure 27 - Primary and secondary index servers

Figure 27 shows an example of this primary and secondary (backup) index server topology. In the illustration, one of the index servers has failed, and the query is redirected to the secondary index for that site.

Adaptive Routing between Primary and Secondary Index Servers

Experience with the backup index server topology demonstrated that in many cases poor performance or failure of an individual server persists over time. For example, a network or server failure is normally not repaired immediately. Rather than continuing to use a failing primary index server, it is preferable that the user interface server "remember" the failure of the respective server and, as a result, change the rank ordering of the index servers.

To implement such behavior we implemented a simple adaptive algorithm at each user interface server that keeps track of the success or failure history of each index server to which a queries are routed. If a specific index server repeatedly fails within a specified period and a secondary index server exists for the organizations indexed by that server, the unreliable server is "demoted" and the appropriate backup index

servers "promoted". This change in rank is left in place for a fixed period, after which the demotion and promotion are undone (but re-instated if the next retry results in another failure). In this manner, the overall response time to queries is relatively insulated from the effects of unreliable servers.

Connectivity regions and distributed collection service

The addition of international partners to NCSTRL, and the resulting global deployment of Dienst servers, required rethinking the ranked index server topology described in the previous section. As is well known, global connectivity varies dramatically. In fact, the latency times between nodes can differ by several orders of magnitude. In addition, the patterns of connectivity are not necessarily geographically related. Points that are coincident in physical space may be "distant" in network space, as measured by reliability and speed of the connection. This disparity between geographic and electronic "proximity" often corresponds to patterns of telecommunication development over the past fifty years, which often corresponded to political and colonial patterns. The exaggeration of this pattern is the fact the phone (and network) connections from a developing country to its former colonial power are in most cases better than to its neighbors (or, in fact, within its own country!).

We model the patterns of global connectivity through the notion of a connectivity region. A connectivity region is defined as a group of nodes on the network that among them have good connectivity⁷², relative to nodes outside of the region. At present, this definition is qualitative, but we plan to develop a more quantitative definition of the concept. The meaning and purpose of the connectivity abstraction is orthogonal to whether the region is statically or dynamically (adaptively) defined.

⁷² For the remainder of this paper we will define the quality of connectivity as a factor of both latency and reliability (resistance to failure).

The concept of connectivity regions allows us to reframe the requirements for distributed searching in the following fashion. In the absence of network or server failures, query routing from a specific user interface site should be restricted to those index servers in the same connectivity region. In case of a failure, an alternative indexing server should be chosen either in the same region or in another region with which there is good connectivity.

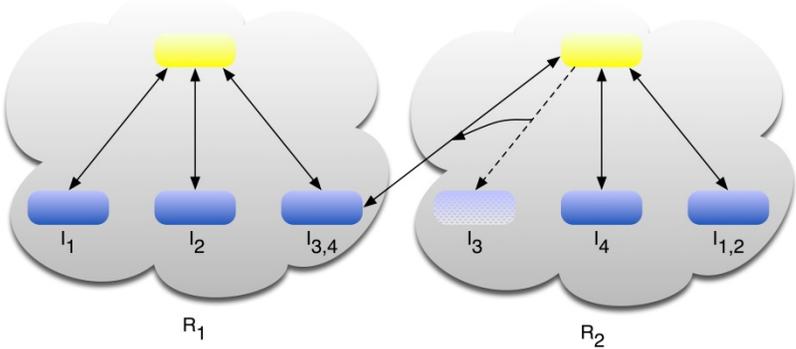


Figure 28 - Connectivity regions

Figure 28 illustrates a simple example of connectivity regions and the motivation behind them. In this figure there are two regions, labeled R_1 and R_2 . Each region contains one user interface server, which dispatches queries and combines responses, and three index servers, which respond to queries. In the example, the indexed data in the collection is divided into four partitions, and the subscript(s) on each of the indexing servers indicates the partition(s) indexed at the index server. For example, index server I_1 holds indexing information in partition 1 and index server $I_{3,4}$ holds indexing information in partitions 3 and 4. As illustrated, indexing information is replicated in a manner that queries can be routed within a region in which a user interface server is located. However, as also illustrated, a failure in an index server in a region (I_3 in R_2) may require routing of a query to an index server outside the region ($I_{3,4}$ in R_1).

A Distributed Collection Service and Collection Views

Earlier in this paper, we described how Dienst user interface services use data from the collection service to determine where to route queries. The routing is both content based - which index servers can answer queries for the organization(s) specified in the query - and priority based - which index server(s) should be considered primary, secondary, etc. for the specific organization(s). As described, the original collection service was implemented within one server. All Dienst user interface servers in the collection used that single server as the source for collection data. Furthermore, the collection data supplied to each Dienst user interface server was identical.

In contrast, the connectivity region concept, illustrated in Figure 28, implies that the routing decisions made by different user interface servers may be based on different collection information. In the example, the user interface server in region R_1 "believes" that the primary source for indexing information on partition 1 of the collection is at the index server labeled I_1 . On the other hand, the user interface server in region R_2 "believes" that the primary source for that information is at the index server labeled $I_{1,2}$. In other words the collection view, the meta-information about the contents of the collection, of the R_1 user interface differs from that of the R_2 user interface. A single collection, such as NCSTRL, may have multiple collection views, corresponding to the connectivity regions that have been defined for the servers in that collection.

In order to support the notion of multiple collection views, we re-implemented the Dienst collection service in a distributed manner. In this new implementation, the distributed collection service was divided into two logical server types.

- 1) *Central Collection Server (CCS)*. There is a single central collection server that serves as the central point of management of the collection. This server stores the following information:
 - (1) A table defining all organizations in the collection, in the same manner as the original collection service implementation.
 - (2) The list of Dienst servers (identified by host and port) that are acting as regional collection servers. There is one regional collection server per connectivity region.
 - (3) A set of collection views, each one corresponding to a defined connectivity region. Each collection view contains the list of index servers that should be used (along with their rank orders) by the user interface servers in that region.
- 2) *Regional Collection Servers (RCS)*. As described above, there is one RCS per connectivity region. An RCS provides the same collection information to the user interface servers in its region as the original single-site collection service. That is, it returns to them the set of rank-ordered index servers that they should use for query routing. Like the original implementation, the RCS gets the information from the CCS, which returns the collection view that corresponds to that region.

Figure 29 illustrates the interactions between the CCS and RCS and Dienst user interface servers. As shown, the central collection server (labeled CCS) contains internal tables that store, for each connectivity region, the server address of the RCS for that region and the collection view that corresponds to that region. In the figure, the RCS labeled S_1 (which is configured with the CCS as its collection server) submits a protocol request to the CCS to fetch a collection view. The CCS, recognizing S_1 as the RCS for R_1 , returns the appropriate collection view. The user interface server in R_1 ,

which is configured with S_1 as its collection server, then receives the correct collection view in response to its collection service protocol request to S_1 . It then uses the information in this collection view to make routing decisions to index servers. It should be noted that as network connectivity changes, the regional view could be re-defined at the CCS level. A region's collection view is modified once the RCS requests and receives new collection data from the CCS.

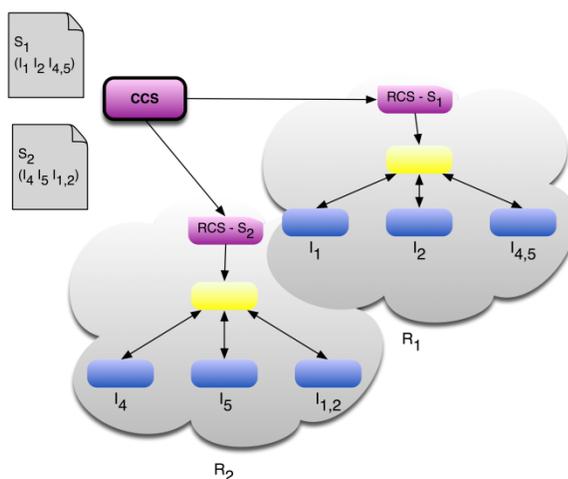


Figure 29 - Interactions of CCS, RCS, and user interface server

There is one final implementation note. If a server not listed with the CCS when an RCS submits a request to the CCS for a collection view, the CCS returns a view that is registered as the "default" collection view. In this fashion, any external service or agent can make use of the collection service for its own internal purposes.

Experience with the Architecture

At the time of completion of this paper in 1998, we had four operating regions within NCSTRL. A server operated by MTA/SZTAKI in Budapest, Hungary acts as an RCS for Dienst servers in Eastern Europe and Italy. A server operated by ICS/FORTH on Crete acts as an RCS for Dienst servers in Greece. A server operated by GMD in Bonn, Germany acts as an RCS for Dienst servers in Northern Europe. A server

operated by Cornell Computer Science in Ithaca, NY acts as an RCS for North American and some European servers with good Trans-Atlantic connections. We have found that connectivity between the West (especially the San Francisco area) and East coasts of the United States is often as bad as between Europe and North America. Because of this, we are investigating breaking up the North American region into two or possibly three regions.

The current configuration of regions was based mainly on conjecture, informal experience, and the willingness of particular Dienst sites to assume the greater reliability responsibilities required by an RCS. Thus, any conclusions based on our experiences are preliminary. In any case, we have found that the perceived reliability of the Dienst system, as measured from any of the user interface gateways, has improved dramatically. In addition, the architecture has proven quite easy to manage and adjust. Modifications to tables in the CCS are quickly propagated to the RCS's and thus to the Dienst servers in those regions. A server can easily be moved from one region to the next and the effect of unreliable servers can be isolated.

Our initial implementation of connectivity regions has uncovered a number of problems that we intend to address in future implementations.

- The implementation was retrofitted on top of a Dienst protocol and Dienst servers that pre-dated the regional architecture. Because of this we had to make a number of implementation compromises to avoid "breaking" legacy systems⁷³. One example of a problem that we have had is the imprecise and insufficient information about database freshness supplied by existing Dienst

⁷³ This is not an uncommon problem with distributed software for which a satisfactory solution will need to be found.

servers. This has made it difficult for us to propagate up-to-date replicas of indexing data between index sites.

- Connectivity problems remain troublesome. For example, the network speed between our Hungarian RCS and other Dienst systems is sometimes so bad that it is impossible to update index servers in that region.
- Server administration problems make it difficult to maintain index server integrity. When the primary source of indexing information is frequently unavailable or the quality of records is inconsistent, it is impossible to maintain useful replicas of that information.

As one strategy for eliminating these problems we are planning to logically segregate our production system from our research test bed. In this manner we can maintain production NCSTRL services, perhaps with a more centralized search strategy, and carry out research on isolated and controlled servers in the test bed. Ironically, the regional architecture can be used to create this segregation - in effect breaking off "production regions" from "research regions".

Finally, researchers outside of Cornell have experimented with the regional idea. For example, the MeDoc project has adapted the concept for defining content-specific regions or collections [16]. Researchers at ICS/FORTH in Greece and at IBM-Watson have used it in experimentation on QoS-based Searching and Retrieval [427].

Conclusions

As stated earlier in this paper future digital libraries architecture will have to address the problems inherent in distributed searching. They will have to do this in the context of global connectivity patterns. Our experience with NCSTRL has shown that the digital library infrastructure must provide information that supports query routing

decisions. Using this information, individual services can then algorithmically or heuristically decide the "best" destination(s) for protocol requests.

In the process of implementing and deploying Dienst and NCSTRL we have developed a number of useful abstractions for addressing this problem. This chapter has described three concepts that together have allowed us to globally distribute the NCSTRL collection.

- The *Collection Service* defines for user interface gateways the location of servers to which resource discovery queries can be routed.
- *Connectivity Regions* define the division of the complete set of servers into groups with relatively good connectivity characteristics.
- A *Collection View* is a definition of the collection, framed as the location of index servers, which corresponds to the connectivity characteristics of a connectivity region.

Chapter 8

Accommodating Simplicity and Complexity in Metadata

Preface

This chapter is based on:

Lagoze, C., *Accommodating Simplicity and Complexity in Metadata: Lessons from the Dublin Core Experience*, in *Seminar on Metadata*, (Archiefschool, Netherlands Institute for Archival Education and Research, The Hague, 2000) [282].

Stuart Weibel, who directed the Dublin Core Metadata Initiative (DCMI) throughout its early years, began his report of the meeting that led to the creation of the DCMI with this sentence:

The explosive growth of interest in the Internet in recent years has created a digital extension of the academic research library for certain kinds of materials. [480]

Therein lies the motivation for the substantial attention to metadata in the digital library community in its first decade⁷⁴. If the Internet is to be considered a “digital extension of the academic research library”, it seemed logical that there should be some form of a “catalog” for this extension.

Admittedly, there were technical reasons for this early interest in metadata – initial web search engines were crude due to the fact that their rankings were based on traditional information retrieval techniques, which were ill suited for the scale and lack of domain specificity of the web corpus [212, 436]. But, as described earlier in this

⁷⁴ Notably, while there is considerably less enthusiasm for metadata relative to this first decade, the Dublin Core Metadata Initiative is still active and work on metadata still continues in various forms.

dissertation, the justification for metadata extended beyond existing technical problems and more generally reflected the legacy of the catalog in the library meme and the belief in its central role in discovery. This belief fueled the continuation of DCMI as an active DL activity even after the huge improvement in search engine effectiveness due the introduction of link analysis.

The paper that forms the basis of this chapter was written in 2000 at a time that was probably the most active period of the Dublin Core Metadata Initiative. It covers three concepts that in combination demonstrate some of the initial assumptions of metadata work and the later tensions that developed in that work.

First, it continues the argument in Weibel's original paper that some form of semi-structured surrogate is a required part of the web discovery landscape and that users prefer the field-based discovery paradigm it supports. It is useful to examine this argument from the perspective of Arm's "viewpoint analysis" [26] that DL requirements "... developed from an organizational viewpoint... [assuming] continuity of existing organizations".

Examine, then, the following text from the paper included in this chapter that states:

The simple "one text box" approach used by existing web search engines, while useful, does not permit even the simplest type of search specificity. There are times that users find it desirable to be more specific in their searches.

Empirical evidence, however, has indicated that there is a clear user preference for the simple one text box approach. Apparently the assertion of what the user found "desirable" was a projection by the library community, which dominated DCMI, and had substantial investment in the catalog and its accouterments, such as fielded bibliographic search.

Second, it articulates some of the existing tension in the Dublin Core community, which developed around the issue of simplicity vs. complexity. Some participants supported maintaining Dublin Core as a simple descriptive format. Others proposed adding complexity via qualification to it for richer semantics. Conspicuously absent from this discussion, but one which surfaces in my later work described in Chapter 11, is the whole issue of the quality of metadata created by nonprofessional catalogers (those for whom metadata was developed as an alternative to rigorous cataloging), which has emerged as one of the most problematic issues for metadata deployment regardless of the simplicity or complexity of the format.

Third, the paper introduces the motivation for and details of more semantically-based modeling work that was taking place in the NSF-funded Harmony Project. This modeling proposal, called ABC, was based on the notion of lifecycle events and can be seen in retrospect as a precursor of the interest in “workflow management and representation” that we now see in the eScience and cyberinfrastructure communities [233]. The ABC/Harmony work is described in considerably greater detail in [166, 283, 292, 293].

Acknowledgements

The content of this paper is based on collaborations and discussions with various colleagues in the metadata and data modeling community. These include Bill Arms, Dan Brickley, Ron Daniel Jr., Martin Doerr, Jane Hunter, John Kunze, Clifford Lynch, Eric Miller, Sandy Payette, John Perkins, and Stuart Weibel. Support for the work in this chapter came from funding through NSF Grant 9905955. I owe special thanks to Tom Baker for his substantial editing of the original paper.

Realities for all occasions

Reality is chaotic. It consists of entities and objects of all types and forms. These entities change over time and sometimes morph into other distinct objects. As a result entities are interrelated in numerous and complex ways. Just limiting our domain to the document world, we see relationships such as translations, derivations, editions, versions, and citations, just to name a few.

People try to understand and work with this chaotic reality by simplifying it. Using categorization and classification they create artificial ordered realities in which entities fit into convenient slots. As noted by Bowker and Star [85], humans are insatiable classifiers who deeply fixate classification schemes into social, political, and scholarly structures. This categorization allows people to ignore the idiosyncrasies of individual entities and manipulate them via their coarse granularity group characteristics.

The modern library, and the Catalog that is at the core of its operations, is arguably the preeminent example of classification. From Melvil Dewey's conception of the Dewey Decimal Classification system [115] in the 19th century to the now preeminent MARC encoding⁷⁵ of AACR2 cataloging rules [213], libraries have been deeply engaged in classifying a variety of physical (and now digital) artifacts. In his excellent study of catalogers and their craft, David Levy described the process of "order making" [336], whereby a veneer of regularity is overlaid on the natural disorder of the artifacts that libraries encounter.

⁷⁵ <http://lcweb.loc.gov/marc/marc.html>.

The emergence of the web and the explosive growth of on-line content has challenged and enriched this order-making task. Interest in *metadata*⁷⁶ has evolved in response to this new and challenging context. While it shares many of the same purposes as cataloging, ordering and therefore simplifying the entities that it describes, metadata plays a somewhat different role due to the way the web differs from the traditional library environment.

The environment within which the Catalog and its standards exist is relatively self-contained and controlled. Creation and maintenance of catalog records is the task of a controlled community of expertise. The interface to Catalog records is generally restricted to specialized Integrated Library Systems (ILS) with little or no published interface to other systems. Finally, exchange of catalog records is regimented, usually in the form of MARC downloads from authorized sources like OCLC or RLG.

In contrast to this controlled community, the web is a bit like the Wild West. The maintenance of information is the purview of diverse communities with a variety of descriptive standards. The content and services provided by these diverse communities coexist in the same common space and their use frequently crosses

⁷⁶ In a rapidly developing and wide-ranging field such as metadata, finding the right common terms is a significant part of the problem. Throughout the remainder of this paper, we use a number of terms in relation to metadata:

Vocabulary – The set of elements (properties) provided by a specific metadata set.

Statement – The result of associating a metadata element with a resource and value (e.g., “Romeo and Juliet has a creator William Shakespeare”).

Record – A set of statements that collectively describe a resource.

Schema – The rules, or data model, for constructing statements in a metadata set.

Metadata Set – A “standard” for metadata that includes both a vocabulary and schema.

As described in this paper, the DCMES defines a limited vocabulary, the fifteen elements, and a simple resource-centric schema.

community boundaries. Stuart Weibel, who has led the Dublin Core Metadata Initiative since the beginning, has characterized this as the *Internet Commons*, where boundaries of control and use of information are blurred or non-existent. This concept is illustrated in Figure 30, where the circles indicate domains – the divisions among which are themselves frequently not distinct – and the arrows represent the exchange of information amongst these domain boundaries.

This environment presents both a challenge and opportunity for the formulators of descriptive metadata standards. The boundary-crossing nature of web use creates the need for descriptive standards that facilitate usability across domain and community boundaries. This requirement is often described under the rubric of *interoperability*.

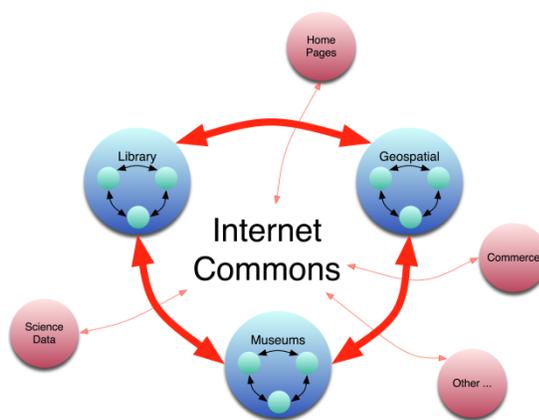


Figure 30 - Mixing information from multiple communities

While the need for interoperability in the Internet Commons is important, the fabulous diversity of the web also provides a unique opportunity for customization and specialization. As we noted in an earlier paper [284], metadata on the web should allow individuals to use and search the global information space conformant to their current roles and needs. We can think of metadata like a database view, capable of projecting multiple order-making schemes on content. This then makes it possible to customize the services that consume that metadata, for example search engines, and

tailor their functionality to differing needs. This concept is illustrated in Figure 31 where the same content, the painting of the Mona Lisa, may be projected via metadata in three views:

1. *geo-spatial* – that describes the specific location of the object and routes to use to find it. Such metadata might be useful for applications such as museum directories or tours on mobile devices⁷⁷.
2. *rights* – that emphasizes the identity of agents and organizations involved in ownership or management of the object. Such metadata might be useful in the production of copies and derivations of the original work.
3. *museum* – that emphasizes facets of the object associated with its exhibition and preservation.

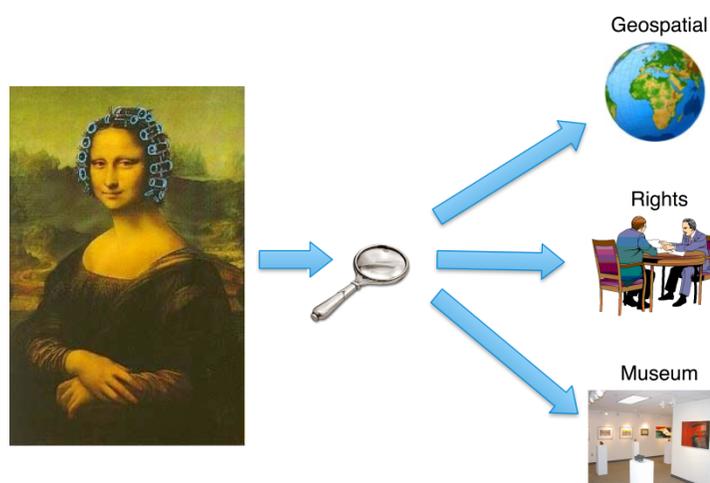


Figure 31 - Multiple views of the same content

Such multiple descriptive views are possible by using a *modular* approach, where separate metadata packages are associated with the resource. This is the approach taken by a number of web metadata architectures including the Warwick Framework

⁷⁷ John Perkins of CIMI has described to me a number of interesting applications of mobile devices in the museum environment.

and Resource Description Framework [92, 300, 323], which permit multiple communities of expertise to associate need and domain specific metadata with web content.

Such modularization has typically been described and justified from the perspective of domain specificity. However, the advantages afforded by such an architectural framework are amenable to other dimensions of specialization. This paper focuses on the simplicity/complexity dimension⁷⁸. This dimension plays an important role in resource discovery because of the manner in which users generally begin the discovery process using basic terms and then “drill-down” with higher levels of specificity in their queries. Simple descriptions, those that have been embodied in a number of “core” metadata sets, are extremely important components in such a strategy. Similarly, complex descriptions play an important role of permitting users to query more granular, often domain-specific, aspects of resources, or by providing information vital to the preservation, administration, and management of access to the resource. We argue that attempting to intermix in a single descriptive schema and vocabulary the simple semantics needed for coarse granularity queries with the complex semantics needed for drill-down queries, and other purposes, leads to metadata sets that are not ideally suited for either purpose. As an alternative, the modularization principles embodied in RDF and the Warwick Framework should be exploited to develop and deploy schema tailored for simplicity and others tailored for complexity.

⁷⁸ Complexity, as covered by this paper, is in the form of richness of description. We recognize that this is but one facet of complexity. Others, which play an important role in metadata and which also deserve examination, include versioning, multiple languages, and multiple encodings.

The paper will use the Dublin Core Metadata Element Set (DCMES) and its development history to illustrate these points. The purpose here is not to disparage the DCMES, which has proven enormously successful for its original purpose as a descriptive metadata set for coarse granularity resource description. We do mean to raise issue with the efforts in the Dublin Core Metadata Initiative (DCMI) over the past several years to re-purpose the DCMES as the basis for richer descriptions. In our opinion, and that of many others in the metadata community with whom we have discussed this issue, such an effort has interfered with the original goal and failed to provide a basis for such rich descriptions. We encourage the DCMI, and other communities involved in metadata developments, to turn to modularity when faced with a variety of descriptive requirements⁷⁹.

A world of document-like objects

The history of the Dublin Core has been well documented in a number of workshop reports[159, 480, 481, 483]. The purpose of this paper is not to replay this history. What follows is a brief review of the development of the DCMES to provide a context for the remainder of this paper.

The DCMI began in 1994 in response to the recognized need for better resource discovery tools for the web. This requirement grew out of dissatisfaction with two extremes:

⁷⁹ Throughout this paper we will clearly distinguish between two things: 1) the DCMES that is the 15-element set, and 2) the DCMI that is the organization that is examining metadata for networked resources and has the DCMES as its most visible result. Other tangible results of the DCMI are the Warwick Framework and much of the work on using RDF for descriptive metadata. This distinction corresponds to recent work of the DCMI that has involved refining and broadening scope.

- The standard cataloging methods in libraries were, and still are, too complex and expensive to provide a reasonable basis for resource description of web content. Whereas such methods may be appropriate for stable entities such as the physical artifacts that libraries collect, they are inappropriate for the dynamic web environment. web content is ephemeral and disseminated by a variety of sources that are often far removed from established publication authorities.
- The simple “one text box” approach used by existing web search engines, while useful, does not permit even the simplest type of search specificity. There are times that users find it desirable to be more specific in their searches. For example, even the simplest queries frequently need to distinguish between *by-ness* (e.g., books by Charles Dickens) and *about-ness* (e.g., books about Charles Dickens).

The product of the early Dublin Core meetings – one that has remained essentially stable and is recognized as the primary result of the DCMI – is the fifteen-element DCMES. These elements include some that are reasonably consistent across all domains – for example, creation of the resource, naming of the resources, subject of the resource – and others that some argue stand on the fringe of “core-ness” – such as, geospatial characteristics and rights management statements. Focusing on the exact composition of the Dublin Core elements is not the purpose of this paper. In fact, any argument about the exact composition of core semantics is rather moot since each community evidently has individual notions of such.

The more relevant task is to use the view metaphor mentioned earlier and thereby understand the nature of the “ordering” that DCMES imposes on content. This understanding can be inferred by looking at the types of resources that DCMES was

targeted at; simple web documents written in HTML. Much of the early literature on the DCMES characterized this type of object as a *document-like object*, or DLO.

The exact nature of a DLO has never been specified and, in fact, lack of specificity about its definition is central to its nature. Drawing from its humble origins, the simple web page, we argue that the essence of a DLO is simplicity in both structure and lifecycle. That is, a DLO is not composed of compound sub-parts nor is it characterized by complex inter-relationships with other resources, either physical or digital.

This simplicity may not actually correspond to any resource. An analysis of many web pages, even those from the earliest days of the Web, shows that very few of them are stand-alone items, and most have subtle and unexpected complexity. Exact correspondence of the DLO view to reality, however, is neither an important issue nor is it relevant to our argument. We take the perspective articulated by Borges who noted that "...there is no classification of the universe that is not fictional and conjectural." [74]

The relevance of the DLO is its usefulness as a simple metaphor for describing a mixture of resources and facilitating cross-domain, cross-genre resource discovery. By "pretending" that a cross-section of resources is uniformly simple we thereby make it possible to search for them in a simple manner. Two useful metaphors have surfaced during the development of the DCMES to express this simplicity:

- Pidgin Languages – Tom Baker, who is a member of the Dublin Core Executive Committee, has compared the DCMES to a pidgin language. Such a language ensues when individuals with different language backgrounds are mixed together (often forcibly as in refugee communities or during the time of American slavery). Inevitably, the members of such communities rapidly

develop a crude, syntax-less language for basic communication amongst themselves.

- Digital Tourist – Ricky Erway of Research Libraries Group (RLG) used the digital tourist metaphor for the DCMES. A tourist who travels to a country with another language brings a basic phrase book and develops a set of rudimentary phrases for communication during the visit. An examination of phrase books reveals that the elements of the basic language are quite uniform despite the target language differences (e.g., “I need a doctor”, “Where is the train station”, etc.).

Key to the simplicity of both of these simple languages is both limited vocabulary and simple structure. This means that the *statements that* can be constructed lack compound syntactic constructions (e.g., sub-phrases and complex clauses) and are mainly stated in the present tense. Events and entities are flattened into simple declarative phrases. Such flattening is central to the simplicity of the DCMES as originally conceived. Basic DCMES descriptions, which are a combination of very simple statements about single resources, project views of the resources that generally hide or obscure the complexity of their origins and derivations.



DC Element Set Record

<i>Title</i>	Mona Lisa in Curlers
<i>Creator</i>	Leonardo da Vinci
<i>Creator</i>	George Castaldo
<i>Date</i>	1506
<i>Date</i>	1994

Figure 32 - Flattening complex reality

We can examine this through an illustrative example shown in Figure 32. The example shows an image called “Mona Lisa in Curlers”⁸⁰. This image, created by George Castaldo in 1994, is clearly a derivation of the famous “Mona Lisa”, created by Leonardo da Vinci in 1506. As described later, such derivations have complex histories that consist of a variety of agents, tools, and events. Also shown in Figure 32 beside the image is, in a syntax free representation, one possible simple DCMES record for this image⁸¹. The record demonstrates the nature of flattening, whereby the creators *da Vinci* and *Castaldo* are placed at the same descriptive level, as are the two dates of “creation”. Without a doubt, the description as shown does not stand alone as a complete description of the artifact. On the other hand, the simple document-like view provides data for indexing that is useful for simple resource discovery. For example, users making a “digital tourist” query for resources by “da Vinci” would find the resource. Of course, such simple metadata does not permit queries about the type of software used for digitally processing the image, but such information falls outside the notion of pidgin and ventures into the more complex descriptive languages inhabited by domain experts.

⁸⁰ Mona Lisa in Curlers is © 1994 the American Postcard Co., Inc. as part of the *Misguided Masterpieces* Series. This image was extracted from the Web page at <http://www.pipeline.com/~rabaron/MONA08.htm>.

⁸¹ The flattening of what are essentially a number of creations into one record as shown in this example is not mandated by the DCMES. In fact, there has been considerable discussion in the DC community about the so-called *one-to-one rule* (due to David Bearman). Briefly stated, this “rule” expresses the notion that creators of DCMES records should recognize the dangers of too much flattening and, where appropriate, should create separate records for what are essentially separate items. Therefore, the single record in this example could be expressed as separate records with linkages between them using the *relation* element. Our view is that either form, the single record or multiple linked records, is not necessarily “correct” and one of the strengths of DCMES is the ability choose the compositions of the records based on what is most deemed most useful for resource discovery. As shown in the remainder of the paragraph, the single record as shown has definite use as a tool for discovery.

The next section of this paper explores further aspects of this example and describes problems that result from trying to extend the flat model in an open fashion.

Confounding the simple model

Agreement on simplicity and the nature of the DLO has certainly not been universal among the parties involved in the DCMI. Indeed, there has been substantial interest from the beginning in schemes to enrich the descriptive power of the DCMES and use it for purposes generally outside the application of cross-domain resource discovery. Rather than think of the DCMES as a simple view of richer descriptions, some have sought to use the DCMES as a mechanism for creating rich cataloging records⁸².

The early discussions about this were described as a division of the DC community into the *minimalists* and the *structuralists*. The former advocated that the most valid use of the DCMES was in their simple, unadorned free-text form; the latter were interested in methods of enriching the capabilities of the DCMES through various mechanisms. The arguments by the latter group centered on the fact that the unadorned elements were simply insufficient for “real description” of any resources. We do not disagree with this argument, but maintain that “real description” cannot be done in a generic manner (it is context-specific) and that neither the schema nor vocabulary of the DCMES is sufficient for such description.

Over time, the minimalist/structuralist discussion evolved into the characteristics of *qualification*. Broadly speaking, qualification consists of mechanisms for adding semantic specificity to DCMES descriptions. While the basic fifteen elements have

⁸² As a corollary, there has been a general presumption that DC elements are fixated in physical records. As we note in [15], mechanisms whereby DCMES elements are computationally projected from more complex descriptions stored in databases may be the more sensible and scalable approach.

remained almost invariant over the first five years of DCMI, the issue of qualification has been a sea of shifting interpretations and models. Much of the difficulty has been devising a means of accommodating complexity and extensibility with the simplicity of the original DLO view. In theory, individual communities should be able to establish qualifiers to elements, tailored to their domain-specific needs. Furthermore, the DC records enhanced with these qualifiers should be able to interoperate with records containing qualifiers devised by other communities and with DC records that employ just the simple unqualified semantics. The key principal for accomplishing this interoperability is the notion that element qualifiers should *refine* rather than *extend* element semantics⁸³.

A simple example of such semantic refinement and its use is illustrative. Table 3 shows, in natural language, qualification of the DCMES *creator* element, which is defined as “An entity primarily responsible for making the content of the resource”⁸⁴. In the example, one community has defined a qualifier *poet* as a specialization of *creator*, and another has defined a qualifier *author*. The nature of semantic refinement makes it possible for a search engine, which has processed the two “facts” and knows about the specialization relationship, to answer more general queries such as that shown in Table 3. This generalizing, stripping off qualifiers and returning to the base element form, should make it possible for diverse communities to essentially ignore

⁸³ We are ignoring here another form of qualification informally known as *value qualification*. This form allows the specification of controlled vocabularies or encoding rules for element values. An example of such is the association of the encoding rule “LCSH” to a DC *subject* value to indicate that the value term is described in the Library of Congress Subject Headings. Work by Tom Baker using sentence construction metaphors indicates that the distinction between the two “types” of qualification may indeed be a red herring.

⁸⁴ This definition is taken from the Dublin Core Metadata Element Set, Version 1.1: Reference Description at <http://purl.org/DC/documents/rec-dces-19990702.html>.

qualifiers that are unknown to them, yet make sense of the records. The Dublin Core community has coined the term *dumbing-down* for this process of stripping off semantic refiners and returning to base forms.

Table 3 - Employing the "dumb-down" principle

- *Joseph Brodsky* is the **poet** of *Discovery*
- *Joseph Brodsky* is the **author** of *Watermark*
- A **poet** is a specialization of a **creator**
- An **author** is a specialization of a **creator**
- Find me objects of which *Brodsky* is the **creator**
 - *Discovery*
 - *Watermark*

We don't doubt that such a qualification model, employed in a controlled fashion, is possible. Control would mean that a central authority, for example the DCMI, defines a fixed and relatively simple qualification set that adheres to the principles outlined above. Yet, this is not the model that has been consistently promoted by the DCMI and therein lies a problem with dumbing-down. Instead, a model that has been frequently proposed by the DCMI is that qualification occurs in a distributed, community-specific manner and can be used for increasing levels of complexity and specificity⁸⁵.

Such an approach is flawed in both its motivation and its execution. Establishing a pathway for extensive and essentially unlimited qualification of the DCMES presupposes a broader scope than the original "simple resource discovery". It frames

⁸⁵ In fact, lack of clear guidelines for qualification and the lack of a clear definition of scope for the DCMES have led to a proliferation of Dublin Core records that are qualified in ways that defy dumbing down.

the DCMES a one-stop cataloging standard with which records can be constructed that describe any and all facets of resources and their related entities (their creators, their intended audiences, etc.) Even if such an expanded scope were acceptable, the execution of it within the framework of the DCMES is problematic for number of reasons.

- *Model* -Building complex descriptions on top of the flat DC data model is fundamentally flawed since the model makes it difficult to distinguish between different entities and their attributes.
- *Vocabulary* -The DCMES elements themselves were not originally engineered for such complex descriptions. The elements are completely non-normalized, ranging from ones that are essentially data types (e.g., *date*) to ones that are facets of a more general concept (e.g., *creator*, *contributor*, and *publisher* are all facets of agency). Other elements such as *rights* appear to contribute no information that is actionable by a computer within the context of user queries. Furthermore, qualification uncovers relationships among the elements that should be expressed structurally. If *published* is a qualifier for the *date* element, how does this relate to the *publisher* element? If *scanned* is a qualifier for the *date* element, how does this relate a *format* element that lists *tiff* as one of its values? Such engineering sloppiness is acceptable, and in fact may be the best method, for fulfilling the original DCMES intent, pidgin metadata. It is not appropriate as the basis for a rich descriptive framework.
- *Process* -As pointed out by John Perkins of CIMI⁸⁶, the notion of refinement is implicitly community-specific. A guideline such as “qualifiers shall only refine

⁸⁶ CIMI, the Consortium for the Interchange of Museum Information, has been one of the leading experimenters with DCMES. CIMI established an XML DTD for DCMES and worked with its members to create a large number of unqualified records. Its experiments

and not extend element semantics”⁸⁷ will inevitably be interpreted by communities in fashions that will make dumbing-down impossible and defeat the interoperability goal. Our fear is a balkanization of the element set with DCMES qualified by community “A” incompatible with that qualified by community “B” and neither compatible with those who wish to use the element set in its simple unqualified form.

Discussions within the DCMI over qualification of the agent elements (*Creator*, *Contributor*, *Publisher*) demonstrate the nature of the problem. One of the qualifiers that was suggested for common use was *affiliation*, indicating the organization with which the individual is affiliated. (This qualifier was subsequently rejected based on the principles described in the next paragraph). From the perspective of many communities, affiliation was a clear refinement of agent semantics. However, there are serious problems with dumbing-down such a qualifier as indicated in Figure 33. The first panel of the figure shows the use of such a qualifier in a simple HTML syntax[291]⁸⁸ for a record associated with a book by *Allison Lurie*, who is affiliated with *Cornell University*. The second panel of Figure 33 shows a simple unqualified record for a book by the author *Gary Cornell*. The third panel shows the result of “promiscuous” dumbing-down⁸⁹ of the record; stripping off the qualifiers and accepting the tokens –*Alison, Lurie, Cornell* – as values for the element creator. A

indicated that DCMES, without qualification, served as a useful basis for coarse level resource discovery. However, later attempts to extend those experiments to the use of qualification indicated that these semantic extensions interfered with the original core interoperability requirement. The comments reported here are personal communication with John Perkins in May, 2000.

⁸⁷ This paraphrasing captures the essence of the qualification guidelines in heretofore unpublished documents of the DCMI.

⁸⁸ This syntax is for illustrative purposes and is not the syntax being considered within DCMI.

⁸⁹ Credit goes to Ron Daniel Jr. for this phrase.

simple query on the creator field would degrade the quality of the search through false hits. While the problem may seem trivial and easily fixable for this simple example, the problem becomes intractable with a huge number of records and unlimited qualification by distributed communities. It is unacceptable to either promiscuously dumb-down – making false hits the rule rather than the exception – or, as an alternative, throw out qualified values – creating the balkanization alluded to earlier.



Figure 33 - Uncontrolled qualification vs. interoperability

Tom Baker and others associated with the DCMI proposed a solution that may prove to be a workable compromise. The solution builds on the notion that qualification of the DCMES should proceed on well-defined *qualification principles* that constrain qualifiers to basic semantic refinement (e.g., defining sub-types of elements such as *illustrator* as a qualifier for *creator*) and value encoding (e.g., defining that a *date* value is encoded according to ISO8601). These interoperability principles will be publicly disseminated, accompanied by a set of exemplary qualifiers that demonstrate the principles, in a DCMI document due third quarter 2000. Baker then proposes that qualification proceeds within distributed communities but that a *usage board*, similar

to that which exists for natural language dictionaries, periodically vets qualifiers in use and maintains a registry of qualifiers with annotation indicating their conformance to the interoperability principles. Such a registry would, over time, be implemented using mechanisms such as RDF schema that would permit implementations that consume DC metadata descriptions to automatically check conformance to published interoperability standards. This solution does not prevent communities from developing qualifiers that confound interoperability (there is no solution to that problem) but is beneficial both because it is grounded in the principle of DCMES as a vehicle for simple resource description and provides a mechanism for monitoring conformance to that principle.

In June 2000 this compromise was adopted as the official policy of the DCMI. Since that time it has served as a useful vehicle for maintaining the scope of the DCMES and thereby facilitating interoperability among DCMES records. The experiences of the past, however, where scope was lost in the context of creeping functionality, suggest that a certain level of vigilance vis-à-vis these principles will continue to be necessary.

Agents of change

We began this paper by mentioning reality and its complexity. In this section we look at this complexity with a finer lens and attempt to understand how this complexity impacts descriptive schema.

The example illustrated in Figure 33, in which attributes of the agent (affiliation) are intermixed with attributes of the resource, provides a glimpse into the core of our argument, which is as follows. Key to the simplicity of descriptive schema such as the DCMES, as it was originally conceived, is their resource-centric data model. This model, as described earlier, has a simple flat structure whereby attributes (e.g. *title*) and their values (e.g., *Mona Lisa in Curlers*) are associated with reasonably mundane

(real-world) objects – e.g., documents, books, pictures, and the like. Richer more complex descriptions confound this model since they inevitably include the attributes of multiple entities. This was shown in Figure 33 where attributes of the creator were intermixed with attributes of the document. Such intermixing, as shown earlier, compromises the effectiveness of the schema as a mechanism for descriptive interoperability.

We suggest that that a framework for building richer descriptions must address two needs:

- *Expanded and Refined Vocabulary* – Qualification of the DCMES effectively expands the available vocabulary that can be used in descriptions. However, as noted earlier, this expansion needs to build on a more refined foundation that accounts for the fact that the core vocabulary will serve as the root of more complex terms.
- *Expressive Structure* – The data model, the rules for assembling the metadata vocabulary into statements, should be able to unambiguously express the boundaries between different entities.

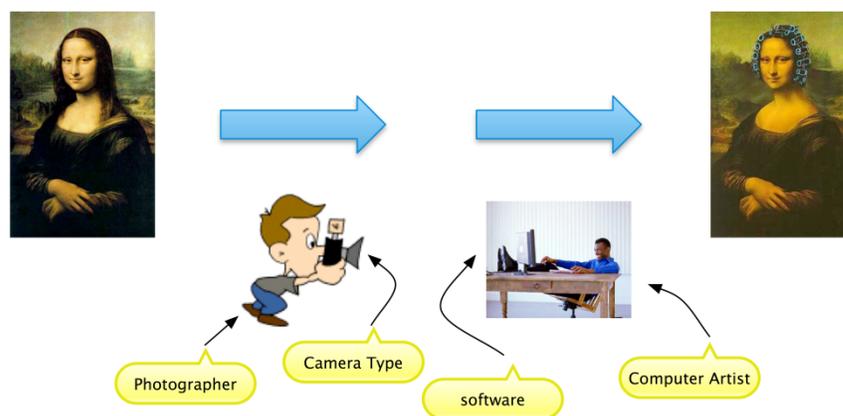


Figure 34 - A closer look at resource, entities, and their relationships

We do not suggest that these more complex needs be met by ignoring the goals that originally motivated the DCMES. Instead we advocate establishing frameworks for the creation of more complex descriptions that can co-exist with simpler ones as separate packages in a Warwick Framework or RDF like container framework⁹⁰.

A closer look at the “Mona Lisa in Curlers” illustrates the nature of more expressive structure. Figure 34 illustrates some of the complexity underlying the “Mona Lisa in Curlers” resource. As indicated, the Castaldo work is a derivation of the original work by Leonardo da Vinci. The derivation process consisted of a number of events and agents and tools related to those events. For example, the image of the original da Vinci painting was digitized perhaps by a photographer using a specific type of digital camera. This digital image was then altered using some image processing software (e.g., Photoshop) by an artist on some type of computer system. There are numerous other details and processes not shown in the illustration. The essential point is that complex descriptions that meet the needs of specific descriptive communities will involve descriptions of these other entities. For example, the digital imaging community would certainly be interested in descriptions of the camera type and imaging software. Yet, providing those descriptive components within the flat data model, as attributes of the Mona image, presents the problem described in the earlier section.

If a flat resource-centric data model is not sufficient for richer descriptions, then what is the better alternative? This is an issue that is being examined by a number of

⁹⁰ As we describe in [284], the notion of separate packages can be a logical, rather than a strictly physical concept. Given a well-defined and expressive underlying representation, such as that expressed in the ABC model [93], it should be possible to project automatically both simple DCMES views and more complex views.

descriptive communities. For example, the bibliographic community (i.e., Libraries) has become increasingly aware of the shortcomings of the generally flat AACR [213] model for describing resource inter-relationships. The IFLA FRBR framework [3] recognizes the lifecycle aspects of intellectual content and distinguishes between abstract works, their manifestations, and the items that are produced from those manifestations. Similarly, the rights management community [424, 425] has noted that representing transactions and the information related to those transactions is essential for metadata concerned with managing intellectual property. Finally, the archival and record-keeping communities have stated the importance of process orientation for descriptions [41, 42].

Our own work in the Harmony Project⁹¹ builds on these earlier efforts and argues that *event-awareness* is vital for the understanding and expression of more detailed descriptions of resources. The details of this are beyond the scope of this paper and are described more fully in other documents [93, 294]. A brief summary of the event-aware concept is as follows.

Resources, as shown simply by the Mona Lisa in Curlers example and as noted in the IFLA FRBR, are the tangible result of an evolution of transformations and derivations. An important aid towards understanding this evolution of an individual resource and the derivative relationships between resources is to characterize the events that are implicit in the evolution or derivation. For example, the evolution from *work* to *expression* may contain an implicit *composing event*. The process of making implicit events explicit – making them *first-class objects* – provides well-defined attachment points for common descriptive concepts such as agency, dates, times, and roles. The

⁹¹ <http://metadata.net/harmony>

clear and uniform definition of such attachment points then makes it possible for automated processes, computer programs, to unambiguously distinguish between entities and their attributes.

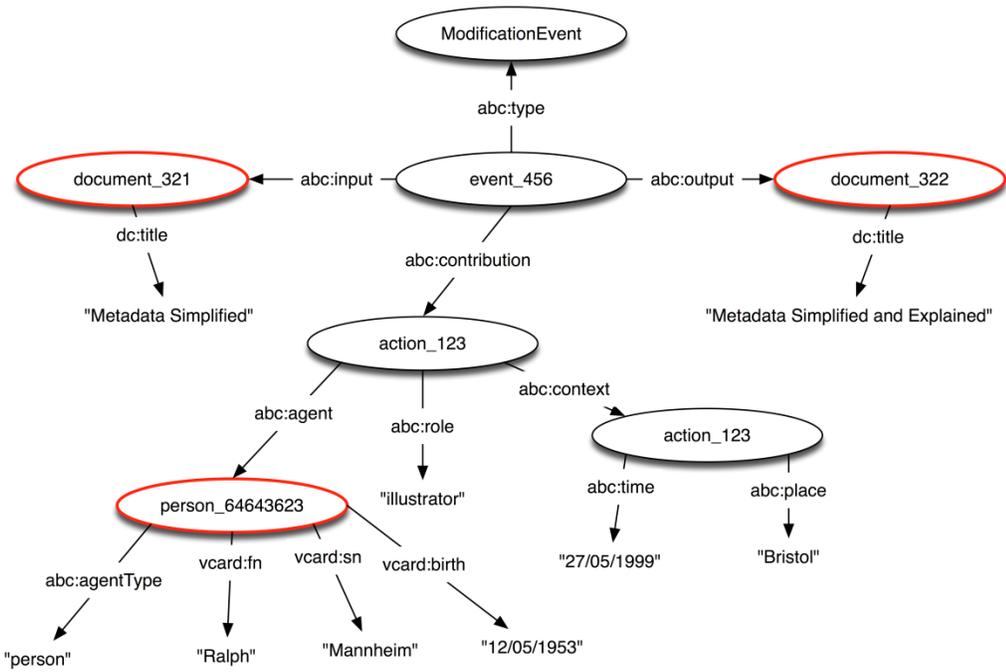


Figure 35 - Event aware descriptive data model

We have been experimenting in the Harmony Project with a highly expressive form of an event-aware model. This is illustrated in Figure 35 using an RDF-like graph representation. The figure shows a transition between two resources, labeled “document_321” and “document_322”. As shown, separate entities such as agents, roles, documents, and contexts are cleanly separated in the model. Such separation makes it possible for programs to cleanly differentiate between dates related to the agent “person_64643623” born on “12/5/1953” from the date related to the modification of the resources (the date “27/05/1999” that is the context of the modification event).

Is it all worth it?

Undoubtedly the model presented in Figure 35 is complex. Furthermore, the representation, manipulation, and querying of such a model will require tools far more powerful than simple HTML META tags or existing relational databases. (Such tools are currently the subject of the extensive work by both the RDF and XML communities in the W3C).

We do not suggest that this level of complexity is the only possible alternative to the simple DC schema. Simplicity and complexity are two endpoints along a spectrum and, correspondingly, we need metadata sets that are well-suited to the varying points along this spectrum. At the same time, we need to seriously consider functionality vs. cost trade-offs when formulating metadata standards and applying them. Bill Arms points out such a trade-off in [28] and raises the issue that reduced functionality (simplicity) and reduced costs may be the proper choice for a large class of resource discovery needs. Applying Pareto's 80/20 rule, perhaps we would see an 80% improvement in overall resource discovery, at a relatively low cost, through the application of very simple descriptive schema. Perhaps the use of a very simple core set of descriptive elements (maybe not even fifteen) makes more sense from a cost/benefit standpoint than extensive work on complex representations and the creation of descriptions that match those representations. Certainly serious investigation and evaluation in the digital library and metadata community on this cost/functionality trade-off would be a sensible path of research.

Whatever the conclusions of such a study, we argue that trying to hide complexity in a simple metadata set such as DCMES leads to unacceptable compromises. The resulting metadata is inadequate for simple discovery purposes – dumb-down principles don't work – nor is it sufficient for complex description, which requires

clear delineation of entities and their properties. Our goals for discovery and description are better served by abiding to principles of modularization and seeking solutions that are bounded by well-defined scope and purpose.

Chapter 9

An Architecture for Complex Objects and their Relationships

Preface

This chapter is based on:

Lagoze, C., Payette, S., Shin, E. and Wilper, C. Fedora: An Architecture for Complex Objects and their Relationships. *International Journal of Digital Libraries*, 6 (2). 124-138 [301].

Fedora, the Flexible Extensible Digital Object Repository Architecture, is an architecture and implementation of a powerful digital object model and repository for storage of and access to those objects. It is perhaps the most successful result of our work on digital libraries at Cornell.

The Fedora work is interesting because of its roots in early digital library concepts, and because of its longevity due to the flexibility of its initial design. Initial work on Fedora began in 1997 as an outgrowth of our earlier work on extensions and implementations of the Kahn/Wilensky Framework [290], Dienst [152], and the Warwick framework [147, 148, 286]. The Kahn/Wilensky Framework, which was described in greater detail in earlier chapters, provided the foundational notion of a *digital object*, an identified container aggregating multiple data streams. The Dienst work contributed the notion of a repository and protocol-based access to digital objects that conform to a document model. Finally, the Warwick Framework work, and its subsequent extension with the notion of *distributed active relationships*, which occurred within the Dublin Core metadata effort, contributed the concept of linking data with services to produce active disseminations of content.

As described in the remainder of this chapter, recent work has extended the original concepts of Fedora into the realm of the semantic Web, service-based architectures, and policy enforcement. As stated in the text the “recent work uniquely integrates advanced content management with semantic web technology.” In this manner, Fedora is both an exemplar of the library meme-based notions described at the beginning of this dissertation blended with the Web 2.0 concepts that developed alongside digital library efforts. Perhaps this is the reason behind the enduring nature of the Fedora Project in the manner in which it is recognized as the technological leader amongst competing projects such as DSpace, Greenstone, and ePrints. Indeed, the words behind the name Fedora – Flexible Extensible Digital Object Repository Architecture – have indeed proven true as it has successfully bridged the digital library and Web 2.0 worlds.

Fedora is still a very active and increasingly influential project, especially in the expanding cyberinfrastructure research area. Under the leadership of Sandy Payette, the Fedora project has now been spun off into a not-for-profit called Duraspace, which focuses on open source projects to support durable digital content⁹².

Acknowledgments

As a project that has spanned more than 10 years, the success of Fedora is based on the efforts of a large number of people. Primary among them is Sandy Payette, the current Chief Executive Officer of DuraSpace, the not-for-profit corporation that is currently responsible for Fedora development and maintenance. Sandy was my chief collaborator on Fedora during the DARPA-funded and NSF-funded research phase. She subsequently led the Fedora project in its transition to widely disseminated open

⁹² <http://www.duraspace.org/>

source software and architectural integration with developing web frameworks.

Through a unique combination of organizational and technical leadership, Sandy has brought Fedora from its humble beginnings as a research project to a leader in open source technology. Other primary contributors to the work over the years include the rest of the Fedora team at Cornell including Chris Wilper, Eddie Shin, and Dan Davis. In addition, significant contributions have been made by the team at University of Virginia led by Thornton Staples and Ross Wayland. Finally, the large and diverse Fedora community should be recognized for their contributions.

Initial support for the Fedora work came from Digital Library Initiative funding from the National Science Foundation and DARPA. More recent funding is due to the generosity and wisdom of the Andrew W. Mellon foundation, especially Don Waters who as program manager has continued to support Fedora activities. Finally, the transformation from Fedora to Fedora Commons has been possible due to extremely generous support from the Gordon and Betty Moore foundation, and the efforts of Jim Omora, a program officer at Moore.

Introduction

At a minimum, technologies for representing digital content should be able to match the richness, and complexity of well-established physical formats. As such, they should allow the representation of a variety of structural organizations, such as chapters and verses; accommodate the flexible combination of different genre of materials, such as text and images; and allow the aggregation of content from multiple sources and the association of metadata with the elements of the aggregation.

However, freed of the constraints of physical media, digital content architectures should do more. Exploiting their networked context, they should allow aggregation of content regardless of its physical location. By leveraging local and remote computing power they should support programmatic and user-directed manipulation of digital content. Finally, they should represent the complex structural, semantic, provenance, and administrative relationships among digital resources.

This paper describes our latest work on Fedora, an open-source digital content repository service, which provides a flexible foundation for managing and delivering complex digital objects. This recent work uniquely integrates advanced content management with semantic web technology. It supports the representation of rich information networks, where the nodes are complex digital objects combining data and metadata with web services and the edges are ontology-based relationships among these digital objects.

The motivation for integrating content management and the semantic web originates from requirements defined by the broader Fedora user community. The most familiar example is the need to express well-known management relationships among digital resources such as the organization of items in a collection and structural relationships such as the part-whole relationships between individual articles and a journal. While the relationships among digital objects in these familiar applications are mainly hierarchical, we are working with other applications where the relationships are more graph-like. For example, in the NSF-funded NSDL (National Science Digital Library) Project⁹³, we are using Fedora to implement an information network overlay that represents local and distributed resources and the provenance, managerial, and

⁹³ <http://nsdl.org>

semantic relationships among those resources. We report on the results of this work later in this paper.

While there are a number of schemes for representing these relationships such as conventional relational databases and formalisms like conceptual graphs [440], the products of the semantic web initiative such as RDFS [92], OWL [155], and highly-scalable triple-stores such as Kowari [496] provide extensible open-source solutions for representation, manipulation, and querying these knowledge networks.

The remainder of this paper describes the details of the Fedora architecture that provides the foundation for these rich applications. The structure of this paper is as follows. We begin by summarizing the historical development of Fedora. The next section provides core background on the Fedora digital object model, articulating a graph-based view of the model that is consistent with the semantic web orientation of our latest work. After that, we described the Fedora relationship model that provides a common framework for describing, storing, and querying relationships among objects and their components. The penultimate section describes results of the deployment of Fedora, focusing on applications that exploit features that distinguish Fedora from related work. The final section is a conclusion.

Background

The Fedora Project⁹⁴ is an ongoing research and development effort to provide the framework for creation, management, and preservation of existing and evolving forms of digital content. The roots of the project lie in DARPA-funded research in the early 1990's that defined the notion of a *digital object* [255] and implemented Dienst [289], a networked digital library architecture with protocol-based dissemination of digital

⁹⁴ <http://www.fedora-commons.org/>

objects in multiple formats. Follow-on research extended these initial concepts with the notion of *active digital objects* [148] and *distributed active relationships* [147]. These concepts were refined and prototyped in a CORBA-based Fedora (Flexible Extensible Digital Object Repository Architecture) [403] as part of research with CNRI [401] and in the context of the NSF-funded Prism Project [295]. This prototype provided the context for a variety of research initiatives most notably in the areas of fine-grained policy enforcement [404] and preservation [405].

The transition of Fedora from a research prototype to production repository software began when the University of Virginia Library, seeking a solution for managing increasingly complex digital content, experimented with the Fedora architecture [443]. This experimentation took place in the context of innovations in humanities research. The experimentation proved successful, providing the basis for subsequent funding from the Andrew W. Mellon Foundation to Cornell and Virginia [406] to jointly develop Fedora and make it available as open source software to libraries, museums, archives, and content managers, facing increasing variety and complexity in the digital content that they manage [444]. Mellon-funded development continues through 2009.

The richness of the Fedora digital object model and extensibility of the Fedora service-based architecture has led to its deployment in a variety of domains including digital libraries [279, 298], institutional repositories [457], electronic records archives [488], trusted repositories for digital preservation [248], library systems⁹⁵, educational technologies [279], web publishing⁹⁶, and distributed information networks [298].

⁹⁵ <http://www.vtls.com/products/vital>

⁹⁶ <http://www.encyclopedia.chicagohistory.org/>

Fedora is implemented as a set of web services that provide full programmatic management of digital objects as well as search and access to multiple representations of objects. All Fedora APIs are described using the Web Service Description Language (WSDL) [124]. As such, Fedora is particularly well-suited to exist in a broader web service framework and act as the foundation layer for a variety of multi-tiered systems, service-oriented architectures, and end-user applications. This distinguishes Fedora from other complex object systems that are turn-key, vertical applications for storing and manipulating complex objects through a fixed user interface (e.g., DSpace [439], arXiv⁹⁷, ePrints⁹⁸, Greenstone [494]).

By providing both a model for digital objects and repository services to manage them, Fedora is also distinguished from work focused on defining and promoting standard XML formats for representing and transmitting complex objects (e.g., METS⁹⁹, MPEG-21 DIDL[245], IEEE LOM [4]). However, Fedora is compatible with these efforts since it has the ability to ingest and export digital objects that are encoded in such XML transmission formats¹⁰⁰. This allows Fedora to comfortably coexist in the archival framework defined by OAIS [418].

As a service-based architecture for complex digital objects, Fedora has some commonality with the aDORe architecture [461] developed at the Los Alamos National Laboratory research library. The aDORe system provides a standards-based

⁹⁷ <http://arxiv.org>

⁹⁸ <http://eprints.org>

⁹⁹ <http://standards.loc.gov/mets>

¹⁰⁰ Fedora currently supports ingest/export of digital objects encoded using METS and also the Fedora XML wrapper format (FOXML). Future releases will support MPEG-21 DIDL and possible other formats.

repository for managing and accessing complex digital objects. Objects are encoded in XML using DIDL [47] and a limited set of object relationships can be expressed using RDF. Object dissemination services are available via OAI-PMH [314] and OpenURL [12].

Fedora model for complex objects

The Fedora object model supports the expression of many kinds of complex objects, including documents, images, electronic books, multi-media learning objects, datasets, computer programs, and other compound information entities. Fedora supports aggregation of any combination of media types into complex objects, and allows the association of services with objects that produce dynamic or computed content. The Fedora model also allows the assertion of relationships among objects so that a set of related Fedora objects can represent the items in a managed collection, the components of a structural object like the chapters of a book, or a set of resources that share common characteristics (defined by semantic relationships).

Fedora defines a powerful object model for expressing this variety of complex content and their relationships. This object model can be understood from two perspectives.

- The *representational* perspective defines a simplified abstraction for understanding Fedora objects, where each object is modeled as a uniquely identified resource projecting one or more views, or *representations*. From this perspective the internal structure of a digital object is opaque; however, relationships among objects are observable.
- The *functional* perspective reveals the object components that underlie the representational perspective and provides the basis for understanding how the Fedora object model relates to the management services exposed in the Fedora repository architecture.

Representational View

The representational perspective of the Fedora object model asserts that each digital object can disseminate one or more representations of itself, and that each object can be related to one or more other objects. A familiar example of a digital object with multiple representations is a document or image where the content is available in multiple formats. All digital objects, and their individual representations, are identified with Uniform Resource Identifiers (URIs). These URIs are specified using the “info” scheme and conform to the syntax described at [462]. This choice frees the architecture from ties with any identifier resolution system (e.g., the Handle System [449]).

This perspective hides complexity and exposes only the access points to content stored in a Fedora repository. Figure 36 depicts the representational view of three inter-related Fedora objects. The diagram shows a directed graph, where the larger nodes are digital objects, and the smaller nodes are representations of the digital objects¹⁰¹. These nodes are linked by two types of arcs – relationship arcs connect digital objects, and representation arcs connect digital objects to their respective representations. This graph can be expressed as RDF, stored in a triple store, and queried. This is discussed a later section.

Each digital object in the diagram has at least one representation, related to its originating digital object by a `hasRep` arc. For example, the node labeled `info:fedora/demo:11` is an image digital object with four representations, identified by their respective URIs:

¹⁰¹ This graph-based overlay model can form the basis for interoperability among heterogeneous object models and repositories. This concept was explored as part of an NSF-funded research project, Pathways, a collaboration between the authors of this paper and colleagues at Cornell, LANL, and others [466].

- -Dublin Core record, identified as `info:fedora/demo:11/DC`
- -High-resolution image, identified as `info:fedora/demo:11/HIGH`
- -Thumbnail image, identified as `info:fedora/demo:11/THUMB`
- -Image with zoom/pan utility, as `info:fedora/demo:11/bdef:2/ZPAN`

We have yet to define the underlying source of these representations. In fact, in this view of the architecture such details are hidden from the client application concerned with access to these representations.

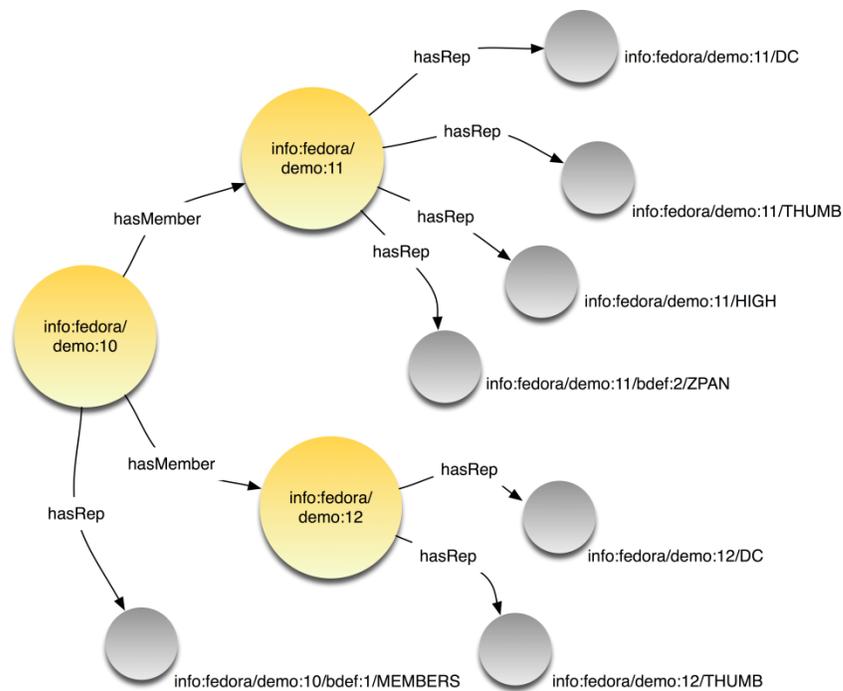


Figure 36 - Representational view of Fedora objects

Figure 36 also demonstrates an example of inter-object relationships. In this example, the node labeled `info:fedora/demo:10` is a “collection” with two “items”, the nodes labeled `info:fedora/demo:11` and `info:fedora/demo:12`. These collection-item relationships are expressed by the `hasMember` arc that emanates from the collection object. The inverse `isMemberOf` relationships are not shown in the diagram for simplification.

This simple representational view forms the basis of Fedora’s REST-based access service (i.e., API-A-LITE), whereby digital object URIs and representation URIs can be easily converted to service request URLs upon Fedora repositories.

Functional View I - Datastreams

While the representational perspective of the Fedora object model provides a simple, access-oriented overlay for digital resources and collections, the *functional perspective* provides a view of the core underlying data model for Fedora. In the following sections, we take one of the digital object nodes depicted in Figure 36, and drill down to unveil the specific components of a Fedora digital object that enable access to representations. We start with the digital object as a container with a persistent unique identifier (i.e., PID). From there, we unveil the components incrementally, first focusing on components that enable simple content aggregation, then on components that enable dynamic and computed content, and finally on components related to digital object integrity. We note again that these underlying details are invisible to clients concerned only with information access.

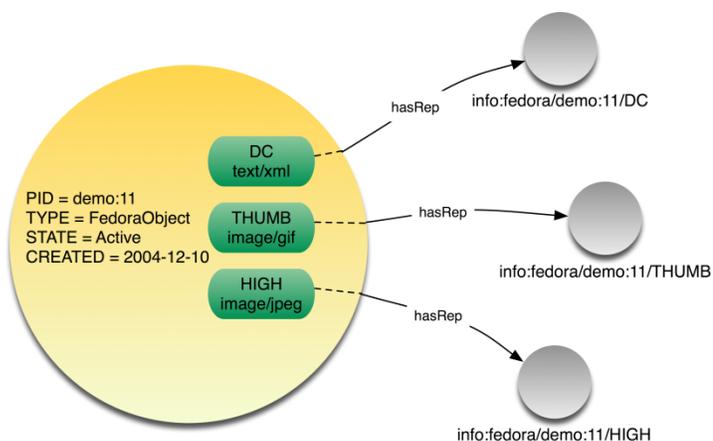
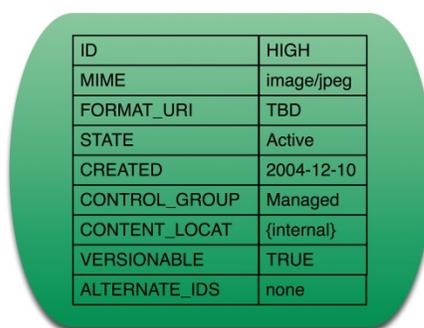


Figure 37 - Fedora object with PID, properties, and datastreams

In its simplest form a Fedora object is an aggregation of content items, where each content item maps to a representation. The Fedora object model defines a component

known as a datastream to represent a content item. A datastream component either encapsulates bytestream content internally or references it externally. In either case that content may be in any media type. Figure 37 shows a digital object as an aggregation of datastreams and the one-to-one correspondence of those datastreams to the representations of the digital object that are exposed to accessing clients. In this simple case, each representation of a Fedora object is a simple transcription of the content that lies behind a datastream component.



ID	HIGH
MIME	image/jpeg
FORMAT_URI	TBD
STATE	Active
CREATED	2004-12-10
CONTROL_GROUP	Managed
CONTENT_LOCAT	{internal}
VERSIONABLE	TRUE
ALTERNATE_IDS	none

Figure 38 - Properties of a datastream component

As seen in Figure 37, a digital object has a unique identifier (PID) and a set of key descriptive properties. Each datastream contains information necessary to manage a content item in a Fedora repository. These are stored as properties of the datastream as shown in Figure 38.

Three datastream properties deserve special attention. The Format URI refines the media type definition and anticipates the emergence of a global digital format registry such as the GDFR¹⁰². Control group defines whether the datastream represents either local or remote content. Datastreams with a control group of “Managed” are internal content bytestreams that are under the direct custodianship of the Fedora repository.

¹⁰² <http://hul.harvard.edu/gdfr/>

Datastreams whose control group is “External” or “Redirected” (the difference between these is outside the scope of this paper) represent content that is stored outside the repository. These datastreams have a content location property that is a URL pointing to a service point outside the repository that is responsible for providing the content. The ability to create digital objects that aggregate locally managed content with external content is a powerful feature of Fedora, and is useful in a variety of contexts. A good example of a hybrid local/remote object is an educational object where local content is an instructor’s syllabus, lecture notes, and exams, and remote content are primary resources included by-reference from other sites.

Functional View II - Disseminators

In addition to the representations described in the previous section, which are direct transcriptions of datastreams, the Fedora object model enables the definition of *virtual representations* of a digital object. A virtual representation, also known as a *dissemination*, is a view of an object that is produced by a service operation (i.e., a method invocation) that can take as input one or more of the datastreams of the respective digital object. As such, it is a means to deliver dynamic or computed content from a Fedora object.

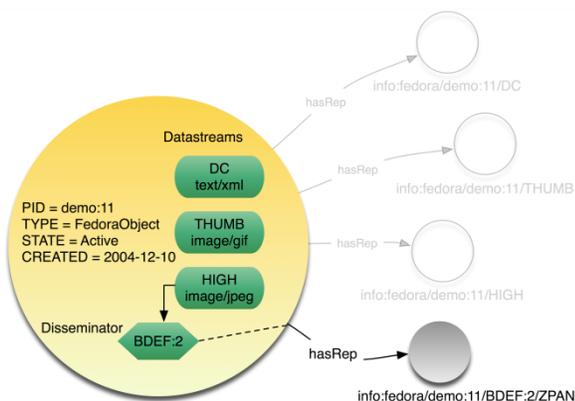


Figure 39 - Fedora object with disseminator added

This is illustrated in Figure 39, where a virtual representation labeled `info:fedora/demo:11/BDEF:2/ZPAN` is highlighted. From the access perspective this representation is an image wrapped in a java application that provides image zoom and pan functions. Note that this representation is not a direct transcription of any Datastream in the object. Instead, it is the result of a service operation defined in the Disseminator component labeled `BDEF:2` inside the object that uses the datastream labeled `HIGH` as input. The light-weight, REST-based interface to Fedora (API-A-LITE) makes it possible for a client application to pass parameters to the invoked service; in this case zoom and pan specifications.

To enable such behavior, a Disseminator must contain three pieces of information: (1) a reference to a description of service operation(s) in an abstract syntax, (2) a reference to a WSDL service description [124] that defines bindings to concrete web service to run operation(s), and (3) the identifiers of any Datastreams in the object that should be used as input to the service operation(s).

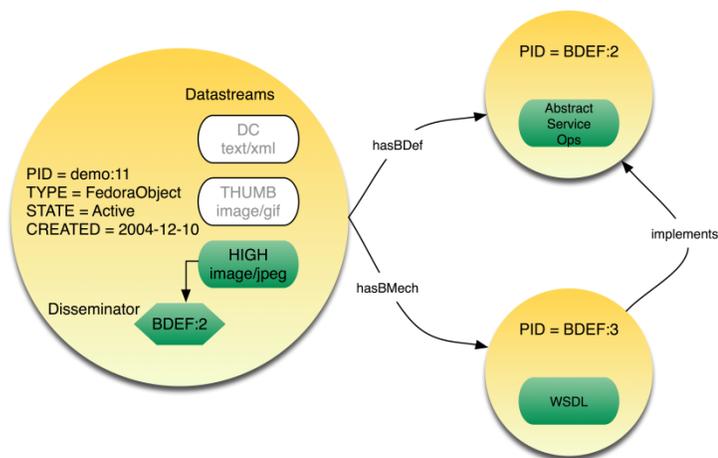


Figure 40 - Disseminators establish relationships to service definition objects

Fedora stores the service operation description and the WSDL service description within special digital objects, respectively known as BDefs (behavior definitions) and

BMechs (behavior mechanisms). Figure 40 depicts a Fedora BDef object and a BMech object along with object-to-object relationships that exist due to the presence of the Disseminator component in the main object (i.e., `demo:11`).

Disseminators are effectively metadata that a Fedora repository uses at run time to construct and dispatch service requests and produce one or more virtual representations of the digital object. From a client perspective this is transparent since virtual representations look just like other representations of the object.

Disseminators are a powerful feature in the Fedora object model. They can be used to create common representational access points for digital objects that have different underlying structure or format. For example, an institutional repository might contain scholarly documents in a variety of root formats (e.g., Word, HTML, TeX), where the root format is stored as a datastream in a Fedora digital object. For interoperability purposes, a virtual representation can be defined on each object that converts the datastream containing the root format to a common format (e.g., PDF). Similarly, a repository manager can decide for archival purposes to convert all documents in a repository to a canonical preservation format without disrupting the manner in which clients access documents for browsing, viewing, etc. Finally, disseminators can add utility operations to digital objects. For example, a Disseminator can be defined for a digital object that provides parameterized query access to the relationships defined for that object. Such a query might return the “members of a collection” or, in the case of an educational digital library such as the NSDL [498], the set of resources that are appropriate for K-12 mathematics education. The implementation of these queries is described later.

Functional View III – Object Integrity Components

The Fedora object model defines several metadata entities that pertain to managing the integrity of digital objects. These entities are the object’s relationship metadata, access control policy, and audit trail. To keep the Fedora model simple and consistent, integrity entities are modeled as datastream components with reserved identifiers. As such, the integrity entities are stored like other datastreams, however the Fedora Repository system recognizes them as special and asserts constraints over how they are created and modified. Figure 41 depicts these integrity-oriented entities as special datastreams in a digital object, identified as Relations, Policy, and Audit Trail.

A Relations datastream is used to assert object-to-object relationships such as collection/member, part/whole, equivalence, “aboutness,” and more. The previously discussed “hasMember” relationship is an example of the type of assertion that can be managed via the Relations datastream, described later in this chapter.

A Policy datastream is used to express authorization policies for digital objects, both to protect the integrity of an object and to enable fine-grained access controls on an object’s content. In Fedora objects, a policy is expressed using the eXtensible Access Control Markup Language (XACML)¹⁰³, which is a flexible XML-based language used to assert statements about who can do what with an object, and when they can do it. Object policies are enforced by the authorization module (i.e., AuthZ) implemented within the Fedora repository service.

¹⁰³ http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml

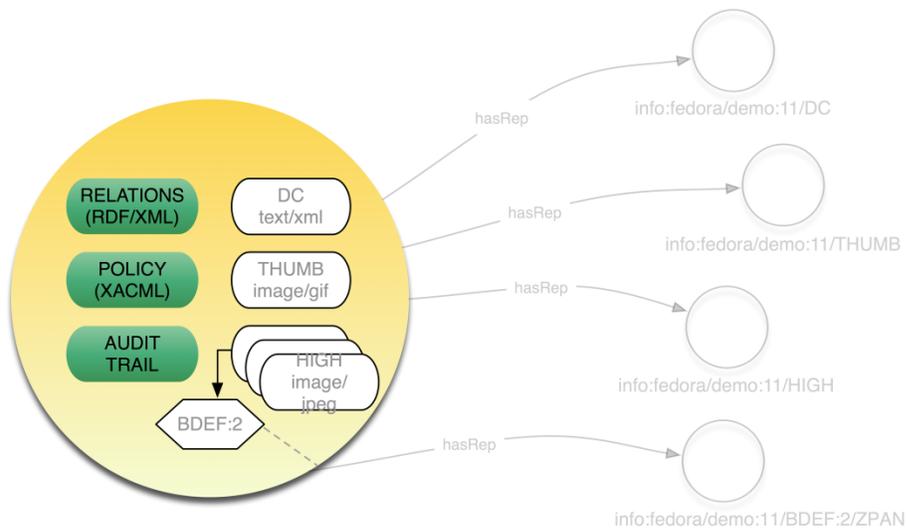


Figure 41 - Integrity datastreams - relationships, policy, and audit trail

The Audit Trail is a system-controlled datastream that keeps a record of all changes to an object during its lifetime. The Fedora repository service automatically creates an audit record for every operation upon an object, detailing who, what, when, where, and why an object was changed. This information is important to support preservation and archiving of digital objects.

Another feature for managing the lifecycle of objects is versioning. Versioning is important for applications where change tracking is essential, as well as for preservation and archiving systems that must be able to recover historical views of digital objects. The Fedora object model supports component-level versioning, meaning that datastreams and disseminators can be changed without losing their former instantiations. Fedora automatically creates a new version of these components whenever they are modified.

This is depicted in Figure 41, which shows a digital object with multiple versions of a datastream (see component labeled HIGH). Also, the versioned datastream is input to the disseminator labeled BDEF:2. Requests for representations of this digital object

can be date-time stamped and the Fedora repository service will ensure that the appropriate component version is returned. This feature applies for representations that are direct transcriptions of datastream content, as well as for virtual representation where datastream content is mediated via a Disseminator.

Relationships in Fedora

As described earlier, the Fedora object model can be abstractly viewed as a directed graph, consisting of *internal* arcs that relate digital object nodes to their representation nodes and *external* arcs between digital objects. In this section we focus on that relationship graph and describe the Fedora *Resource Index* module, which allows storage and query of the graph. This module builds on RDF (Resource Description Framework) [273] primitives developed within the semantic web community. The Fedora system defines a base relationship ontology that, in the fashion of any RDF properties, can co-exist with domain-specific ontologies from other namespaces. Each digital object's relationships to other digital objects are expressed in RDF/XML [44] within a reserved datastream in the object. The Resource Index is a relationship graph over all digital objects in the repository that is derived by merging the relationships implied by the Fedora object model itself with the relationships explicitly stated in an object's relationship datastream. The triples representing this graph are then stored in a triple-store providing the capability for searching over the graph.

The combination of representing explicit relationships as RDF/XML in a datastream of a digital object and then mapping them to the triple store offers the "best of both worlds". The explicit representation provides the basis for exporting, transporting, and archiving of the digital objects with their asserted relationships to other objects. The mapping to the triple store provides a graph-based index of an entire repository and the basis for high-performance queries over the relationships. An added advantage

of the dual representation is that the entire triple store can be rebuilt by importing and parsing the XML-based digital objects.

Representing object-to-object relationships

There has been a significant amount of work in the area of structural metadata for complex objects. These efforts have been focused on developing XML schema for expressing structural relationships with individual digital objects. One early example was the Making of America [161] project that formalized structural metadata and defined a set of templates that correspond to well-known physical artifacts such as a book composed of chapters and diaries consisting of entries. The current exemplar of this encoding of structural metadata is in METS [347].

Our focus in Fedora has been to decompose these structural units into separate digital objects. The motivation for this is that the units can then be reused in a variety of structural compositions. In addition, this lays the basis for expressing other types of non-structural relationships among digital objects such as:

- The organization of individual resources into larger *collection* units, for the purpose of management, OAI-PMH harvesting [314], user browsing, and other uses.
- The relationships among *bibliographic* entities such as those described in the Functional Requirements for Bibliographic Relationships [3].
- *Semantic* relationships among resources such as their relevance to state educational standards or curricula in an educational digital library like the National Science Digital Library [498].

- Modeling more complex forms of *network overlays* over the resources in a content repository such as citation links [202, 236], link structure, friend of a friend¹⁰⁴, etc.

All of these relationships, including structural relationships, should be expressible both within individual digital objects and among multiple digital objects. For example breaking the components of a structural entity, such as the chapters of a book, into separate digital objects provides the flexibility for reuse of those individual components into other structural units. This is even more important for the other forms of relationships. For example, a single resource may be part of multiple collections or may be relevant for multiple state standards.

The remainder of this section describes a relatively simple example of inter-object relationships to demonstrate how these are expressed in Fedora. The simple techniques illustrated here can be used to express more complex inter-object relationships. In a later section, we will describe a more complex example in the context of our NSDL work.

The expression of arbitrary, inter-object relationships in Fedora is enabled by a reserved datastream known as the Relations datastream. This datastream allows for a restricted subset of RDF/XML where the subject of each statement must be the digital object within which the datastream is defined.

Since predicates from any vocabulary can be used in Relations, the repository manager has considerable flexibility in the kinds of relationships that can be asserted. Table 4 shows an example Relations datastream in a Fedora digital object identified by the URI, `info:fedora/demo:11`. The RDF/XML refers to three different relationship

¹⁰⁴ <http://www.foaf-project.org/>

vocabularies (hypothetical for the purpose of this example) and asserts the following relationships:

- `demo:11` is a member of the collection represented by the object `demo:10`,
- `demo:11` fulfills the state educational standard represented by the object `demo:Standard5`,
- `demo:11` is a manifestation of the expression represented by the object `demo:Expression2`.

Table 4 - Example relations datastream

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:nsdl="http://nsdl.org/std#"
  xmlns:rel="http://example.org/rel#"
  xmlns:frbr="http://example.org/frbr#">
  <rdf:Description rdf:about="info:fedora/demo:11">
    <rel:isMemberOf rdf:resource="info:fedora/demo:10"/>
    <std:fulfillsStandard rdf:resource="info:fedora/demo:Standard5"/>
    <frbr:isManifestionOf rdf:resource="
  info:fedora/demo:Expression2"/>
  </rdf:Description>
</rdf:RDF>
```

Object representations and properties in the Resource Index

As described earlier, a Fedora digital object consists of a number of core components such as datastreams and disseminators, which bind to BDefs and BMechs. In addition each Fedora digital object has system metadata or properties. The architecture provides a system-defined ontology to represent the relationships among these core components. For example, the relationship of an object to its representations is expressed using the `<fedora-model:disseminates>` predicate as shown in the triple in Table 5.

Table 5 - Object-representation relationship

```
<info:fedora/demo:11>  
<fedora-model:disseminates>  
<info:fedora/demo:11/HIGH>
```

In addition to these relationships, the system-defined ontology also represents object data properties whose range contains date and boolean datatypes, as shown in the triple in Table 6.

Table 6 - Data type properties

```
<info:fedora/demo:11/HIGH>  
<fedora-view:lastModifiedDate>  
"2004-12-12T00:22:00"^^xsd:dateTime
```

Unlike the relationships expressed in the Relations datastream, these relationships are not explicitly asserted within the digital object. Instead they are derived from the structure of the object itself and mapped into the Resource Index, alongside the relationships represented in the Relations datastreams. This is described in the next section.

Storing and querying the relationship graph

All these relationships – the relationships explicitly stated in the Relations datastream, the relationships implied by the object structure, and the data relationships contained in the object properties – are stored in the Resource Index. This index is automatically updated by the repository service whenever an object structure is modified or its Relations datastream is changed.

The Resource Index handles queries over these relationships. The combination of all relationships into a single graph, and the automated management of that combined graph, enables a powerful and flexible service model. External services may issue queries combining relationships from different name spaces, since they are all RDF properties. For example,

Table 7 shows a query listing all the representations of all objects that are members of a particular collection.

Table 7 - Sample RDF query using iTQL

```
select $dissemination
from<#ri>
where ($object <fedora-view:disseminates> $dissemination)
and $object <rel:isMemberOf><demo:10>
```

An early design goal of the Resource Index was to allow the use of different triplestores and thus permit the Fedora repository administrator to choose the most appropriate underlying store. To that end, the Resource Index employs a triplestore API similar in spirit to JDBC, to provide a consistent update and query interface to a variety of triplestores. Extensive testing of both query performance time and query language features ultimately led to the selection of Kowari as the default triplestore¹⁰⁵.

The query interface to the relationship graph currently supports three RDF query languages, RDQL¹⁰⁶, iTQL¹⁰⁷, and SPARQL [410]. Both RDQL and iTQL share a superficially similar syntax to SQL, with RDQL enjoying broader implementation support, but iTQL providing a richer feature set [223].

The RDF query results naturally take the form of rows of key-value pairs, again similar to the result sets returned by a SQL query. However, it is often useful to work with a sub-graph or a constructed graph based on the original. To this end, the query API may also return *triples* instead of *tuples*.

105 Our preliminary report on query performance of various triple store technologies is available at <http://tripletest.sourceforge.net/2005-06-08/index.html>.

106 <http://www.w3.org/Submission/2004/SUBM-RDQL-20040109/>

107 <http://kowari.org/271.htm>.

Using the relationship graph

The Resource Index is exposed as one of the interfaces of the core Fedora repository service. This facilitates the development of other services in the Fedora Service Framework¹⁰⁸. The Resource Index interface is exposed in a REST architectural style to provide a stateless query interface that accepts queries by value or by reference. The service has been implemented with an eye toward eventual conformance to the W3C Data Access Working Group's SPARQL protocol for RDF [127], as it matures. One example of a service exploiting the Resource Index is the OAI Provider Service that exposes metadata about resources in a Fedora repository. This OAI Provider Service is quite flexible in that it can be configured to allow harvesting not only of static metadata formats, but those that are dynamically produced via service-based disseminations of Fedora objects.

Table 8 - A query to build an OAI response

```
select $member $collection $dissemination
from<#ri>
where $member <rel:isMemberOf><info:fedora/demo:10>
and $member <rel:isMemberOf> $collection
and $member <rel:isMemberOf> $dissemination
and $member <fedora-view:disseminates> $dissemination
and $dissemination <fedora-
view:disseminationType><info:fedora/*/bdef:OAI/getQualifi
edDC>
```

An example of the interaction of this service with the Resource Index is as follows. An external OAI harvester requests qualified Dublin Core records for a particular set

¹⁰⁸ <http://fedora.info/download/2.1/userdocs/server/features/serviceframework.htm>

of resources from the repository. The OAI Provider service processes this by issuing the query to the Resource Index listed in Table 8. This query effectively requests “all *disseminations* of qualified Dublin Core records of resources that are members of the collection identified as `demo:10`. The significance of requesting disseminations is that the Dublin Core records may not statically exist as datastreams within the object, but they may be derived from another metadata format such as MARC. The Resource Index query would return the tuples shown in Table 9 that can provide the basis of an OAI response. Note that the OAI representations were not shown earlier in Figure 36.

Table 9 - The query response as triples

member	collection	dissemination
<code>info:fedora/demo:11</code>	<code>info:fedora/demo:10</code>	<code>info:fedora/demo:11/bdef:OAI/getDC</code>
<code>info:fedora/demo:12</code>	<code>info:fedora/demo:10</code>	<code>info:fedora/demo:12/bdef:OAI/getDC</code>

Results

Fedora has been tested in the field with the real-world collections of our collaborators. These applications demonstrate the flexibility of the Fedora object model and repository service to accommodate a diverse set of information management problems. They distinguish Fedora from seemingly similar architectures such as DSpace, arXiv, and ePrints, whose focus is primarily on institutional repositories for scholarly publications. The applications supported by Fedora include not only complex digital library collections¹⁰⁹ [279] and institutional document repositories [457], but also

¹⁰⁹ <http://www.lib.virginia.edu/digital/collections/>

electronic records archives [488], trusted repositories for digital preservation [248], and distributed information networks such as the NSDL. This section describes results of Fedora deployment in three of these contexts: the University of Virginia digital library collections; the Encyclopedia of Chicago, a multimedia cultural heritage resource; and the National Science Digital Library, a distributed information network of educational resources and contextual information about those resources.

The core functionality of Fedora has proven effective for integrating rich digital collections at the University of Virginia Library (UVA). The UVA digital repository is built upon well-defined “content models” for digital content, where a content model specifies the number and types of datastreams and disseminators for particular genre of complex digital objects, including images, books, letters, archival finding aids, and data sets. The result is a seamless integration of content in a repository that enables consistent management of digital objects, consistent interfaces to access digital objects, and easy re-use of digital materials in different contexts. The architecture provides a means to easily aggregate materials from different collections and create new views on content using both Fedora relationship metadata and custom disseminators. One example is a cross-collection object that builds upon multiple objects: one that disseminates architectural drawings about historical buildings, another that contains recent photographs of those buildings from art collections, and another that contains historical letters that mention the buildings from the electronic text collections.

Northwestern University’s use of Fedora demonstrates how Fedora’s flexibility allows the management and publication of rich multimedia objects. Most compelling is the

electronic version of the Encyclopedia of Chicago¹¹⁰ produced in collaboration with the Chicago Historical Society. The encyclopedia is a multi-media resource that exploits the Fedora disseminator capability in novel ways. A notable feature of this application is the design of digital objects and disseminators to create rich, clickable maps. These maps are linked to disseminators that provide multi-dimensional views and contextual information about a location in Chicago. For example a street map of Chicago highlights sites of labor unrest. By clicking on the map, a user can discover and launch numerous disseminations that link to population statistics, newspaper articles, and historical data that relate to a particular place on the map. This is all done using well-designed digital object content models and rich, service-based disseminators to produce dynamic transformations of digital object content. Nearly every piece of content on the web site is a dissemination of a Fedora digital object, and interestingly, the entire web site itself is published via a single dissemination of a master collection object.

Our work in the context of the NSF-funded NSDL (National Science Digital Library) Project [498] is perhaps the most interesting example of the power of Fedora's relationship architecture. Our goal in the NSDL is not only to provide a digital library allowing search and access to distributed resources, but also to augment NSDL resources with context that defines their usability and reusability in different learning and teaching environments. By "context", we mean information such as the provenance of the resources, the manner in which resources have been used, comments by users that annotate and explain primary resources, and linkages between the resources and relevant state educational standards. While the NSDL work is

¹¹⁰ <http://www.encyclopedia.chicagohistory.org/>

specifically targeted at the education domain, we argue that the notion of *contextualization* is increasingly important as a means of adding value to digital content and defining its quality based on provenance, utility, and other factors.

Using the content management and semantic web tools in Fedora we have implemented an *information network overlay* [298]. This architecture represents the data underlying the NSDL as a graph of typed nodes, corresponding to the information entities in the NSDL, and semantic edges representing the contextual relationships among those entities. The nature and variety of these relationships will evolve over time and, thus, any fixed schema approach for representing the network overlay would be too restrictive. Our results thus far indicate that the semantic web approach of Fedora is particularly well-suited for this application.

Figure 42 illustrates a fragment of the network overlay. The nodes in the overlay graph correspond to Fedora digital objects – each shape corresponding to an information entity in the NSDL. These entities include agents, resources, metadata, and the like. The edges are relationships among these entities, which are represented in the Resource Index. For example, Figure 42 shows the grouping of resources in collections, and the provenance trail of who originally recommended those resources and who manages them. Relationships from other ontologies, such as state education standards, are overlaid on this base graph. These are similarly represented in the Resource Index alongside the base ontology relationships. The entire knowledge base can then be queried by external services to build rich portals for users and tools for inferring quality, usability, and educational value.

foundation for the deployment of extended services and user/client applications that apply Fedora in a variety of domains.

Increasingly rich digital content is placing greater demands on the institutions responsible for the creation, storage, management, and preservation of that content.

Fedora is well positioned to meet those demands and its open architecture provides the basis for meeting new requirements as they develop in the future.

Metadata Aggregation and “Automated Digital Libraries”

Preface

This chapter is based on:

Lagoze, C., Krafft, D., Cornwell, T., Dushay, N., Eckstrom, D. and Saylor, J., Metadata aggregation and "automated digital libraries": A retrospective on the NSDL experience, in *Joint Conference on Digital Libraries (Best paper award)*, (Chapel Hill, NC, 2006), ACM [296].

This chapter describes two large, influential projects, NSDL and OAI-PMH, both of which demonstrate a number of the themes that are central to this dissertation. Each project reflects the influence of library organizational principles and practices on digital library work, reveals the problematic nature of those principles and practices in the digital domain, and illustrates efforts to move away from that traditional meme towards a model that reflects developments in the broader web context.

The National Science Digital Library Project (NSDL)¹¹¹ grew out of a late 1990s vision [479] of a network-based technology and organization to support that technology with the goal of facilitating innovations in science, engineering, and mathematics education in the United States. The program was initiated in 2001 with funding from the NSF Directorate for Education and Human Resources (EHR). Over

¹¹¹ The inherent tensions within the NSDL project are evident even in the varying words attached to the mnemonic NSDL over the years. These words include “National STEM (Science, Technology, Engineering, and Mathematics) Digital Library”, National Science Digital Library, and most recently “National STEM Education Distributed Learning”. The last of which indicates the disenchantment of the NSF program managers as a whole with the “digital library” concept as a vehicle for advancing teaching and learning.

the course of the past eight years NSDL grants have been given to over 200 institutions for a variety of purposes including collection development, service prototyping and implementation, and basic research. The work described in this chapter occurred under the auspices of the core integration (CI) program that funded a collaboration between Cornell University and the University Corporation for Atmospheric Research (UCAR), who were charged with creating a technical and organizational infrastructure for NSDL¹¹².

As described in this chapter the work of CI was explicitly and purposefully modeled in the manner of a traditional library transferred to the digital domain. This included an architecture built on the fundamentals of a metadata-based catalog that “resembled a well-exercised union catalog model”. The decision to use this historically-based architecture was based on both expediency and “reflect[ed] the principles within the CI team” who “felt that structured metadata should be at the core of a production Digital Library”. Notably, several of the key members on this team were professional librarians. Furthermore, it reflected the belief then widespread in the digital library community, that “simple metadata”, in the form of Dublin Core, was the proper vehicle for translating traditional library cataloguing principles to the more highly scaled and distributed digital library environment.

The other technology central to be NSDL architecture was the Open Archives Initiative - Protocol for Metadata Harvesting (OAI-PMH), a project that I co-directed

¹¹² In the latest round of funding the notion of core integration has been replaced with an organization known as technical network services (TNS), which along with a companion organization known as the resource center, is responsible for original core integration services and for the development of the sustainability plan for the NSDL. TNS is a collaboration between Cornell, UCAR, and the University of Colorado with myself as the principal investigator.

with Herbert Van de Sompel from 1999 through 2004¹¹³. OAI-PMH is perhaps one of the most widely deployed and well-known technologies for digital library interoperability. It reflects four threads of thinking prevalent in digital library research and development at that time.

- First, it is based on the premise that discovery occurs within the boundaries of the digital library, relying on a technology distinct from that in the mainstream web architecture. As stated earlier, this assumption has proven to be false, with users employing mainstream web search engines as their primary discovery mechanism.
- Second, the work on OAI-PMH reflects the realization that federated searching, to which considerable effort was expended in the Dienst Project (see Chapter 7) and other digital library projects, was impractical and should be replaced with some form of centralized indexing.
- Third, OAI-PMH demonstrates the awareness that simplicity, at the cost of some functionality [28], was a principle that should be followed in digital library infrastructure design. OAI-PMH was intentionally crafted as a step-back from the full functionality of the Dienst protocol, which preceded it.
- Finally, OAI-PMH presumes that structured metadata (i.e. surrogates) are necessary and fundamental to a digital library environment, and that the Dublin Core vocabulary is the key to “semantic” interoperability.

It is interesting to note that the OAI-PMH work occurred at the same time as similar work in the web community on syndication formats, in the guise of RSS [407] and later ATOM [386]. These two syndication technologies, both of which are

¹¹³ The Open Archives Initiative (OAI) still exists and is responsible for the Object Reuse and Exchange (ORE) work described in Chapter 12.

fundamentally semi-structured data transmission formats, are arguably two of the most important technological components of Web 2.0, enabling information aggregation and mashup. In retrospect, the decision to develop in parallel a special digital library metadata packaging format was ill-founded, and efforts to integrate with the parallel developments of the broader web community should have been undertaken.

The paper that forms the basis of this chapter describes in great detail the problems that arose with this metadata-centric architecture. The overriding lesson learned from the early NSDL experience is that assumptions underlying library principles and practices, such as the existence of a well-controlled and contained environment, break down when put in the context of the decentralized, anarchistic networked information environment. Fundamentally, the traditional library works so well because of the implicit and explicit agreements amongst a uniform professional community. These agreements are both organizational and technological. As has been shown by the history of the Web, the best that can be expected in the largely anarchistic web world is basic agreement on simple technical infrastructures and there can be little reliance on any uniform practices or use of these technologies.

Acknowledgments

Work on NSDL described here was supported by the National Science Foundation under Grants No. 0227648, 0227656, and 0227888. There were a large number of collaborators on this work but the following people deserve special acknowledgment: Bill Arms, Dean Krafft, Dave Fulker, Susan Jesuroga, Kaye Howe, Naomi Dushay, and Diane Hillmann.

Work on OAI-PMH described here was supported by the Andrew W. Mellon Foundation, the Coalition for Networked Information, and the Digital Library Federation. There were a large number of collaborators on this work including the

global OAI-PMH community. Special acknowledgment goes to: Herbert Van de Sompel, Michael Nelson, and Simeon Warner. Special thanks to Clifford Lynch for his continued support of OAI-PMH.

Introduction

Starting in 2002, the NSDL Core Integration team (CI) developed and administered an expanding education-focused digital library. The visible presence of this digital library is the main NSDL portal¹¹⁴. Underlying this portal is an architecture based on the aggregation of metadata from multiple sources, the storage of that metadata in a metadata repository (MR), and the provision of services that consume and process that metadata. One of these services is a Lucene-based search engine that indexes metadata in the MR and, if possible, the full-text content that the metadata references. The NSDL architecture was initially described in an earlier paper [288]. Our choice of this architecture was motivated by a mixture of factors:

- *Expediency*: The NSF grant to CI mandated the launch of a production NSDL presence soon after the initiation of funding. This required that the system use established tools and standards, and that it embody familiar practices rather than innovative techniques. We adopted OAI-PMH [314] and Dublin Core [2] based on these criteria. Similarly, we implemented the MR in an Oracle® RDBMS because it permitted the use of familiar “enterprise” system management techniques. Finally, because the metadata-based architecture resembled the well-exercised union cataloging model, we believed that production methods from that model could be used in the NSDL environment.

¹¹⁴ <http://nsdl.org>

We recognized that these methods would have to be modified due to the differences in complexity between Dublin Core records and library cataloging records and because metadata creators in the NSDL were both widely distributed and were generally not professional catalogers. These design choices were successful in meeting the rapid deployment mandate – the “initial launch” of the NSDL occurred in December 2002, a little over a year after the initiation of CI funding.

- *Philosophy:* The choice of structured metadata and the union catalog paradigm reflected principles within the CI team. From the beginning we intended that the initial architecture would evolve to a “spectrum of interoperability” [32], which would accommodate other less traditional paradigms (e.g., focused crawling, automated classification). However, many members of the CI team felt that structured metadata should be at the core of a production digital library. Like many mainstream digital library efforts, they had confidence that structured metadata was a well-known and easily exploited means of making precise information available to library services, such as search and discovery.
- *Finances:* Finally, the initial decisions about how to build the NSDL reflected the nature of the CI budget. Over the years, CI has received approximately 4M USD annually from the NSF, with the expectation that most of this would be used for library development, rather than day-to-day operations. This mandated an operational strategy that relied on automation, exploiting relatively inexpensive computers and networks, rather than on expensive human effort [28]. In general, cataloging has been a human-intensive activity in libraries [187]. “Low-barrier” standards such as Dublin Core and OAI-PMH were designed to reduce and distribute this cost, and the NSDL built a library based on such expectations.

As noted, the availability and relative simplicity of the individual architectural components facilitated rapid deployment. However, our three years of experience with the NSDL have contradicted our original expectations of automation and low people cost. We have learned that “low-barrier” standards have been more difficult for contributors to use than expected. Moreover, despite the relative simplicity of the individual components in the NSDL, the combination of these components, plus maintaining them on a 24x7 basis, adds up to a system of surprising complexity. There are multiple data feeds, many software components, and multiple machines that are distributed over multiple organizations and locations. The number of components and variables to be managed has frequently interfered with our efforts to handle the process automatically, forcing us to fall back on expensive human intervention. At times, this human effort has consumed developer time that otherwise could have been used to widen the spectrum of interoperability and innovation.

This paper provides a retrospective on three years (2003-2005) of running a relatively large-scale digital library (over a million objects) by collecting, processing, storing, and using metadata. Our intent is not to argue for or against the utility of metadata aggregation as the basis for a digital library. Such an argument needs to take into account metrics on how metadata actually improves services such as information retrieval (in the manner of the seminal Cranfield experiments [129]), and contrast the costs and benefits of metadata aggregation against other approaches. What we do provide is quantitative and anecdotal data on the operational costs of a metadata-based digital library. Both costs and benefits need to be accounted for in a final evaluation of the architecture.

The organization of this paper reflects the stages in the flow of metadata through the NSDL architecture. It examines each stage, from original provision of metadata to the

use of the metadata by services, exemplified by the search service. The description of each phase describes impediments encountered and the success and/or failure of various tools to overcome those impediments. We purposely omit a discussion of user interface portals, since evaluation of user interfaces is by nature different from the system issues that are the focus here.

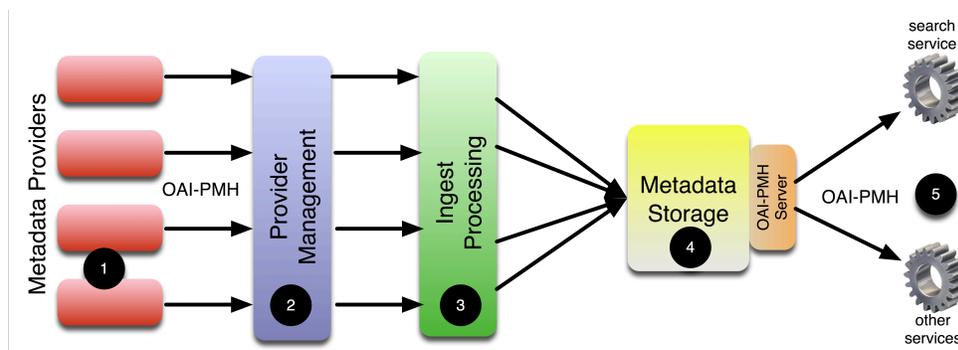


Figure 43 - NSDL metadata flow

The metadata flow is shown in Figure 43. The components of this flow, which are labeled with circled numbers in the figure, and which correspond to the order of the remaining sections in this paper, are as follows.

- *Metadata Providers* – These are the organizations from which CI harvests metadata via OAI-PMH. In some cases these metadata providers also manage the content described by the metadata; in others, they either exclusively or additionally aggregate metadata about resources managed by other organizations. The NSDL architecture does not distinguish among these roles; everyone is treated as a metadata provider. We describe problems that providers have encountered with metadata and OAI-PMH, and some tools and techniques that simplified that process.
- *Provider Management* – CI has developed a software component known as the Collection Registration Service (CRS) that maintains knowledge of all data

providers, descriptions of their collections, their OAI servers and harvest information, harvest schedules, and logs of harvests performed. The intent of this system is to automate, as much as possible, periodic harvests from a large number (potentially hundreds) of OAI-PMH data providers. We will describe issues that arose that have interfered with this automation.

- *Ingest Processing* – CI developed back-end services to process the raw OAI-PMH feeds and normalize the metadata before storage in the RDBMS metadata repository. We describe these processes and their efficacy in automating the OAI feeds and improving the metadata.
- *Metadata Storage and OAI Re-Exposure* –We describe some aspects of table design of the Oracle-based metadata repository, especially related to the exposure of the metadata via OAI-PMH¹¹⁵, allowing CI to act as an “OAI-PMH aggregator” [313], effectively combining the roles of data provider and service provider. We describe our experiences running a relatively large-scale (1.2 million metadata records) OAI-PMH server and our techniques for linking that server to an RDBMS.
- *Search Service*¹¹⁶ – CI runs a search service that uses Lucene¹¹⁷ for indexing and query processing. Lucene indexes both the metadata, consumed from the MR via OAI-PMH, and if possible the full-text resources, crawled via Nutch¹¹⁸ using resource URLs provided in metadata records (if present and accessible). By and large, users of the NSDL (and most libraries) are concerned with

¹¹⁵ The baseURL of the NSDL OAI server is <http://services.nsdl.org:8080/nsdloai/OAI>.

¹¹⁶ Although there are other services in the NSDL, such as an archive service, we will not describe them in this paper.

¹¹⁷ <http://lucene.apache.org/>.

¹¹⁸ <http://lucene.apache.org/nutch/>.

finding and accessing resources. As such, the search service (and many other services) needs to translate the metadata-centric data model (where metadata originates from both content holders and metadata aggregators) to a resource-centric view. We will describe issues related to presenting a resource-centric view of the library over a metadata-centric architecture.

The paper concludes with some broader comments on the overall utility of this digital library architecture. Our recent work in the NSDL and other projects [298, 299] focuses on a resource-centric architecture that integrates less structured forms of information, which collectively add value and context to digital library resources. Traditional structured metadata plays a role in such information contextualization. However, it exists as a component of a resource-centric model, rather than being the focus of the information model itself.

Related work

The architecture of the NSDL and the issues of metadata creation, harvesting, and aggregation have been described in earlier papers by the CI team. The initial prototype of the architecture was described in [32]. The current NSDL production architecture was introduced in [287]. Some of the processes described in this chapter and related issues with metadata aggregation in the NSDL were described in [31]. This paper logically follows after those papers, providing an overview of the costs, problems, and experiences in supporting the metadata aggregation model over the past three years. It also is written at a time when the CI team is engaged in a major project to shift the architecture to a different, resource-centric, paradigm [298]. As such, it provides the opportunity to look back on the initial architecture from the perspective of lessons learned.

The issue of metadata quality is an important factor in the system described here. Even if all other aspects of the system worked perfectly, poor quality metadata would degrade the quality of the resulting library. Diane Hillmann, who was instrumental in the deployment of the NSDL, has written extensively on this issue [100, 235]. With Naomi Dushay, she has written about visualization tools for analyzing the quality of metadata [172]. Other papers focus on the quality of metadata harvested and federated from distributed sources [131, 435]. This paper does not cover metadata quality per se, but touches on it as one of the system design issues, complexities, and costs in maintaining a relatively large-scale metadata aggregation site.

Finally, OAI-PMH, upon which the NSDL system is built, is a de facto standard for metadata sharing about which much has been written. The OAIster system [224] is another example of a large-scale aggregation system. [225] reports findings on a number of metadata harvesting experiments. There has been some research related to normalizing and enhancing large-scale harvests. [190] describes the use of harvested collection metadata records to enhance harvested item records. [230] provides preliminary findings on eliminating duplicates in harvested OAI-PMH records. This paper briefly touches on these issues, but does not focus on them.

Metadata providers

According to the NSDL Collection Development Policy , the “NSDL Collection is a *collection of sets of resources*. These sets of resources are also referred to as *collections*.” Furthermore: “As a general rule, collections that are considered to be part of the NSDL Collection are not actually held within NSDL-owned computers or storage systems. Instead, individual collections typically are held and managed by their owners or providers.”

In lieu of storing the resources that make up the NSDL collection, the decision was made to develop and manage a repository of metadata surrogates for these resources. Intentionally operating without a cataloging staff, CI assumed that metadata records would be contributed by external parties, both those that wanted to contribute their content to the NSDL collection, and those that had metadata about other organizations' resources.

The practice of collecting resource surrogates from distributed parties and cataloging them is well established in the library community. OCLC's WorldCat¹¹⁹ collects and distributes many library catalog records. Our plan was to adapt this model with Dublin Core as a minimalist metadata format, which could be supplemented by richer metadata formats, and OAI-PMH as a low-barrier transport technology. Our expectation was that Dublin Core and OAI-PMH were relatively simple and that surely every collection provider would be able to implement them and be integrated into the NSDL.

In fact, reality fell far short of our expectations. We discovered that the WorldCat paradigm, which works so well in the shared professional culture of the library, is less effective in a context of widely varied commitment and expertise. Few collections were willing or able to allocate sufficient human resources to provide quality metadata. A mandate from the NSF in 2004 that NSF-funded NSDL collections had to share metadata addressed some of the "willingness" problems of those collections. Unfortunately, commercial providers of STEM resources were especially resistant to sharing their metadata; they had yet to learn (e.g. from Google) that open access to discovery information leads to more use (i.e. sales).

¹¹⁹ <http://www.oclc.org/worldcat/>

But more problematic was the reality that the personnel requirements to share metadata were deceptively high due to what can be characterized as a “knowledge gap”. Successful provision of metadata actually involves three distinct skill sets:

- *Domain expertise* – knowledge of the resources themselves and their pedagogical goal.
- *Metadata expertise* – knowledge of cataloging practices such as use of controlled vocabularies and proper formatting of data such as names and dates.
- *Technical expertise* – knowledge of tools involved in setting up and running an OAI-PMH server including XML, XML schema, UTF8, and HTTP.

We found that very few NSDL collections had a single person, let alone a team, with these three skill sets. In fact, the “team” for many collections consisted of one person working part-time. Thus, the CI team, which indeed had the combined expertise, had to provide intensive consultation. Documentation on Dublin Core [173] and OAI-PMH helped somewhat, but still the amount of hand-holding was well beyond what was anticipated. An analysis of our collection development email logs indicates that for a large number of collections the time lag between first contact and successful provision of metadata *exceeded several months*, and in one exceptional case spanned two years! Throughout this interim, the CI team had to engage in frequent training and persuasion to move metadata providers into the production cycle.

Some of the technical barriers were overcome by funding from NSF for the development and deployment of the Collection Workflow Integration System¹²⁰ (CWIS) “... software to assemble, organize, and share collections of data about resources, like Yahoo! or Google Directory but conforming to international and

¹²⁰ <http://scout.wisc.edu/Projects/CWIS/>

academic standards for metadata.” The CWIS software comes complete with an OAI-PMH server, so that metadata stored within a CWIS installation could be readily ingested into the MR (or any other OAI-PMH aggregator). CWIS has proven effective for some collections and has been deployed on a relatively modest basis. At last count, sixteen NSDL collections, out of the approximately 85 OAI servers, are running CWIS.

Obviously massively scaled web search engines such as Google and Yahoo do not incur either the resistance or costs of metadata provision and harvesting. Although there are limits to automated crawling and indexing – e.g., deep web invisibility and indexing non-textual resources¹²¹ - we recognize that the future of collection development in the NSDL relies on deploying these technologies as a supplement and, in many cases, a replacement for the harvesting model. We are currently working with the iVia project [374] at the University of California-Riverside, which has developed technology for focused crawling, automated metadata generation, and “rich-text” generation (intended for resource discovery). CI has started to use this tool for collection building.

Provider management

From a technical perspective, an NSDL *collection* is an entity from which metadata is harvested via OAI-PMH. The Collection Registration Service (CRS) provides a set of services for identifying and managing these collections and for managing the processes that harvest metadata from them. The CRS accomplishes this by

¹²¹ We note that these are not insurmountable limitations and future considerations about metadata and metadata harvesting must consider rapid technical advancements in these areas.

maintaining both Dublin Core metadata about the collection itself and the information needed to automatically harvest OAI metadata from the collection's OAI provider.

In this section we describe the design of the automated harvesting system, enumerate some problems with automation, and then describe some statistics related to our harvesting experience.

Automated harvesting model

In the original model, harvesting of metadata was intended to be almost completely automated, with the following workflow:

- New collections validate their OAI-PMH server¹²².
- A metadata record describing the collection is created and stored in the CRS, which then ports it to the MR.
- A metadata harvesting record for the collection is created that lists the OAI source, OAI set and format information, provider emails, and a harvest schedule. This record is the basis for automated harvesting
- An initial full harvest of the collection is initiated.
- Subsequently, incremental harvests happen on a schedule appropriate to the collection (e.g. weekly, monthly, quarterly), with automatic emails to the provider describing any problems encountered, allowing the provider to correct the problem and schedule an updated harvest.

¹²² The NSDL wrote its own OAI validator (publicly available at http://repository.comm.nsd.org/prs_web/harvest_server_val.php), which provides stringent checks to facilitate automated harvesting using the same code used for validation at ingest.

Automation problems

In a few cases this workflow proceeds smoothly, but the vast majority of cases require significant manual intervention. A detailed enumeration of these problems is impossible due to the constraints of this chapter, but we highlight the following.

The process of initially validating a new OAI provider is extensive, typically requiring several email exchanges and repeated harvest attempts. Validation errors run the gamut, including UTF-8 errors, XML schema validation problems, URL and XML encoding problems, improper date stamping, bad resumption tokens, and the like. We provide more details on validation statistics in [476]. Compared to other protocols OAI-PMH may be “low barrier”, but deploying it requires reasonable technical sophistication with protocols, XML, schema, and the like.

Providers often fail to use available OAI validation tools, and rarely perform routine self-validation. This places the burden on harvesters like CI to notify providers of problems.

Often validation of an OAI server will fail over time. Because OAI-PMH responses are structured as a set of packages (e.g., “about” containers, metadata) that are variable across OAI-PMH transactions, validation may break down as the content of a package varies, or due to web server upgrades or other software changes.

The notion of *incremental harvesting* is a fundamental part of the OAI-PMH model. Theoretically, a data aggregator should only need to do one initial full harvest, followed by repeated harvests that include modifications, deletions, and additions to the metadata from the data provider. In practice, incremental harvesting is often not possible due to two main problems:

- First, support for “deleted” records is inconsistent. As documented in the OAI Registry at UIUC¹²³, less than 50% of OAI-PMH data providers purport to persist (“forever”, as defined in the OAI-PMH specification) deleted records. We have found, moreover, that some data providers that claim “persist” actually have less stringent perspectives on persistence and that a complete harvest is often the only reliable way to get an accurate snapshot of a data provider. As described by [101], detecting changes and deletions across distributed digital library collections is generally problematic.
- Second, when OAI servers fail on any record during an incremental harvest, the start date cannot be updated. Similarly, any server instability can cause problems in determining an appropriate start date. The result is that a full harvest needs to be performed to “re-sync” the repository’s view with that of the data provider.

As a result of these problems, initial harvest setup and regular harvests require constant monitoring. Emailed harvest results are sent to the CI harvest production team, who interpret them and contact the providers as necessary to correct OAI server protocol, XML, schema, and other errors. Weekly production meetings of the ingest team, together with a careful process of tracking harvest results and provider email exchanges, keep things relatively smooth, but the ongoing people cost is significant despite all efforts to automate.

¹²³ <http://gita.grainger.uiuc.edu/registry/>

Harvesting statistics

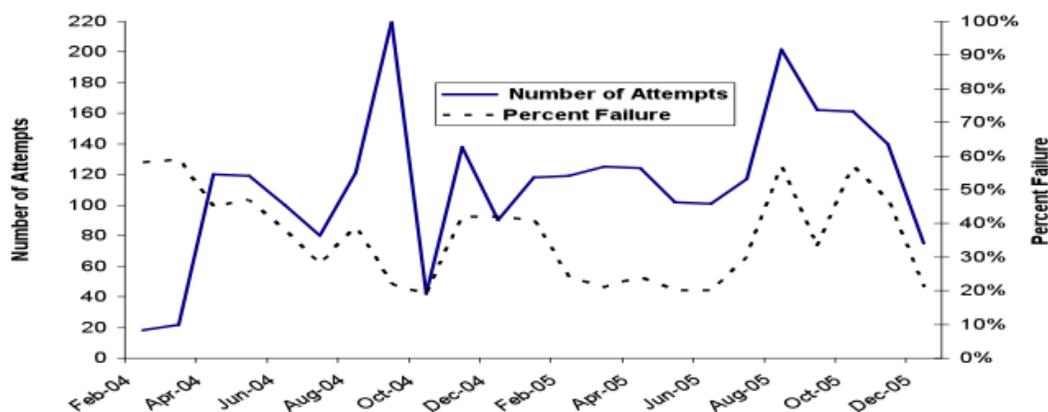


Figure 44 - NSDL harvesting failure rate

NSDL routinely harvests metadata from 113 collections via OAI. The harvesting discovers an average of 9250 items per collection. Each collection is re-harvested on an interval of between 1 to 3 months depending on the needs of the collection. Over the past two years NSDL has made over 2,600 harvest attempts.

We should note that not all collections run their own OAI service. Of the 114 collections, 37 are harvested from 8 OAI servers. This has resulted in economies for some collections. Additionally, many of the servers are based on shared code such as OCLC's OAICat¹²⁴ or Scout Portal Toolkit¹²⁵.

Our overall harvest success rate for the years 2004-2005 is 64%. On a monthly basis our harvest failure rate has stubbornly hovered between 25-50%. This is illustrated in Figure 44. Periodically, major efforts have been made to reduce these failures (Sept

¹²⁴ <http://www.oclc.org/research/software/oai/cat.htm>

¹²⁵ <http://scout.wisc.edu/Projects/SPT/>

2004, Aug. 2005). While these efforts have pulled a great deal of new content into the repository, they did not succeed in lowering the failure rate over the long term.

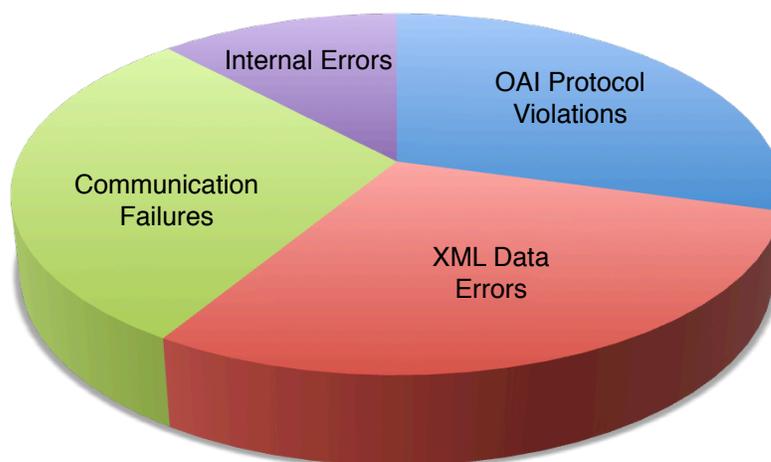


Figure 45 - Harvest failure categories

The reasons for individual failures vary over time. Harvest failure categories are illustrated in Figure 45. Summarizing, our actual experience has shown that failures present themselves in equal measure within 3 broad categories: (i) a communications or system failure either at their server or with our OAI harvester, (ii) OAI protocol violations, and (iii) invalid XML data, XML schema non-compliance, or XML, URL, or UTF-8 character encoding. In fact, many of the OAI protocol violations are the result of these sorts of format errors, which result in an inability to process or even complete the OAI harvest.

Even the best maintained collections have difficulties at one point or another in their life cycle. Network and host availability issues are expected to impact harvesting, yet only 23% of harvest failures are due to such transient failures. Most harvest failures are due to data and protocol problems that require intervention by the metadata

providers. In these cases, the specific causes of the failure must be thoroughly diagnosed, and the provider personnel contacted with actions needed to bring their OAI metadata back into compliance. Generally, harvesting cannot resume until the provider rectifies the problem(s). Note that CI staff attempt to find all co-occurring errors before contacting providers. While this is time consuming up front, it prevents repeated dialog on related errors.

Email is the primary method of interaction with all providers. While some of this email is essentially templated machine-interpretable communication, much of it is human correspondence. A cross section of email archives of 8 representative providers revealed over 2,700 messages, or around 170 messages per provider per year.

The subjects of these emails are indicative of the difficulties in automating ingests from these feeds. 25% of all messages analyzed were automated reports of harvest failures. 39% were human or diagnostic messages, usually in response to failures. The remaining 36% were routine messages of successful harvests. The difficulty of setting up an OAI server and establishing harvesting is also apparent. On average, 98 messages were transmitted before NSDL was able to successfully retrieve its first harvest of a collection. In some cases, there were hundreds of email messages exchanged before a successful harvest occurred. Each of these messages corresponds to considerable human effort to resolve the problem.

Ingest processing

The goal of ingest processing is to transform the raw OAI feeds from metadata providers into metadata ready for storage in the Oracle-based metadata repository. These transforms address some metadata quality issues. Following the transforms, the metadata is staged in an XML file we call `dbInsert`. This is a list of metadata records

similar to an OAI-PMH `ListRecords` response. The major difference is that a single metadata “record” in the `dbInsert` file contains both the originally harvested record (which remains unchanged) and the newly created normalized `nsdl_dc` generated by the transform process.

Whereas the plan has always been to ingest multiple rich metadata formats, the harvest-ingest process currently processes only two formats:

- `oai_dc`: the required OAI-PMH schema for unqualified Dublin Core.
- `nsdl_dc`: an NSDL-specific application profile for qualified Dublin Core that includes extensions relevant for educational materials, as recommended by the DC Education Working Group [364].

There are two reasons for the delay in ingesting additional richer metadata formats:

First, as mentioned already, we have experienced considerable difficulty working with collections as they implement the minimal harvesting scenario: running an OAI-PMH server that provides the required `oai_dc` format. Many collections simply do not have the resources to take the next step and provide richer metadata after that initial implementation. `nsdl_dc` is considerably more expressive than `oai_dc`, yet only 50% of the collections provide metadata in that format. Only about 10% provide metadata in any of the other NSDL-supported metadata formats, providing little justification for CI to expend the effort necessary to process these formats.

Second, metadata quality, even with this minimal format, remains vexing. Our experience in improving the quality of DC records has been mixed. As described in [235], an initial approach involved “collection-specific” transforms, whereby we processed and corrected metadata on a collection-by-collection basis. We found, however, that in practice there was little consistency to the types of problems that

arose within an individual collection’s metadata, and “collection-specific” often evolved to “harvest-specific” corrections. Clearly this was not scalable.

We therefore evolved a more scalable strategy known as “safe transforms”, a process that takes `oai_dc` or `nsdl_dc` as input, fixes some common errors, applies some simple refinement techniques, and generates `nsdl_dc` as output. These transforms include:

- removing metadata fields with no information value (e.g., “no abstract submitted”),
- removing extraneous white-space,
- removing duplicate elements,
- qualifying easily recognized encoding schemes (e.g., URIs, well-known DCMI types, normalizing `dc:language` values), and
- correctly specifying and encoding URIs.

The last transform deserves additional explanation. As we describe later, we need to “harden” the link between the metadata and the actual resource, so we can use that link for additional indexing. Reliable metadata->resource links are also useful for a number of other services. Thus, the MR ingest process “smartens up” `dc:identifier` fields: those that start `http://` or `ftp://` are designated with the DC URI encoding scheme if they can be automatically scrubbed into fetchable URLs. The ingest process also does an XML schema validation (via Xerces) and some additional validation on `dc:identifier.dct:URI` fields provided in Qualified Dublin Core by a collection. Those that fail these validation steps are downgraded to plain `dc:identifier` fields.

Some of these transforms, such as URI corrections, apply across formats, but many are specific to a metadata format, such as specific DCMI encodings and types. As a result, the introduction of each new metadata format requires expensive analysis of

common problems and potential fixes in order to extend the “safe transform” philosophy. The scalability of this is questionable.

In the end, all of these transforms don’t enhance the richness of the information in the metadata. Minimally descriptive metadata, like Dublin Core, is still minimally descriptive even after multiple quality repairs. We suggest that the time spent on such format-specific transforms might be better spent on analysis of the resource itself – the source of all manner of rich information.

Metadata storage and OAI exposure

The MR is implemented as an Oracle database. We chose an ‘enterprise level’ data store to allow rapid deployment of a repository capable of handling a very large number of metadata records. Also, local expertise made Oracle an attractive choice.

As a part of the redundancy and backup plan, two file servers are used to house the database. Metadata is processed through these two servers running separate Oracle instances. The metadata is inserted in and the XML metadata records are generated on one system, and the ready-to-expose records and supporting index tables are transferred to a separate system that feeds the OAI-PMH service. This separation of ingest, XML generation, and exposure has allowed for flexibility in configuration and backup of the source data and the served XML records.

Along with various logging and administrative data tables, the MR database schema contains three sets of tables: a set of five tables for storing data as it is parsed on input into the system, a set of four tables that contain the generated XML formats that are used for OAI-PMH serving, and a set of seven tables that contain the combined OAI-formatted data and index tables optimized for retrieval by our java-based OAI server.

Data flow through the MR database

An initial (SAX) parse of the inbound `dbInsert` XML metadata records separates them into the two sets of records that the safe-transform process creates – the normalized `nsdl_dc` records created by the safe-transform, and the original records harvested from the OAI provider. These original records are then stored as a single string with their own originator date stamp and schema identification.

The `nsdl_dc` records generated by the normalization process are shredded into element-value pairs and stored. Element names and their source `nsdl_dc` schema and schemes are coded and identified in reference tables. This structure was chosen to facilitate analysis and modifications of specific elements within the normalized `nsdl_dc` records across all metadata records. It is also used to generate a rudimentary resource index by extracting all identifier elements of all metadata records that are URI-like.

As metadata records are inserted, the records for OAI exposure are also generated.

NSDL currently produces five distinct OAI formats from this metadata:

- `nsdl_dc` is the normalized Qualified Dublin Core version of the harvested metadata records.
- `oai_dc` is the simple Dublin Core record required of any OAI-PMH data provider. This is a dumbed-down version of the normalized `nsdl_dc`.
- `nsdl_links` indicates relationships between metadata records. Currently the only relationship represented is collection membership – each record is a member of a collection, which is represented by a metadata record for the collection.
- `nsdl_search` is a combined format that includes the above three formats as well as the original ‘native’ harvested format. This format is not released to

the public, as the provider of the native format may not wish to share their metadata, and it is currently used only for the NSDL's search-index process.

- `nsdl_all` is the same as `nsdl_search` except that the 'native' metadata record will not be present if the metadata provider has requested that their metadata not be made publicly available.

All of these five served formats are generated as large strings and stored in the staging tables on the ingest server. A timed process runs on the serving database that queries the staging database for new entries. New entries are gathered and the tables required to serve OAI-PMH are populated with the new or updated entries. The serving Oracle instance contains views and some level of de-normalization of table data in order to optimize the queries that the java OAI server uses to service requests.

Lessons from the MR implementation

Oracle has proven to be a flexible data storage tool, but the cost for configuration and operations has been high. Configuring and tuning the database to perform optimally has taken considerable time and effort, and the on-going management of the database has required more-than-expected personnel resources as well. The ingest-staging database contains about 55GB of data, and the current OAI serving database contains approximately 53GB of data.

Because OAI records have timestamps that are used in incremental harvests, it is crucial that the timestamp associated with an OAI record be calculated appropriately. This required some rather arcane processing. OAI-PMH idempotency requirements mandate that a request for records between the dates D and $D+\Delta$ will always return the same records, if they have not been updated in the meantime. Since the record timestamps must be generated significantly before we expose the OAI record in our tables, to meet the OAI-PMH idempotency requirements we must post-date all

inbound metadata records by three hours. If we didn't post-date into the future, then harvest requests for very recent metadata could potentially be missing any OAI records that had not yet been generated for exposure.

Throughput for processing harvested records in the current production environment runs between 5000-10000 records per hour. This depends greatly on the transfer rate from the originating OAI server, the number of records to process, and the density of those records.

Early in the life of the repository, some errors in content propagated through to the data store. As our error detection and correction efforts have improved, most of these errors have been corrected, but some, particularly from collections that are no longer available, are still in the system. The people cost of correcting these errors is too high, so we continue to serve a small percentage of OAI records with XML schema errors.

Our current ingest process, fairly robust after two major rewrites and numerous bug fixes, is still vulnerable to occasional UTF8 encoding and XML Schema validation errors creeping into newly stored records. These errors often go unobserved for weeks until some downstream user or service stumbles on them.

Functionality of the RDBMS-based MR

Overall, the Oracle RDBMS has been successful as a tool for metadata storage, meeting the original requirements. However, as the NSDL has matured, the requirements have grown. We note two areas where the RDBMS has been problematic in extending the functionality of the NSDL.

We increasingly find that storing and querying an expanding set of relationships among library entities – resources, metadata, annotations, standards, and providers – is essential. Handling queries such as “find all the resources contributed by DLESE that

meet the California middle school standard for earth science” is critical for building the types of customized applications of the NSDL that we envision. While the RDBMS design contains a “links” (relationships) table, it lacks the expressiveness of ontology-based relationships. Furthermore, composing transitive queries across entity-relationship graphs is cumbersome and may encounter expensive blow-ups in the number of joins.

The table design of the MR is based on the notion of structured data – metadata elements and values. However, following the initial release, CI has tried to move to less structured forms of data and, in fact, into the creation and storage of content itself – e.g., lesson plans, curricula, annotations, etc. The MR-based architecture, which imposes a strict bifurcation between “metadata” and “data”, has interfered with the effort to create a unified data repository that can flexibly accommodate a range of structured, un-structured, and semi-structured data.

Search

The NSDL Search Service is, essentially, the first customer of MR records exposed via OAI, and the information in the NSDL search index determines whether a resource can be discovered via searching at the main NSDL portal. In fact, many collections check to see if their metadata has been integrated into the NSDL by doing “known item” searches at nsdl.org. Thus, the search service is sometimes used to discover ingest errors such as missing or incorrect metadata.

The current production search service is based on the metadata aggregation model, and it is sometimes referred to as “metadata-centric.” As the limitations of this model have come to light, and additionally, as nsdl.org users have complained about finding duplicates in their search results, we have moved to create a “resource-centric” search service, both to avoid duplicates in search results and to position us to include richer

information, such as context and less structured metadata, in determining nsdl.org search results.

Metadata-centric search

The metadata-centric search service starts with an OAI harvest from the MR. The XML metadata received is parsed, then indexed using Lucene, an open source search engine. The index contains the normalized `nsdl_dc` metadata, the “raw native” metadata, and some additional information, such as NSDL collection membership. Each metadata record becomes a document in the Lucene index – a document roughly equates to a “hit” in search results. Thus metadata-centric search results contain a “hit” for each relevant metadata record in the index. The indexed metadata is updated incrementally – only records modified since the previous harvest are requested from the MR OAI server, and the results are used to update the existing Lucene index.

We also fetch and index the textual content of resources described by the metadata. The search service looks in the `dc:identifier.dct:URI` fields exposed in the normalized `nsdl_dc` from the MR for URLs we can fetch. Then the search service uses Nutch, an open source web crawler, to fetch the content and manage it. (Currently the Nutch software comes with code to retrieve content via http and ftp). Nutch stores the URLs and the fetched content (both as received and as text ready to be indexed) for efficient storage and access and also provides a mechanism to refresh stale content automatically. As of January 25, 2006, the production search index contained 1,056,407 Lucene documents, representing all the “active” metadata records from the MR. (Note that this number does not reflect approximately 280,000 MR OAI metadata records marked deleted.) Approximately 7500 of these Lucene documents have no URL resource identifier, meaning there was no resource URL that passed our validation.

Resource-centric search

We have a number of reasons for moving from a data model that is metadata-centric to one that is resource-centric. For example, we receive metadata records from a large number of providers, and some of those are about the same resource. Rather than a simple metadata repository that stores these as separate records, they should be related to a common resource “entity”, which is currently not represented in the MR data model. In the future, we also want to express the relationships between resources and other information, such as annotations and standards alignments. Finally, we wish to inter-relate resources themselves, such as their co-existence within a lesson plan or curriculum. That resource-centric model is the subject of current work on an NSDL Data Repository (NDR), which will replace the MR [298].

Independent of that work, we have been transitioning to a more resource-centric search service, currently using the metadata repository, but later the NDR. Whereas the current search engine has a one-to-one mapping from metadata record to “hit”, this work will map hits to resources – independent of the number of metadata records about that resource.

In order to do this, we need to infer resource equivalence from the MR, which sits at the end of a data flow that up to this point is entirely metadata focused. We determine equivalence by exploiting the identifiers in the `nsdl_dc` records that we harvest from the MR.

We should note, however, that the URL in an item record does not automatically correspond to an actual link to the real digital resource described by the metadata. We have found that some metadata providers shortcut the effort to actually insert a unique item URL in the DC record by using the same collection “splash page” URL for a set of item records. This indicates that the methods we describe below for

determining resource equivalence need to also account for “fuzzy equivalence” between metadata records – i.e., whether two records that purport to describe the same resource (measured by URL equivalence) are really “about” the same content [230].

At this point, however, we are taking two approaches to determining equivalence using the URLs in the metadata records.

Resource equivalence phase I: URL normalization

The URI specification [53] enumerates steps to normalize URLs to determine if they are equivalent. This includes ensuring the scheme and hostname are lower case, the default port is not specified, an empty absolute path is represented as a trailing slash, and so on. The search service addresses most of this URL normalization with `java.net.URI` methods; the remaining pieces are addressed with additional java code.

Initially we took a naïve approach that created a “resource” (a Lucene document) for each `dc:identifier` and `dc:identifier.dct:URI` in the OAI metadata. However, this naïve approach had the undesirable effect of increasing the number of documents in the Lucene index by almost 50%: at that time we had slightly more than 1 million documents in the metadata-centric index, and almost 1,500,000 documents in this naïve resource-centric index. In examining the causes, we learned that there are approximately 1,500,000 `dc:identifier` fields (in various flavors) but the number of fetchable URLs is closer to 1 million.

Before choosing a different algorithm and making a similar mistake, we chose to examine our `dc:identifier` fields and our metadata records more carefully. This involved writing some tools to examine Lucene index contents, as well as performing SQL queries against our Oracle database. Because of our decision to split our normalized `nsdl_dc` into elements in the Oracle DB, getting information such as

“what do records with multiple fetchable resource identifiers look like?” and “how many metadata records have 2 or more fetchable resource identifiers” has been difficult and is still in progress. As of January 26, 2006, we count approximately 180,000 metadata records with 2 or more fetchable resource identifiers.

Resource equivalence phase II: comparison of fetched content via MD5Hash

The Nutch application creates an MD5Hash for fetched content to facilitate comparison. In our current work, we will use these checksums as an initial means to determine if fetched content is equivalent. If so, the normalized resource URLs (and matching NDR resource digital objects) will be marked as part of an equivalence class, and the corresponding Lucene documents in the search index will be merged into a single Lucene document for the resource. This phase has not yet been implemented. Furthermore, we recognize that for textual content there are more advanced methods for determining equivalences. [99]

Conclusion

After three years of work, the NSDL CI team has learned that a seemingly modest architecture based on metadata harvesting is surprisingly difficult to manage in a large-scale implementation. The administrative difficulties result from a combination of provider difficulties with OAI-PMH and Dublin Core, the complexities in consistent handling of multiple metadata feeds over a large number of iterations, and the limitations of metadata quality remediation.

More problematic are the shortcomings of the architecture as the basis for a service-rich digital library. As noted in previous sections, the centrality of structured metadata interferes with the intermingling of potentially more valuable unstructured and structured data and the rich relationships among these data entities. Even the

implementation of search, a basic digital library service, is hampered by the need to recover a resource-centric view from a dataflow that is solely metadata-centric.

Arguably, it makes more sense to create an architecture that begins with a resource-centric view (e.g., a set of resource URIs from a web crawl) and carries that view through the entire model. The CI team is now implementing such an architecture, based on the notion of an information network overlay. This architecture emphasizes the integration of multiple information entities and their rich relationships, while focusing on creating and expressing context for resources.

Representing Contextualized Information in the NSDL

Preface

This chapter is based on:

Lagoze, C., Krafft, D., Cornwell, T., Eckstrom, D., Jesuroga, S. and Wilper, C., *Representing Contextualized Information in the NSDL*, in *ECDL2006 (Best paper award)*, (Alicante, Spain, 2006), Springer [297].

As described in Chapter 10, the initial architecture of NSDL was modeled upon the notion of the traditional library union catalog. In that architecture, metadata was harvested from content repositories via OAI-PMH, processed for normalization, indexed in a central search engine, and subsequently used as the basis for searching through a portal interface. The technical problems with this architecture, in particular its management costs and overhead, are described in detail in Chapter 10.

This chapter describes work motivated in part by the aforementioned operational problems with the metadata-centric architecture. However, a more significant motivation for the work was the recognition that the traditional “search and access” digital library paradigm embodied in the early NSDL implementation was insufficient relative to the specific aims of the NSDL. Apropos of this, the text notes that:

... in order to provide an educationally-focused digital library, the information infrastructure must support flexible integration of information, ranging from highly structured metadata to unstructured comments and observations. It needs to be dynamic, expanding both in the manner that the standard library collection expands, but also based on the collective experience and input of the user community.

To meet these goals of being dynamic and incorporating collective experience the paper describes a new architecture for the NSDL that “...combines traditional digital

library notions of resources and structured metadata with service-oriented architecture and semantic web technology...” Furthermore, it incorporates Web 2.0 concepts to encourage a “culture of participation” and provide an interface “to its accumulated information for innovative mashups that exploit library information in innovative ways”. The basis for this new architecture is Fedora, which is described in Chapter 9.

These observations about the need for a richer digital library model specific to educationally-focused applications fit into a number of broader observations I made about digital libraries in general in a companion paper titled “What Is a Digital Library Anymore, Anyway?” [299]. In this other paper, I stated:

... free of the constraints of physical space and media, digital libraries can be more adaptive and reflective of the communities they serve. They should be collaborative, allowing users to contribute knowledge to the library, either actively through annotations, reviews, and the like, or passively through their patterns of resource use. In addition, they should be contextual, expressing the expanding web of inter-relationships and layers of knowledge that extend among selected primary resources. In this manner, the core of the digital library should be an evolving information base, weaving together professional selection and the “wisdom of crowds”.

The contrast between this thinking and that included in earlier papers that form the basis of Chapter 7 and Chapter 8 is indeed quite striking and is representative of a fundamental shift in thinking about the nature of digital information. The web had come a long way from the network of hypertext documents that existed at the time that Dienst was implemented and the Dublin Core vocabulary was formulated. Clearly, digital libraries needed to leap beyond their search and access routes into the new read/write information paradigm of Web 2.0.

Acknowledgements

The work described in the remainder of this chapter was supported by the National Science Foundation under grants 0227648, 0227656, and 0227888. Support for the

Fedora work in collaboration with NSDL came from the Andrew W. Mellon Foundation. The entire NSDL CI team made contributions to this work. Contributors are of note were Dean Krafft, Susan Jesuroga, Tim Cornwell, Dean Eckstrom, and Chris Wilber. Finally, Sandy Payette deserves particular recognition for her formulation of the notion of an “information network overlay” and for her leadership and technical contributions to Fedora, which made all this work possible.

Introduction

Libraries, traditional and digital, are by nature information rich environments - the organization, selection, and preservation of information are their *raison d'être*. In pursuit of this purpose, libraries have focused on two areas: building a collection of all the *resources* that meet the library's selection criteria, and building a catalog of *metadata* that facilitates organization and discovery of those resources.

This is the approach that the NSDL (National Science Digital Library) Project took over its first three years of existence, when it focused mainly on the location and development of resources appropriate for Science, Technology, Engineering, and Mathematics education, and the creation of quality metadata about those resources. This focus was reflected in the technical infrastructure that harvested metadata from distributed providers, processed and stored that metadata, and made it available to digital library services such as search and preservation.

The value of an excellent collection of resources as a basis for library quality is undeniable. And, even after years of advances in automatic indexing, metadata remains important for a class of resources and applications. However, our three years of effort in the NSDL have revealed that collection building and metadata aggregation

are necessary but not sufficient activities for building an information-rich digital library. In particular, our experience has led to two conclusions. First, the technical and organizational infrastructure to support harvesting, aggregation, and refinement of metadata is surprisingly human-intensive and expensive [296]. Second, in a world of increasingly powerful and ubiquitous search engines, digital libraries must distinguish themselves by providing more than simple search and access [299]. This is particularly true in educationally-focused digital libraries where research shows the importance of interaction with information rather than simple intake.

Based on these conclusions, we have redirected our efforts over the past year towards building a technical infrastructure that supports a more refined definition of information richness. This definition includes, of course, collection size and integrity, and it accommodates the relevance of structured metadata. But it adds the notion of building *information context* around digital library resources. Our goal is to create a knowledge environment that supports aggregation of multiple types of structured and unstructured information related to library resources, the instantiation of multiple relationships among digital library resources, and participation of users in the creation of this context. We are creating an infrastructure that captures the wisdom of users [450], adding information from their usage patterns and collective experience to the formal resources and structured metadata we already collect.

Our technical infrastructure is based on the notion of an *information network overlay* [298] – a directed, typed graph that combines local and distributed information resources, web services, and their semantic relationships. We have implemented this infrastructure using Fedora [301], an architecture for representing complex objects and their relationships.

In this paper we describe the motivations for this architecture, present the information model that underlies it, and provide results from one year of implementation. We note for the reader that this is still a work in progress. The results we provide in this chapter relate to the implementation and scaling issues in creating a rich information model.

The organization of this paper is as follows. The initial section describes related work and situates this work in the context of other digital library efforts. Following that we describe the importance of information contextualization for educational digital libraries. The next section provides a brief background on the NSDL and establishes the application context in which this work occurs. The subsequent section describes the information model of the information network overlay. We then describe the results of our implementation experience, and close with a concluding section.

Related work

The work described in this chapter builds on a number of earlier and ongoing research and implementation projects that investigate the role of user annotations in information environments, the importance of inter-resource relationships, and the integration of web services with digital content. We believe that our work is distinguished from these other projects in two ways. First, it combines traditional digital library notions of resources and structured metadata with service-oriented architecture and semantic web technology, thereby representing the rich relationships among a variety of structured, unstructured, and semi-structured information. Second, it implements this rich information environment at relatively large scale (millions of resources), exercising a number of state-of-the-art technologies beyond their previous deployments.

Perhaps the most closely related work is the body of research on information annotation. Catherine Marshall has written extensively on this subject [361] in the digital library and hypertext context. A number of systems have been developed that implement annotation in digital libraries. For example, Roscheisen, Mogensen, and Winograd created a system call ComMenter [423] that allowed sharing of unstructured comments about on-line resources. The multi-valent document work at Berkeley provides the interface and infrastructure for arbitrary markup and annotation of digital documents, and storage and sharing of that markup [489]. The semantic web community has also examined annotation, with the Annotea project [254] being the most notable example.

The importance of annotation capabilities for education and scholarly digital libraries has been noted by many researchers including Wolfe [495]. The ScholOnto project [376] created a system for the publication and discussion/annotation of scholarly papers, arguing for the importance of informal information along-side established resources. Constantopoulus, et al. [137] examine the semantics of annotations in the SCHOLNET project, a EU-funded project to build a rich digital library environment supporting scholarship. Within the NSDL effort, there have been a number of projects that support annotations, most notably DLESE¹²⁶ (Digital Library for Earth System Education).

Annotations and their association with primary resources are one class of the variety of relationships that can be established among digital content. Ever since Vannevar Bush invented hypertext [105], researchers have been examining tools for inter-linking information. Faaborg and Lagoze [184] examined the notion of semantic browsing

¹²⁶ <http://www.dlese.org/>

whereby users could establish personalized and sharable semantic relationships among existing web pages. Huynh, et al. [243] have recently done similar work in the Simile project.

There is also related work on resource linking specifically for pedagogic purposes within the educational research community. Unmil, et al. [459] describe Walden's Paths, a project that allows teachers to establish meta-structure over the web graph for creation of lesson plans and other learning materials. Recker, et al. have created another system called Instructional Architect¹²⁷ that similarly allows integration of on-line resources by teachers into educational units.

Finally, an important component of the work described here is the integration of content and web services. In many ways our digital library "philosophy" resembles that of the Web 2.0 philosophy [390]. Key components of this are the collection and integration of unique data, the participation of users in that data collection and formulation process, and the availability of the data environment as a web service that can be leveraged by value-add providers. Chad and Miller [113] extend Web 2.0 to something they call Library 2.0. We hope that our work demonstrates many of the principles they describe, notably the notion that Library 2.0 encourages a "culture of participation" and provides the interface to its accumulated information for innovative "mash-ups" that exploit library information in innovative ways.

The need for context and reuse

Research shows that education-focused digital libraries (and digital libraries in general) need to support the full life cycle of information [362]. Reeves wrote "The real power of media and technology to improve education may only be realized when

¹²⁷ <http://ia.usu.edu>

students actively use them as cognitive tools rather than simply perceive and interact with them as tutors or repositories of information.” [417]

One requirement that appears frequently in the learning technology literature is the reuse of resources for the creation of new learning objects. This involves integrating and relating existing resources into a new learning context. A learning context has many dimensions including social and cultural factors; the learner’s educational system; and the learner’s abilities, preferences and prior knowledge [363].

Most digital libraries, including the NSDL, currently rely on forms of metadata to describe learning objects and enable discovery. Metadata standards abstract properties of learning objects, and abstraction can lead to instances where learning context is ignored or reduced to single dimensions [399]. Metadata is often focused on the technical aspects of description and cataloging, not on capturing the actual context of instructional use. Recker and Wiley write “a learning object is part of a complex web of social relations and values regarding learning and practice. We thus question whether such contextual and fluid notions can be represented and bundled up within one, unchanging metadata record.” [416]

McCalla also argues that there is no way of guaranteeing that metadata captures the breadth and depth of content domains. He writes that, ideally, learning objects need to reflect “appropriateness” to address the differences between learners’ needs [366]. In addition, questions remain as to whether these logical representations (e.g. metadata and vocabularies) created primarily for use by computer systems will make the most intuitive sense for learners [133].

Several approaches have been suggested to help supply the rich context for learning object creation and reuse. These include capturing opinions about learning objects and descriptions of how they are used [399]; recording the community of users from which

the learning object is derived [416]; collecting teacher-created linkages to state education standards [415]; tracking and using student-generated search keywords [14]; and providing access to comments or reviews by other faculty and students [368].

We see, therefore, that in order to provide an educationally-focused digital library, the information infrastructure must support flexible integration of information, ranging from highly structured metadata to unstructured comments and observations. It needs to be dynamic, expanding both in the manner that a standard library collection expands, but also based on the collective experience and input of the user community.

A suite of contextualized NSDL services

We are creating the infrastructure to meet notions of information richness outlined in the previous section. This work follows more than three years of work by the NSDL Core Infrastructure (CI) team, and has been described in a number of other papers [287, 297]. Stated very briefly, this earlier work used OAI-PMH to populate a metadata repository (MR). This metadata was indexed by a CI-managed search service, which was accessible by users through a central portal at <http://nsdl.org>.

Our goal is to move beyond the search and access capabilities provided by the MR. The creation of the NSDL Data Repository (NDR), built on the architecture described in the next section, provides a platform for a number of exciting new NSDL applications focused directly on increasing user participation in the library. In addition to creating specific new capabilities for NSDL users, these applications all create context around resources that aids in discovery, selection and use. Four specific applications that exploit the infrastructure described in this paper are currently in various phases of development, testing, and deployment.

Expert Voices (EV) is a collaborative blogging system that fully integrates the resources and capabilities of the NDR. It allows subject matter experts to create real-time entries on critical STEM issues, while weaving into their presentation direct references to NSDL resources. These blog entries automatically become both resources in the NSDL library and annotations on all the referenced resources. EV supports Question/Answer discussions, resource recommendations and annotations, the provision of structured metadata about existing resources, and the establishment of relationships among existing resources in the NSDL, as well as between blog entries and resources.

On Ramp is a system for the distributed creation, editing, and dissemination of content from multiple users and groups in a variety of formats. Disseminations range from publications like the NSDL Annual Report to educational workshop materials to online presentations like the Homepage Highlights exhibit at NSDL.org's homepage. Resources created and released in OnRamp can become NDR content resources, and NDR resources and other content can be directly incorporated into On Ramp publications, creating new context and relationships within the NDR.

Instructional Architect, described by Recker [414], "... enables users (primarily teachers) to discover, select, and design instruction (e.g., lesson plans, study aids, homework) using online learning resources. ". Currently, IA supports searching the NSDL for resources and incorporating direct references to those resources into an IA project. The IA team is currently working with the NDR group to support both publication of IA projects as new NSDL resources and the direct capture of the web of relationships created by an IA project in the NDR.

The *Content Alignment Tool* (CAT), currently in development by a team led by Anne Diekema and Elizabeth Liddy of Syracuse University, uses machine learning

techniques to support the alignment of NSDL resources to state and national educational standards [160]. Initially (2Q2006), users will be able to use the tool to suggest appropriate educational standards for any resource they are viewing. Later versions of the system will allow experts and other users to provide feedback, incorporated into the NDR, on the appropriateness of the assignments. This tool, and the overall incorporation of educational standards relationships into the NDR, will allow NSDL users to search and browse the NSDL by "standards", starting either from a standard or from any relevant resource.

Design and information model

To provide the foundation for this rich array of user-visible services, we have implemented the NSDL Data Repository (NDR). The NDR implements all features of the pre-existing MR such as metadata harvesting, storage, and dissemination. However, it moves from the restrictive metadata-centric focus of the MR to a resource-centric model, which allows representation of rich relationships and context among digital library resources.

The NDR implements a data abstraction that we call an information network overlay (INO). Like other overlay networks [20] the INO instantiates a layer over another network, in this case the web graph.

Specifically, an INO is a directed graph. Nodes are identified via URIs and are packages of multiple streams of data. This data stream composition corresponds to compound object formats such as METS [345] and DIDL [47], allowing the creation of compound digital objects with multiple representations. The component data streams may be contained data or they may be surrogates (via URLs) to web-accessible content. This allows nodes to aggregate local and distributed content, for example the reuse of multiple primary resources into new learning objects. Web

services may be associated with information units and their components, allowing service-mediated disseminations of the data aggregated in a digital object. This advances the reuse paradigm beyond simple aggregation, allowing, for example, a set of resources written in English to be refactored into a Spanish learning object through mediation by a translation service. Edges represent ontologically-typed relationships among the digital objects. The relationship ontology is extensible in the manner of OWL-based ontologies [155]. This allows the NDR to represent the variety of application-based relations described earlier such as collection membership, aggregation via reuse into a learning object, and correlation with one or more state standards. Nodes (digital objects) are polymorphic - they can have multiple types in the data model, where typing means the set of operations that can be performed on the digital object. In the digital library environment, this flexibility overcomes well-known dilemmas such as the data/metadata distinction, which conflicts with the reality that an individual object can be viewable through both of these type lenses.

The NDR is implemented within a Fedora repository. A complete description of Fedora is out-of-scope for this paper and the reader is directed to the up-to-date explanation at [301]. Each node in the INO corresponds to a Fedora digital object. Fedora provides all the functionality necessary for the INO including compound objects, aggregation of local and distributed content, web service linkages, and expression of semantic relationships. Fedora is implemented as a web service and includes fine-grained access control and a persistent storage layer.

Length constraints on this paper prohibit a full description of the information modeling in the NDR and the use of Fedora to accomplish this modeling. This modeling includes the design of Fedora digital objects to provide the different NDR object types – resources, agents, metadata, aggregations, and the like – and the

relationships among these types for common use cases such as resource and metadata branding and resource annotation.

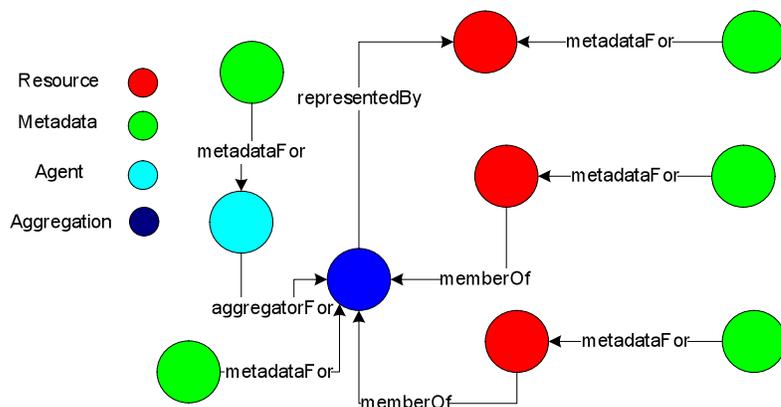


Figure 46 - Modeling an aggregation

The example shown in Figure 46 demonstrates how the NDR represents aggregation. Examples of aggregations include conventional collection/item membership, but also aggregations with other semantics such as membership of individual resources in a compound learning object or alignment of set of resources with a state educational standard. Each node corresponds to a Fedora digital object, with the key at the left showing the type of the object. The labels on the arcs document the type of the relationship. As shown, `memberOf` arcs relate resources to one or more aggregations. Aggregations can have arbitrary semantics, with the semantics documented by the resource that is the object of the `representedBy` arc. For example, this resource may be a surrogate for a collection, or may represent a specific state standard. Lastly, the person or organization responsible for the aggregation is represented by the agent that is the source of the `aggregatorFor` arc.

Results from implementation of the NSDL data repository

For over a year, we have been designing, implementing, and loading data into the NDR. The major implementation task was the creation and coding of an NDR-

specific API for manipulation of information objects in the NDR data model – specific “types” of digital objects such as resources, metadata, agents, and the like and the required relationships among them. Note that this API is distinct from the SOAP and REST API in Fedora that provides access to low-level digital object operations. The NDR API consists of a set of higher-level operations such as `addResource`, `addMetadata`, and `setAggregationMembership`. Each of these higher-level operations is a composition of low-level Fedora primitive operations. For example, the logical NDR operation `addResource`, which adds a new resource to the NDR, translates to a set of low-level Fedora operations including creating the digital object that corresponds to the resource, configuring its datastreams so that they match our model for the resource “type”, and establishing the relationships between that resource and its collection digital object and to the metadata digital objects that describe it.

We have implemented in Java an API layer that mediates all interaction with the NDR, by calling on the constituent set of low-level Fedora operations. In addition to providing a relatively easy-to-use interface for services accessing the NDR, the API performs the vital task of ensuring that constraints of the data model are enforced. For example, the data model mandates that no metadata digital object should exist that does not have one (and only one) `metadataFor` relationship to a resource digital object.

We have used this API to bootstrap the production NDR with data from the pre-existing MR, thereby duplicating existing functionality in the new infrastructure. At the time of writing of this paper, this process is complete. The platform for our NDR production environment is a Dell 6850 server with dual 3Ghz Xeon processors, 32Gb of 400Mhz memory and 517Gb of SCSI RAID disk with 80MB/second sustained

performance. This server is running 64-bit LINUX, for reasons outlined later. We note that the 2006 cost for this production server is about 22K USD.

The NDR has over 2.1 million digital objects – 882,000 of them matching metadata from the MR, 1.2 million of them representing NSDL resources, and several hundred representing other information objects – agents, services, etc., - in the NDR data model. The representation of the relationships among these objects (those defined by the NDR data model and those internal to the Fedora digital object representation) produces over 165 million RDF triples in the triple-store. We have found that ingest into the NDR takes about .7 seconds per object – making data load for this rich information environment a non-trivial task.

This bootstrapping process has been a learning process in scaling up semantically-rich information environments. In order to understand the results, it is necessary to distinguish three components: core Fedora, the triple-store it uses to represent and query inter-object relationships, and the Proai¹²⁸ component that supports OAI-PMH.

Core Fedora is a web service application built on top of a collection of file-system resident XML documents (one file for each digital object) and a relational database that caches fragments and transformations of those documents for performance. These XML documents are relatively small and stable, and at present we are using about 21 GBytes of disk space to store these files across 39,000 directories. We have not experienced any scaling problems nor do we foresee any with this core architecture. In fact, as we expected from our knowledge of the Fedora implementation, basic digital object access is not really dependent on the size of the Fedora repository. For example, our tests on dissemination performance show that requests for metadata

¹²⁸ <http://www.fedora.info/download/2.1/userdocs/services/oaiprovider-service.html>

formats that are stored literally in the NDR are about 69 ms. Requests for formats that are crosswalked from stored formats using an XSLT transform service take about 480 ms.

The more challenging aspect of our data loading and implementation work has involved the triple-store. Relationships among Fedora digital objects, and therefore among nodes in the NDR graph, are stored persistently as RDF/XML in a datastream in the digital object and are indexed as RDF triples in a triple-store, which provides query access to the relationship graph. In the case of the NDR, this provides query functionality such as “return all resources related to a state standard, a specific collection, or in an OAI set”.

Triple-store technology is relatively immature. Scaling it up to accomplish our initial data load has been especially challenging. As part of our implementation of the Fedora relationship architecture (known as the resource index), we experimented with scaling and performance of a number of tripe-store implementations. Our extensive tests comparing Sesame, Jena, and Kowari are available online¹²⁹. One particular target of our testing was the performance of complex queries that involve multiple graph node joins – these are the types of queries we issue to perform OAI-PMH List Records operations that select according to metadata format, set, and date range. We found that Jena would not scale over a few tens of thousands of triples with complex query times approaching 20 minutes for complex queries over .5 million triples. Sesame can be configured in both native storage mode or on top of mysql. We found that Sesame-mysql, like Jena, was unable to return large results sets, producing an out-of-memory error due to accumulating the entire result set in memory. Our remaining

¹²⁹ <http://tripletest.sourceforge.net/>

tests comparing Sesame native to Kowari showed that for a database of several million triples Kowari was faster by a factor of 2 for simple queries, and by a factor of over 9000 for complex queries.

Although the Kowari implementation proved capable under controlled tests of high performance and scalability, we encountered a number of hurdles along the path of our data load. The apparent reality is that neither Kowari nor any other triple-store has been pushed to this scale. Such scale revealed unpleasant and previously undiscovered bugs, such as a memory leak that took months of effort to verify and find. Furthermore, we have found that the hardware requirements to run a large-scale semantic web application are non-trivial. Kowari uses memory-mapped indexes, which are both disk and memory-intensive. Presently the Kowari-based resource index requires over 54 GB of virtual memory, which is significantly larger than the 5 GB addressable by standard 32-bit processors and operating systems (thus the configuration of our production server described earlier).

In order to understand our results on semantic queries to the NDR resource index (storing 165 million triples), it makes sense to divide these queries into two classes. The first class of queries is relatively simple, such as those issued by a user application seeking all resources correlated with a state standard or another accessing all members of a collection. We have found that query performance in this case is on the order of 25ms for the simplest examples (no transitive joins over the graph) to about 250 ms for examples with 2-3 joins. The second class of queries are those that populate the NDR OAI server, Proai, which is a part of the Fedora service framework. Proai is an advanced OAI server that supports any metadata format available through the Fedora repository via direct datastream transcription or service-mediated dissemination. It operates over a MySQL database that is populated via resource-index queries to

Fedora (in batch after an initial load and incrementally over the lifespan of the Fedora repository). The resource-index queries to populate Proai are quite complex with semantics such as “list all Fedora disseminations representing OAI-records of a certain format, and get their associated properties and set membership information”. Such a query takes about 1 hour, when issued in batch over the fully loaded repository, and the combination of queries to pre-load the Proai database after the batch NDR load takes about 1-2 days. We note, however, that this load is only performed once on initial load of the NDR and that incremental updates, as information is added to the NDR, are much quicker.

Proai performance is quite impressive. Throughput on an OAI-PMH ListRecords request is about 900 records per/second, and we have been able to harvest all Dublin Core records from the NDR (to populate our search indexes) in about 3 hours.

Our results provide hardware guidelines for large Fedora implementations that use the resource index. We have found that they greatly benefit from a machine with large real memory, high-speed disks, and high-performance disk controllers. The Dual Xeon processors provide an excellent match for Fedora processing allowing uniform execution partitioning of core Fedora, the NDR API, Proai and MySql processing among the 4 hyper threaded CPU cores available. CPU clock rate is a minor performance factor compared with the overall memory and I/O performance of the chassis.

Conclusions

We have described in this paper our initial work in implementing an advanced infrastructure to support an information-rich NSDL. This infrastructure supports the integration and reuse of local and distributed content, the integration of that content with web services, and the contextualization of that content within a semantic graph.

The work described in this paper has advanced the state-of-the-art in two areas. First, it involves the innovative use of Fedora to represent an information network overlay. This data structure combines local and distributed content management, service-oriented architecture, and semantic web technologies. At a time when digital libraries need to move beyond the search and access paradigm, the INO supports contextualized and participatory information environments. Second, this work pushes the envelope on scaling issues related to semantic web technologies. Although RDF and the semantic web have existed for over 8 years, large-scale implementations still need to be demonstrated. Our experience with scaling the NDR is instructive to a number of other projects looking to build on top of semantic web technologies.

Chapter 12

A Web-Based Resource Model for Scholarship 2.0

Preface

This chapter is based on:

Lagoze, C., Van de Sompel, H., Nelson, M., Warner, S., Sanderson, R. and Johnston, P. A Web-Based Resource Model for Scholarship 2.0: Object Reuse and Exchange. *Concurrency and Computation: Practice and Experience* (Special Issue - Success in Furthering Scientific Discovery) [315].

The Open Archives Initiative – Object Reuse and Exchange project grew out of work to develop new models and infrastructure for scholarly communication, undertaken within the NSF-funded Pathways project [477] that began in 2004. The trajectory of this work over the years and the nature of its ultimate result reveal the evolution of thinking about digital libraries and web information.

The work began with an exclusively repository-centric focus. The problem as originally stated was to develop standards for the exchange of *compound digital objects*¹³⁰ among participating repositories (i.e., institutional and disciplinary). Although at this early point we envisioned the transformation of scholarly communication by this new document paradigm, we had yet to understand the importance of integrating the new paradigm with the general web architecture. Nor have we realized the utility of a general compound document model to a broad class of

¹³⁰ The nature of a compound digital object was described earlier in the context of the Dienst work (Chapter 7) and Fedora Project (Chapter 9). In essence, it is an aggregation of multiple streams of data package together in a single identified unit. As stated earlier in Chapter 5 there has been considerable work on formats that express this packaging notion.

Web 2.0 applications such as Flickr and blogs, which already had implicit aggregations of resources.

As the work progressed through 2006 and 2007 it became clear that this repository-centric focus, which was common across digital library projects, was problematic because the technologies and applications motivated by it were de facto isolated from the mainstream web. We were increasingly aware that the work of the Digital Library community and the Grid community [189], while reasonably successful in their own contexts, had failed to gain widespread deployment largely because of their divergence from this dominant information space. Furthermore, as stated earlier, they had essentially produced separate and parallel information spaces often invisible to crawler-based search engines.

We therefore changed course and began to focus on data models and standards that were strongly based on the fundamental notions of web architecture; resources, representations, URI's, and semantic links. The results are a set of resource-centric standards [304-312] that integrate well with both the RDF-based semantic web world and with the XML-based syndication domain (Atom).

Currently, OAI-ORE is being deployed in a number of high profile cyberinfrastructure and eScience projects. The oreChem project [285], funded by Microsoft, is using OAI-ORE as a modeling basis for information sharing of chemistry data and publications. The Data Conservancy project, recommended for funding within the NSF Office of Cyberinfrastructure Datanet program¹³¹, will be deploying OAI-ORE-based applications for a federated network of data repositories and associated information. This project is described in greater detail in Chapter 14.

¹³¹ <http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm>

Acknowledgments

Work on OAI-ORE has been supported by NSF award number IIS-0430906 (Pathways) and by the Andrew W. Mellon Foundation and Microsoft Corp. Special thanks go to Lee Dirks and Tony Hey at Microsoft for their efforts to fund this project. The OAI-ORE work is the result of a collaboration of many people worldwide. Special acknowledgment goes to the members of the OAI-ORE editorial team for their painstakingly long efforts to get this correct: Herbert Van de Sompel, Michael Nelson, Simeon Warner, Pete Johnston, and Rob Sanderson.

Introduction

Despite the high hype to reality ratio, Web 2.0 [390] represents a fundamental change in the Web. The early Web, which was by-and-large a network of read-only hyperlinked documents, only partially fulfilled Tim Berners-Lee's original vision [55] for the Web. In recent years, this vision has been realized as the Web has morphed into a "read/write" social space built on open standards, linked data[64], distributed services, authoring environments such as blogs and wikis, and social sites such as Facebook and Flickr.

Web 2.0 has had a profound effect on all aspects of society. This paper focuses on its impact on scholarship, its process and the manner in which it is communicated. The principles of Web 2.0 have enabled an emerging form of scholarship known as eScience, eScholarship, or Scholarship 2.0 [80, 239-241, 475, 485], the essence of which is illustrated in Figure 47 in a meme map, a notion introduced by Tim O'Reilly to illustrate Web 2.0 [390]. The center blue box of the meme map defines its strategic positioning, its positioning of the scholar users, and its core competencies relative to competing technologies (e.g., journals, traditional Web-based scholarly publishing).

The yellow ovals at the bottom enumerate the core principles underlying Scholarship 2.0. The outward facing, application-level aspects are in the red-hued ovals at the top. In the emerging Scholarship 2.0 paradigm traditional read-only journal and conference papers and peer review are being supplemented and even replaced by online mechanisms for contribution, participation, and feedback [88, 210] producing an “[a]rchitecture of participation that encourages user contribution” [193]. An open world [150] principle has emerged with manifestations such as open access [247, 491], open data [88], and open standards [451], all of which facilitate a remix and reuse culture allowing new scholarship to fully leverage past results. Finally, embedded in the principles of Scholarship 2.0 is the notion that the availability of the artifacts of the scholarly process, the building blocks of the scholarly value chain [463, 466], is as important as the dissemination of the products (e.g., journal articles, conference papers) [437].

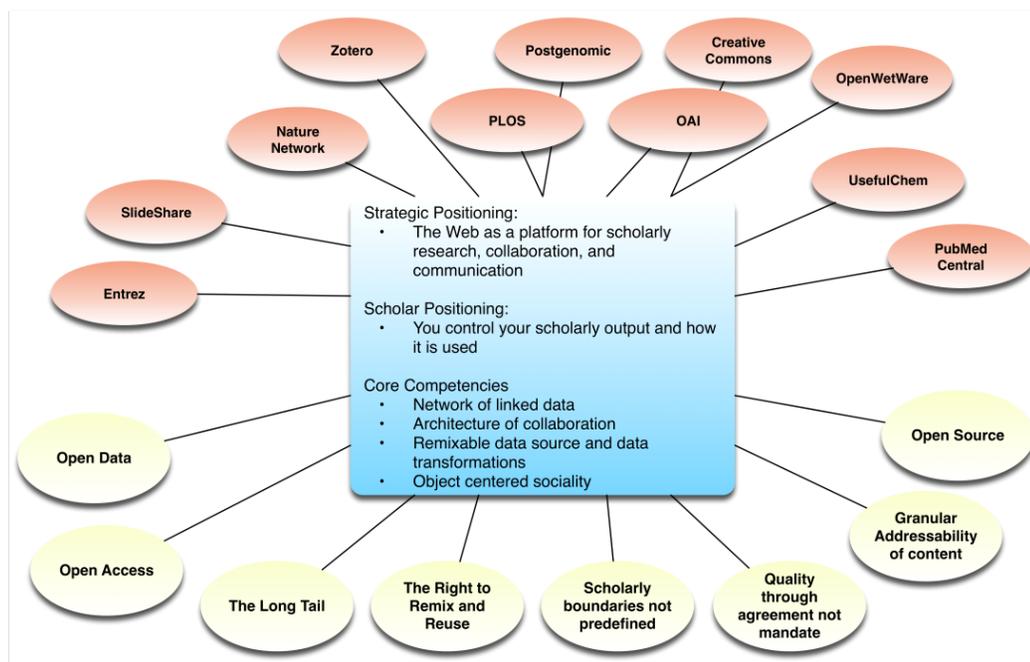


Figure 47 - Scholarship 2.0 meme map

We share the enthusiasm for the realities and potential of Scholarship 2.0. However, we argue in this paper that in order for this new scholarly paradigm to fully mature, there need to be standardized models for integrating the requirements of scholarship with the web architecture, and the manner in which it is used in Web 2.0. In particular, the integrity of scholarship depends on the notions of *citation*, the ability to unambiguously refer to named objects that form the building blocks and provenance of a scholarly result, and *fixity*, the means for defining the structure or boundary of a stable evidential artifact so that scholarly results can be tested and verified. Although the advantages of the componentized, mashup world of Web 2.0 have been demonstrated via numerous examples, standardized mechanisms of identity and fixity have yet to be defined. These mechanisms must be grounded in and fully compatible with the web architecture so that the products of scholarship are accessible to mainstream web applications, such as crawler-based search engines (e.g. Google, Yahoo!, Live Search), and reusable in other Web 2.0 contexts – for example, for teaching and learning [275].

We describe in this chapter our work on Open Archives Initiative – Object Reuse and Exchange (OAI-ORE) [306], a set of standards that address these limitations by providing a mechanism for describing, identifying, and sharing information about compositions, or aggregations, of web resources. The OAI-ORE standards leverage web architecture primitives and protocols such as Resources, URI's, and HTTP and emerging Semantic Web standards such as Linked Data [464]. They provide the foundation for a new generation of networked scholarly applications that exploit Web 2.0 while preserving the mechanisms vital for the integrity and verifiability of scholarship.

This paper is structured as follows. The next section very briefly summarizes the web architecture and Semantic Web technologies, a high level understanding of which is necessary to understand the motivations and mechanisms upon which the OAI-ORE standards are constructed. The subsequent section describes the notion of the compound document, or aggregation, that is implicit in scholarship but inadequately modeled on the web. The next section describes the OAI-ORE standards for identifying and describing compound documents as web resources. That is followed by a section describing implementations of OAI-ORE. The penultimate section describes related work. The final section includes some concluding remarks.

The architecture of the World Wide Web

Full details on the Web Architecture are described in [246]. Stated briefly, it provides the following notions.

A Resource is “an item of interest”. In the early Web, Resources were in most cases documents, such as HTML text or JPEG images, which were viewable in a browser. However, the notion of a Resource is fully generalizable to any entity, either physical or digital.

A URI is a uniform global identifier for a Resource. URIs conform to URI schemes (e.g., http, ftp, gopher) and each scheme defines the mechanism for assigning URIs within that scheme. Within the common http scheme, the URI is an identifier key in an HTTP (hypertext transfer protocol) request message, which may result in the return of information about the respective Resource. However, the ability to automatically de-reference an HTTP URI is not true for all URIs (nor even for all http URIs). For example, the URI for a physical entity such as a person is obviously not automatically dereferencable.

A Representation is a data stream corresponding to the state of a Resource at the time its URI is dereferenced. The Web Architecture allows for multiple Representations of a Resource with access mediated by Content Negotiation. A Resource may have no Representations and may exist only as an abstract “item of interest” (i.e., it may not be “retrievable”).

The relationship between Resources, URIs, and Representations is illustrated in Figure 48.

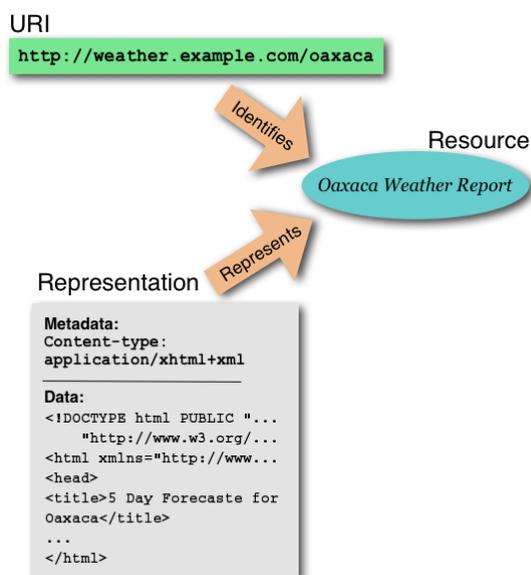


Figure 48 - Identifier, resource, and representation (from [246])

A Link is a directed connection between two Resources. In most common usage, a link is expressed via link or anchor tags (a hyperlink) in an HTML Representation of the originating Resource to the URI of another Resource.

The web is frequently modeled as a graph, where the nodes are Resources labeled with their respective URIs and the edges are links between those Resources. This notion of a web graph is illustrated in Figure 49. As illustrated, the graph is not fully connected

between the Resource identified by the subject (a URI) and the object (a URI or a Literal).

A set of RDF triples is referred to as an RDF Graph because it can be represented as a node and directed-arc diagram, in which each triple is represented as a node-arc-node link. The nodes of an RDF Graph are the subjects and objects of the constituent triples. Multiple triples form a connected graph when they share subjects and/or objects with the same URI's.

An example of an RDF Graph is shown in Figure 50. As shown, the subject and Predicate of a triple are always URIs (the URI is indicated by the text in the yellow circle and shown with bracketed syntax <A> in the table) and the object may be a URI or a literal (shown as a blue rounded rectangle in the graph and in quotations in the table). Note that while this example shows the RDF graph as connected, this not necessary since a triple may include subjects or objects with any URI, regardless of the existence of those URIs in any other triples.

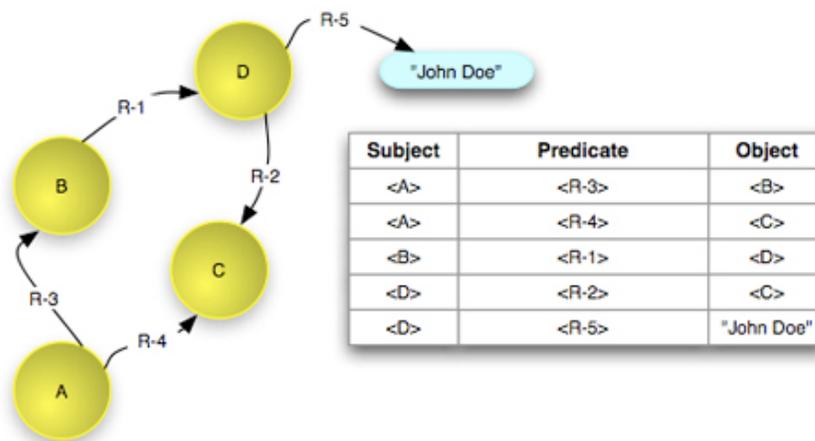


Figure 50 - RDF triples and graph representation

Another tool from the Semantic Web, the RDF Vocabulary Description Language [91], provides the mechanisms to specify vocabularies for defining the types of these

relationships. In combination with the RDF-defined relationship `rdf:type` this vocabulary makes it possible to express types for Resources. Figure 51 shows an example of this. As shown, the objects of the triples with `rdf:type` predicates are URIs that denote classes or types.

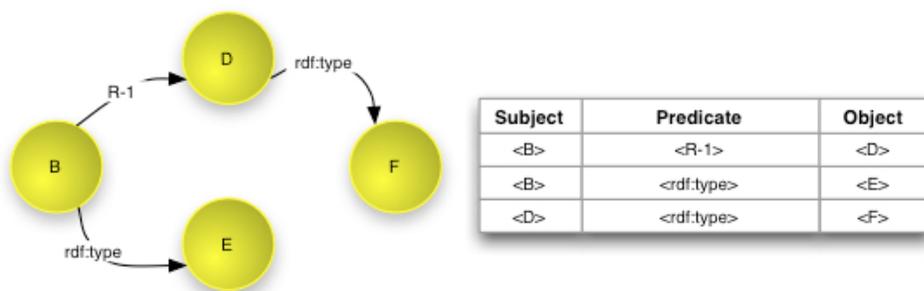


Figure 51 - Expressing types in RDF

These Semantic Web concepts provide the foundation of the recently introduced notion of linked data, which expresses a method of publishing and connecting data on the Web. Linked data provides a formal underpinning [71] for the mashup (i.e., Web Resource re-use and re-aggregation) principle that has become popular in Web 2.0 [390]. Interest in the open linked data concept as a vehicle for fueling reuse and new scholarship has recently accelerated as indicated by the size of the “linked data cloud” shown in Figure 52.

Scholarly documents – Pre-Web to Web 2.0

Scholarship has always been data-centric. Regardless of the subject of a scholarly investigation, its common, and distinguishing characteristic relative to popular culture, is its basis in evidence. This evidence, or underlying data, varies according to the scholarly field, and includes numerical data, images such as astronomical observations, illustrations such as the design of an experiment, transcripts of ethnographic investigations, and various other genres. Figure 53 shows an example of a multi-genre aggregation of evidence of scholarship.

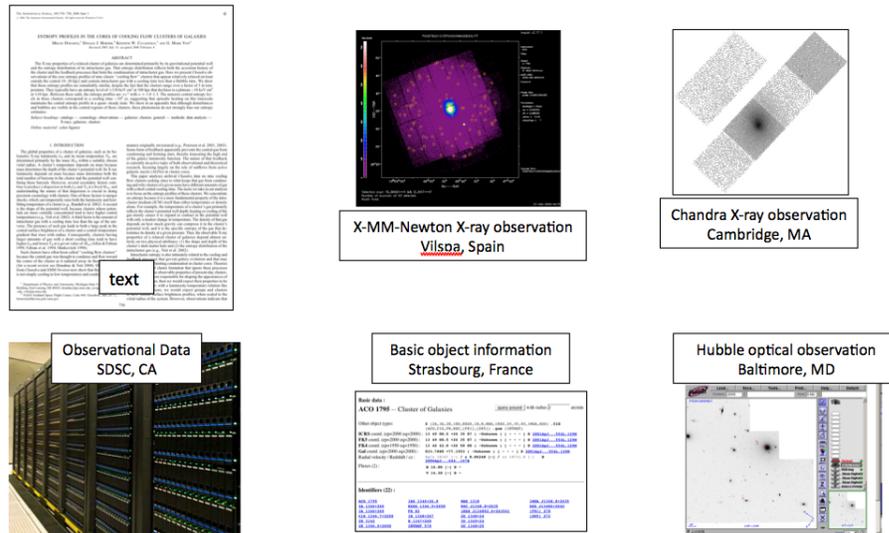


Figure 53 - Aggregation of evidence of scholarship

The emergence of the Web and the subsequent transition of publications from physical forms to digital forms have dramatically changed the accessibility of scholarly publications. Whereas in the pre-digital age monographs, journals, and conference proceedings were generally available only in research libraries, access is now possible to anyone with a computer and Internet connection. In contrast to this rather revolutionary change in accessibility, however, there has until quite recently been very little change in the actual nature of the scholarly article. PDF has replaced ink and paper, but the form of the documents as static mainly textually based artifacts has been preserved. In almost all cases, the actual data underlying the work, the complete evidence of the scholarship, remains unconnected and generally inaccessible in private notebooks or, even worse, on personal, unpreserved magnetic media. Furthermore, although these PDF documents may have embedded images, references, descriptions of software packages, and the like, these components are generally unavailable as individual information units for reuse and re-factoring. In summary, although the

actual process of scholarship frequently produces aggregations of text, images, underlying data, software, and other genre of materials, the mechanisms for expressing this rich substrate in most visible artifacts of the scholarly process, the publications, have until recently been limited.

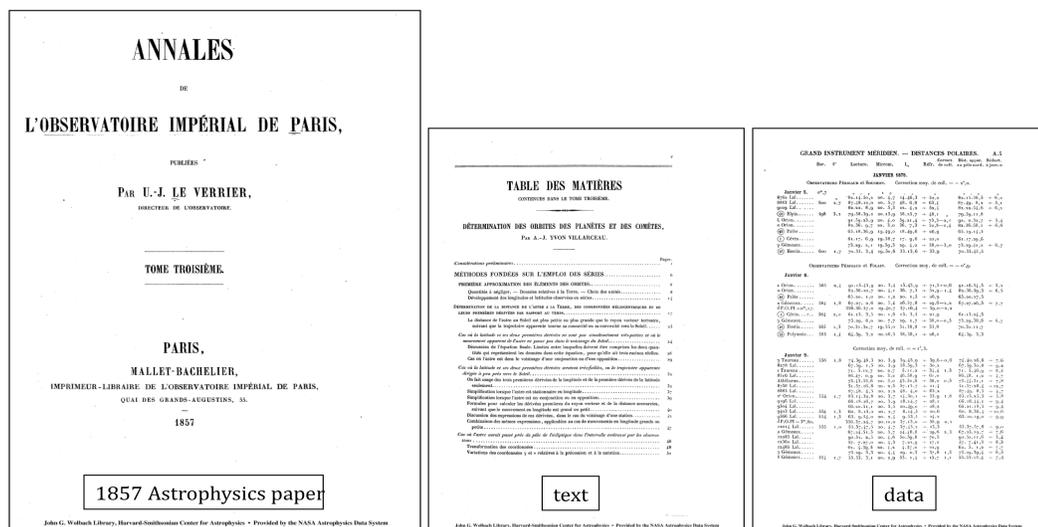


Figure 54 - Pre-Web Scholarly Publication

The digital library research community has dedicated considerable effort to develop mechanisms that overcome the limitations of physical artifacts and their digital clones. This work has produced a variety of models of compound digital objects [255, 357, 404] and architectures of repositories that store and provide access to them [301, 439, 461]. The results of this research are deployed in the increasing number institutional repositories [143, 253, 351] and disciplinary repositories, such as arXiv¹³⁴, which are increasingly the preferred location for storage of and access to scholarly publications.

These repository and digital object architectures are all distinguished by their ability to model, store, and provide access to compound documents, aggregations of multiple

¹³⁴ <http://arxiv.org>

data streams, thereby making it possible to represent the full scholarly record. Although the actual mechanisms for doing so are unique to the particular architecture, they all follow a common paradigm: they provide a human-readable (HTML) “splash page” or entry point for the individual document, that then contains hyperlinks to the components of the compound document. An example of a splash page (from the arXiv) with embedded hyperlinks to compound document components is illustrated in Figure 55.

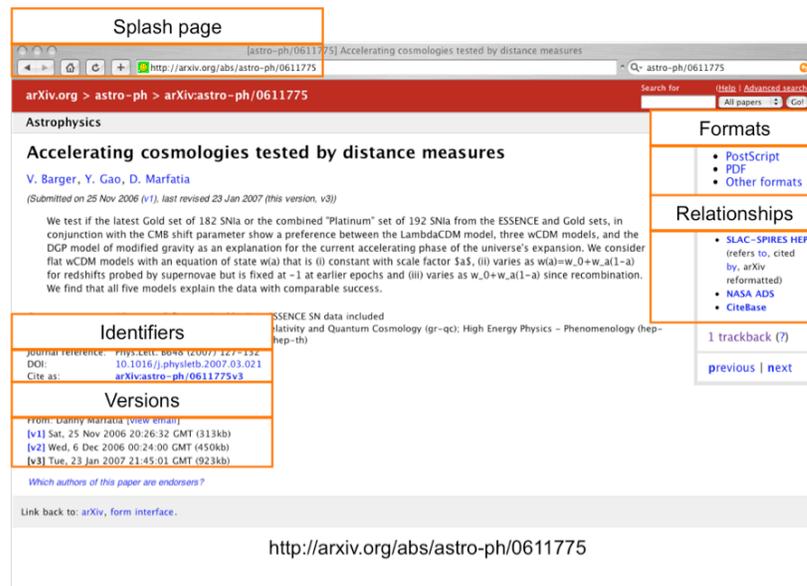


Figure 55 - HTML splash page.

As illustrated, the HTML splash page links to the various components of the document it represents including the text in multiple formats (PDF, PostScript, etc.), multiple versions of the document, and related data objects. While this mechanism is useful for the human user who can view the rendered splash page in a browser and interpret the implied compound document, it is opaque to a machine agent (e.g., a crawler for a search engine) that cannot distinguish this splash page from any Web page with embedded hyperlinks (not all of which are the “root” of a compound document).

This problem is illustrated in Figure 56 that shows a Web graph with an embedded compound document rooted with a slash page. In this illustration, the splash page node is red, linked components are green, and other nodes are yellow. The implied (but not explicit) document “boundary” is denoted by the red dashed line.

Assume that a crawler begins its graph traversal at the node labeled “S”. It will eventually traverse to the splash page and the components but the relationship among them and their co-existence within a document boundary is not evident in any machine-readable form.

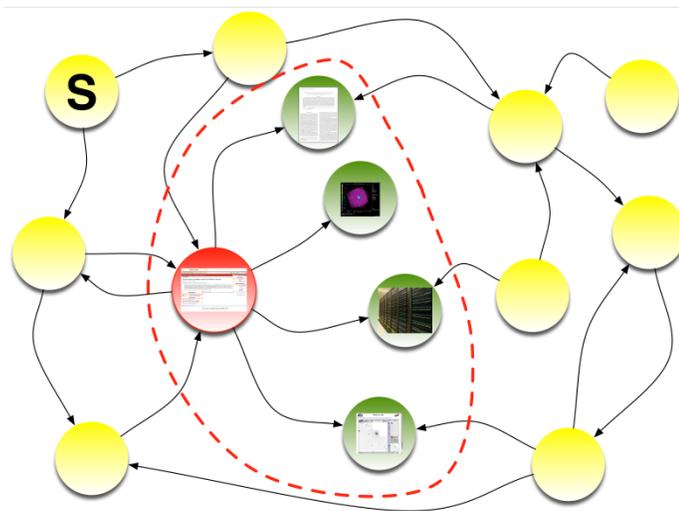


Figure 56 - Web graph with embedded compound object.

The crawler problem described above is really symptomatic of two more general problems.

- *Identity* – As described earlier, the notion of URI-tagged Resources is fundamental to the Web architecture. However, the compound document (the entity within the implied boundary in Figure 56) is **not** a Resource and has no unique identity, and is therefore not a component of the Web graph. Note that as illustrated each of its components including the splash page are Resources with URIs, but none of those URIs, in the terms of the Web architecture,

denote the document (aggregation) as a whole. This is problematic for citation, which is arguably one of the core functionalities in scholarly communication. Without an identity it is impossible to unambiguously establish a citation to the document as a whole (rather than to an individual component). Furthermore, without unique identity it is impossible to fold these compound documents back into the Web 2.0 framework. The aggregation is not a constituent of the linked data cloud, and is therefore not available for reuse, re-aggregation, and object-centered social interaction (e.g., collaboration, annotation, etc.) [181, 182].

- *Description* – Without an explicit and machine-readable description of the structure of the compound document, it is impossible to create user-centered or utility services on these aggregations. Examples of user-centered services include browser-based plug-ins for visualizing and navigating over a compound object, printing all components of an object, or saving a Web-based compound object to local disk. Examples of utility services include facilities for transferring objects among cooperating repositories for preservation purposes in the manner of LOCKSS [419] or search engines that improve ranking and results presentation due to understanding of document structure.

Our solution to these two problems, the Open Archives Initiative Object Reuse and Exchange (OAI-ORE) standards for identifying and describing aggregations on the Web, is examined in the next section.

OAI-ORE: Identifying and describing compound objects

The essence of the OAI-ORE solution is illustrated in Figure 57. In the illustration, the blue node is the Resource denoting the Aggregation. The purple node is a Resource

Map, which asserts the existence of the Aggregation and describes its structure (boundary).

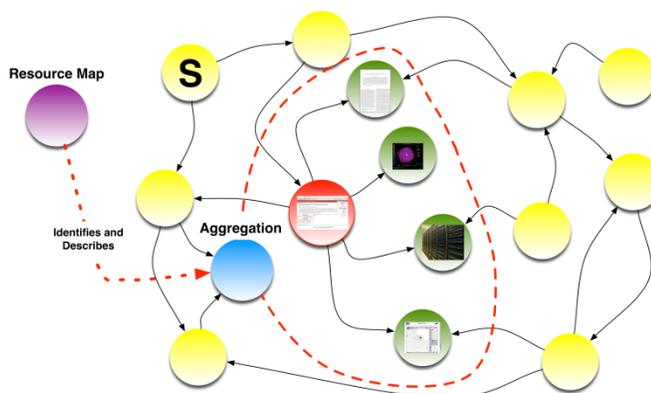


Figure 57 - Identification and description of an Aggregation.

As illustrated, OAI-ORE adds two nodes to the Web graph, each with a unique URI. The first denotes the Aggregation, and can thereby used for citation, as the target of any other links, or as a component of a subsequent Aggregation or mashup. The second denotes the description of the Aggregation, the Resource Map, and has a URI that when dereferenced emits the serialized triples describing identity of the Aggregation and its structure.

This solution is based on the primitives defined in the previously described Architecture of the World Wide Web [246] that defines a Resource as an item of interest; a URI as a global identifier for a Resource; and a Representation as a datastream corresponding to the state of a Resource at the time its URI is dereferenced via some protocol (e.g. HTTP).

In addition, the solution is grounded in the principles introduced by the Semantic Web, in which URIs are also used to identify non-document Resources, such as real-world entities (e.g. people or cars), or even abstract entities (e.g. ideas or classes). These non-document Resources have no Representation to indicate their meaning. OAI-ORE

adopts the following approach, proposed by the Linked Data effort [64], for obtaining information about those Resources:

- Use of HTTP URIs to identify non-document Resources, in this case the Aggregation;
- Publication of another Resource, in this case the Resource Map, with a Representation that provides information about the non-document Resource at a HTTP URI other than the HTTP URI of the non-document Resource;
- Leverage of HTTP mechanisms to allow discovery of the HTTP URI of the published resource from the HTTP URI of the non-document resource.

The OAI-ORE standards include an RDF-based data model for Aggregations, syntaxes for serializing instances of the data model, and mechanisms for providing HTTP access to those serializations. The remainder of this section will summarize those standards. Complete details are available through the OAI-ORE documentation suite [306].

Data Model

The essence of the RDF-based data model is described here and is illustrated in Figure 58. The full details are available in the OAI-ORE Abstract Data Model specification [304].

In order to be able to unambiguously refer to an aggregation of Web resources, a new Resource is introduced that stands for a set or collection of other Resources. This new Resource, named an Aggregation, has a URI just like another Resource on the Web. And, since an Aggregation is a conceptual construct, it is a non-document Resource that does not have a Representation.

Following the Linked Data guidelines, another Resource is introduced to make information about the Aggregation available. This new Resource, named a Resource

Map, has a URI and a machine-readable Representation that provides details about the Aggregation. In essence, a Resource Map expresses which Aggregation it describes (the `ore:describes` relationship in Figure 58), and it lists the Aggregated Resources that are part of the Aggregation (the `ore:aggregates` relationship in Figure 58, a subproperty of `dcterms:hasPart`). But, a Resource Map can also express relationships and properties pertaining to all these Resources, as well as metadata pertaining to the Resource Map itself, e.g. who published it and when it was most recently modified (the `dcterms:creator` and `dcterms:modified` relationships in Figure 58). A Resource Map can also express relationships of the Aggregation, Aggregated Resources, and the Resource Map itself with any arbitrary other Resource, as long as the resulting RDF graph is connected.

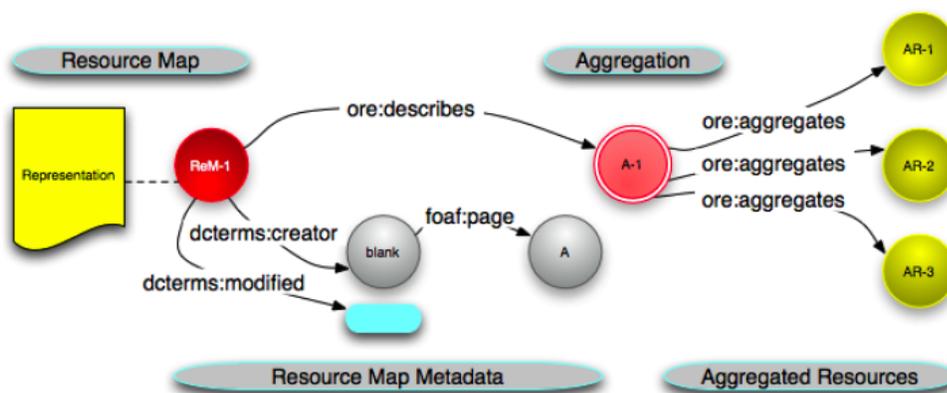


Figure 58 - A Resource Map and Aggregation with 3 Aggregated Resources

In addition, for discovery purposes, the data model allows a Resource Map to express that an Aggregated Resource of a specific Aggregation is also part of another Aggregation. This is achieved by means of the `ore:isAggregatedBy` relationship (the inverse of `ore:aggregates`) between the Aggregated Resource and that other Aggregation. Also stating that an Aggregated Resource is itself an Aggregation (nesting Aggregations) is supported. To that purpose, an `ore:isDescribedBy` relationship (the inverse of `ore:describes`, and a

subproperty of `rdfs:seeAlso`) is expressed between the Aggregated Resource and a Resource Map that describes it as being itself an Aggregation. Furthermore, the use of non-protocol-based identifiers (such as DOIs) that can be expressed as URIs is quite common for referencing scholarly assets. In order to support this practice, the `ore:similarTo` relationship between an Aggregation and a somehow equivalent resource identified by a non-protocol-based URI is expressed. The specificity of `ore:similarTo` is situated between `rdfs:seeAlso` and `owl:sameAs`.

Proxies: Aggregated Resources in Context

We note that the URI asserted in a Resource Map to denote an Aggregated Resource of a particular Aggregation is no different than the URI that denotes that Resource independent of the Aggregation. However, it is important for citing and expressing provenance in scholarly communication and other applications that a resource such as a dataset included in some context, for example a specific article, be distinct from the same dataset outside the context of that article, or in the context of another article.

To accomplish this differentiation, OAI-ORE introduces the notion of a Proxy. A Proxy is a Resource that stands for an Aggregated Resource in the context of a specific Aggregation. The URI of a Proxy provides a mechanism for denoting a Resource in context. Figure 59 shows the `ore:ProxyFor` and `ore:ProxyIn` relationships between a Proxy and an Aggregated Resource and an Aggregation, respectively. It also illustrates how citing the Aggregated Resource is different from citing its Proxy: the former cites a Resource “as is”, the latter cites that Resource as it exists in the context of a specific Aggregation. In order to work seamlessly in the Web and to provide context information to OAI-ORE aware clients, resolution of HTTP URIs assigned to Proxies must lead to the Aggregated Resource, and the response must include a HTTP Link Header [385] that points to the Aggregation.

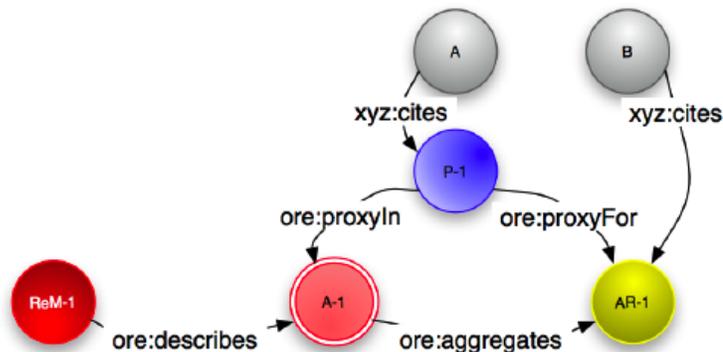


Figure 59 - Citing a Resource in the context of an Aggregation

Resource Map Serializations

A Resource Map has a Representation that describes an Aggregation in some serialization syntax. OAI-ORE explicitly specifies three serialization syntaxes, Atom XML [309], RDF/XML [310], and RDFa [311], and other serialization syntaxes are possible. Which one to choose will largely depend on the use case and on the technical environment available to a Resource Map publisher. For example, in cases where an expressive HTML splash page exists an RDFa approach might be attractive. Note that multiple Resource Maps, each using a different serialization syntax can describe the same Aggregation, and that these may differ in expressiveness.

Although the data model is based on RDF, we were committed to also specify a serialization based on Atom, to allow Aggregations to become the subject of Web 2.0 reuse scenarios and of workflows based on the Atom Publishing Protocol [218]. The Atom Publishing Protocol adds a uniform read/write approach to Web 2.0, which could be of significant benefit in scholarly communication scenarios.

However, the task of reconciling the data model with the Atom model proved to be non-trivial due to tensions between the RDF model and the XML-oriented Atom specification. The former is graph-based, with precise semantics that are global rather than local to a specific document. The latter is hierarchical, (XML) document-centric,

and has intentionally loose element definitions. It took several, dramatically different iterations of the Atom serialization to arrive at an acceptable solution.

The resulting approach expresses an Aggregation by means of an Atom entry, and makes use of Atom's extensibility mechanisms in much the same way as Google Data¹³⁵ does. For example, Atom's link element with an OAI-ORE-specific value for the `rel` attribute is used to aggregate resources. And, lacking a mechanism from the Atom community to express triples, an `ore:triples` element was introduced to act as a wrapper for RDF descriptions. To support unambiguous interpretation of Atom serializations of Resource Maps, a GRDDL transform was implemented that extracts all contained triples that pertain to the OAI-ORE data model, both from the native Atom elements and from the `ore:triples` extension element, and expresses them in RDF/XML.

Leveraging HTTP

In order to make OAI-ORE work in the HTTP-based Web, both the Aggregation and the Resource Map are assigned HTTP URIs, and the Cool URIs for the Semantic Web guidelines [430] are adopted to support discovery of the HTTP URI of a Resource Map given the HTTP URI of an Aggregation. Figure 60 illustrates a situation in which the arXiv Aggregation is described by both an Atom XML and an RDF/XML Resource Map, and in which a client is led to the Atom version via an HTTP 303 redirect and Content Negotiation.

Authoritative Resource Maps

After one party has published a Resource Map that contains a description and a URI for a new Aggregation, any other party can publish competing or even conflicting

¹³⁵ <http://code.google.com/apis/gdata/>

Resource Maps that describe the same Aggregation. To address this we distinguish between Authoritative and Non-Authoritative Resource Maps in the same way as the Linked Data guidelines. An Authoritative Resource Map is one that is accessible by dereferencing the URI of the Aggregation that it describes, for example using the aforementioned Cool URI mechanisms. A Non-Authoritative Resource Map is one not reachable in this manner. The rationale for this approach is that the party that introduces a new Aggregation simultaneously mints URIs for both the Aggregation and the Resource Map, and actually controls both.

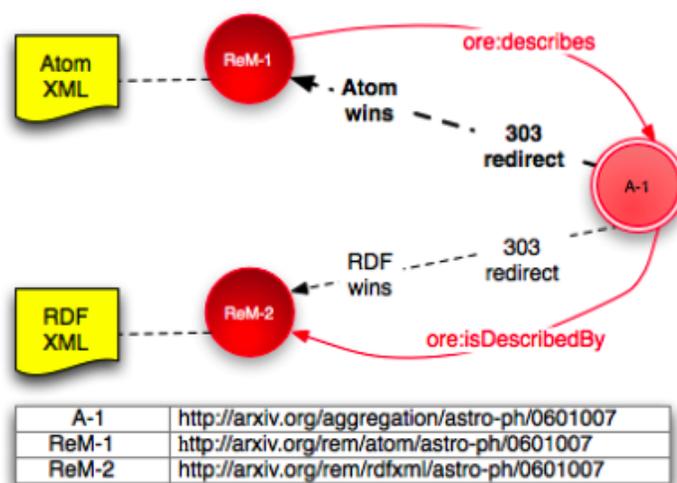


Figure 60 - Resource Map discovery from an Aggregation using Cool URIs

Deployment, experimentation, and implementation

The OAI-ORE specifications were released in October, 2008. An in-depth evaluation of functionality, adoption, and impact is still premature. This section describes deployments by early adopters to leverage the specifications.

Foresite: Revealing Aggregations

In order to provide feedback on the evolving OAI-ORE specification, the UK's Joint Information Systems Committee (JISC) funded an experiment to investigate applying it to an extensive scholarly collection: the approximately four million articles that are

part of the JSTOR¹³⁶ collection. By developing open source OAI-ORE libraries¹³⁷ and applying them to produce interlinked Resource Maps, the Foresite project effectively demonstrated the feasibility of exposing common scholarly artifacts to the Data Web in the manner proposed by OAI-ORE. The project provided valuable feedback that helped refine the OAI-ORE specifications, and had a significant impact on the Atom serialization of Resource Maps.

The overall structure of the Aggregations, and associated Resource Maps, produced for the JSTOR collection mirrors the journal-issue-article hierarchy of the JSTOR content. Each journal is modeled as an Aggregation of journal issues; each issue is an Aggregation of articles; and each article is an Aggregation of individual page images and a PDF-formatted version of the entire article (Figure 61). The Aggregated Resources at each level are also the subject and/or object of a `fst:followedBy` relationship introduced to preserve the page-turning order for pages within an article, articles within an issue and so forth. Because `fst:followedBy` is not a global relationship, but rather only applies within the context of a specific Aggregation, Proxies for these Aggregated Resources were introduced. The article Aggregations interlink via `dcterms:references` relationships for citations, further confirming the necessity of the graph-based nature of the OAI-ORE data model, even though the main JSTOR content hierarchy is tree-shaped. The Resource Maps were published on a Web server at the University of Liverpool.

The resulting OAI-ORE descriptions are of immediate business importance to JSTOR. While JSTOR stores the OCR-ed full-text of each article, it is only able to openly

¹³⁶ <http://www.jstor.org>

¹³⁷ <http://foresite-toolkit.googlecode.com>

expose this kind of topological metadata, and would lose its market advantage (and the participation of contributing publishers) if the full-text were exposed. Having the topology of their collection available in a standardized format that provides links back to their protected full-text documents and images, facilitates reuse in third party applications that can help drive traffic to the JSTOR site and increase its customer base.

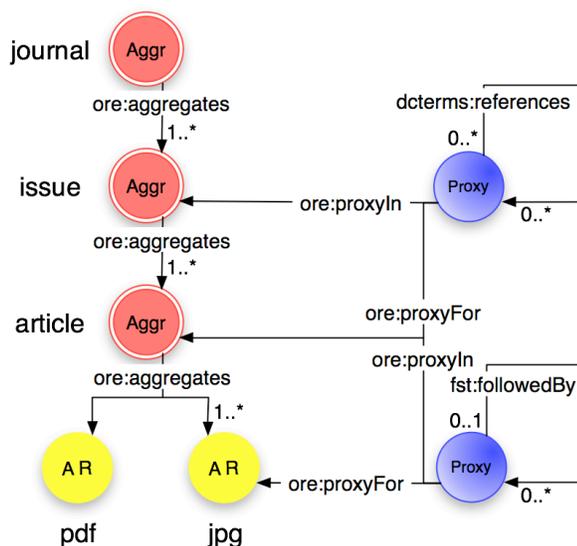


Figure 61 - JSTOR collection mapped to the OAI-ORE data model

Astronomy publication workflow

Datasets are of fundamental importance in observational sciences such as astronomy. The astronomy community has developed sophisticated repositories and data standards, exemplified by the Sloan Digital Sky Survey¹³⁸ and the National Virtual Observatory¹³⁹, which provide excellent facilities for registering and accessing large datasets. However, when submitting an article, both new datasets that were created to

¹³⁸ <http://www.sdss.org>

¹³⁹ <http://www.us-vo.org>

arrive at findings reported in an article, and data citation information that reveals the reuse of existing datasets are often lost, “left behind”, on the personal computer of the author.

A team at Johns Hopkins University is collaborating with the American Astronomical Society to capture datasets as part of the publication workflow [121]. In the newly devised publication workflows, OAI-ORE Aggregations are used to glue an article and its associated datasets together, and Resource Maps that describe these Aggregations are the tokens that move around between author, publisher and dataset repository as the publication process proceeds [162]. At each stage of the publication workflow, the Resource Map is used to convey the current state of the Aggregation, and is then updated to reflect the new state that is then passed on to the next workflow phase. For example, as a Resource Map is passed from the publisher to the dataset repository and back again, it is updated to contain the URIs of datasets that are registered in the repository, and that were used for the article. This allows the publisher to link to the datasets that were used for a specific article, and the repository to link to papers that used a specific dataset.

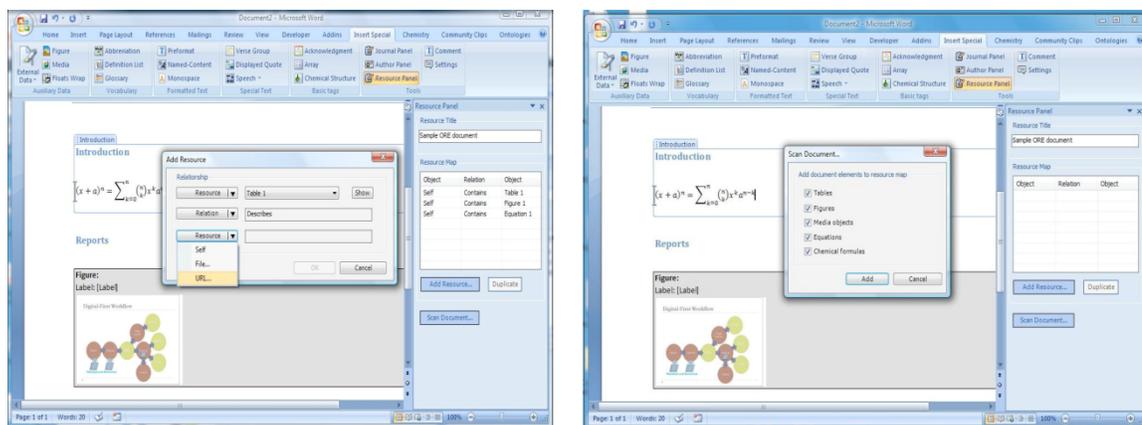
Generally, the availability of these Aggregations enables new services to be built on both the publishing platform and the data repository. If the practices proposed by this novel publication workflow became commonplace, it would represent a significant improvement in the efficiency of scientific communication.

Authoring, editing and reusing

The success of OAI-ORE depends on the ease with which Aggregations and Resource Maps are authored and disseminated on the Web. In many cases, they will be generated automatically based on information that is available in an information system. For example, the arXiv.org database contains all information that is necessary

to automatically generate Aggregations and their associated Resource Maps. And, in the astronomy project described above, the ability to create Resource Maps is built into familiar authoring environments in a manner that makes it a side effect of the authoring process and thus minimizes the burden on authors.

Like all cyberinfrastructure, the success of such authoring environments depends on the manner in which assembling all resources that relate to a particular research task or publication fits into the normal scholarly workflow. Two authoring environments that demonstrate this are the Literature Object Reuse and Exchange (LORE) tool created by Gerber et al. [208], and the SCOPE work of Cheung et al. [118]. LORE is a Firefox extension that communicates via Ajax with a Sesame2 data store for maintaining the OAI-ORE graphs that are generated. LORE allows for the generation of fine-grained metadata and relationships allowing, for example, the designation that a certain resource is contextual information about the literature work that is being studied. The SCOPE work led to the development of the Provenance Explorer [242], a stand-alone Java application with functionalities similar to those of LORE, but aimed at the creation, editing and publication of scientific compound objects.



Manual Addition of Components

Automated Recognition of Components

Figure 62 - Screenshots of Word OAI-ORE plug-in

Work at Microsoft to integrate OAI-ORE into its popular Office® product line offers the promise of making the authorship and deposit of compound documents completely mainstream. An OAI-ORE plug-in for Microsoft Word® is scheduled for third quarter 2009. This plug-in will combine both automatic generation of compound documents, by identifying internal structures such as tables, figures, and citations as document components, with manual assemblage, by providing an interface that allows the author to designate network or file based new components. Figure 62 shows preliminary screenshots of this plug-in.

Enhanced publications

The Dutch SURFshare program¹⁴⁰ and the European DRIVER II project¹⁴¹ are collaborating on cyberinfrastructure to join a multitude of scientific repositories that hold publications and research data. The goal is to give researchers better means to share and access scientific materials through innovative services. One of the envisioned services relates to enhanced publications, composites of textual publications and supporting resources such as research-data, visualizations, annotations, related websites, etc. To ensure the integrity and usability of such enhanced publications it is important that all its components and their interrelations are being preserved.

¹⁴⁰ <http://www.surffoundation.nl/en/>

¹⁴¹ <http://www.driver-community.eu/>

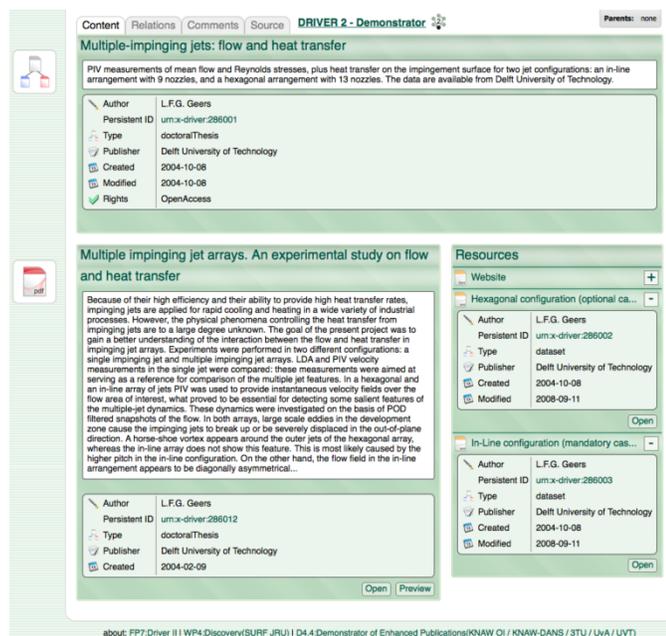


Figure 63 - The splash page dynamically rendered from Resource Map

A study into object models suitable for the representation of enhanced publications recommended the use of OAI-ORE. As a result, a demonstrator project [238] was launched. This project used OAI-ORE to model enhanced publications for multiple scientific disciplines ranging from engineering to journalism. The project investigated approaches to meet a variety of requirements, including presentation, navigation, persistent identification, granularity of referencing, handling of sequentially ordered resources, visualization of interrelationships, etc. The results are available at the project site¹⁴². The project chose RDF/XML to express Resource Maps and uses an XSLT-based approach to dynamically generate an HTML “splash page” from them. In each splash page, a Content tab (Figure 63) lists all crucial metadata about the enhanced publication, shows its textual component and associated metadata, and lists additional resources with their metadata. Many of these resources are themselves

¹⁴² <http://driver2.dans.kna.nl/demonstrator/html/>

modeled as Aggregations, and hence also have their own splash page. To support an understanding of the relationships among resources of an Aggregation and of nested Aggregations, a Relations tab that loads a Java applet fueled by Resource Map content is introduced. Overall, the demonstrator is remarkable because of the elegance and simplicity of the ORE implementation. It clearly illustrates that ORE can be used as a basic model for enhanced publications, and points at the need for community-defined vocabularies to convey expressive relationships among scientific resources.

Chemistry scholarship 2.0

The oreChem Project [285] is a two-year collaboration (initiated in January 2009) funded by Microsoft to investigate and deploy Scholarship 2.0 infrastructures and applications for Chemistry research. The project brings together chemists and information and computer scientists from University of Cambridge, Cornell University, Indiana University, Penn State University, University of Queensland, and University of Southampton. A key aspect of this project is the design and implementation of an interoperability infrastructure that will allow chemistry scholars to share, reuse, manipulate, and enhance data that are located in repositories, databases, and Web services distributed across the network. The foundations of this planned infrastructure are the OAI-ORE specifications.

At the time of writing of this paper (April 2009) the project has been working for three months so only preliminary work has been done. Initial work involves the design of an ontology for formally describing Chemistry research [15] and using that ontology as the basis for extending the OAI-ORE aggregation model for three Chemistry-specific aggregations: publications, molecules and associated properties, and experiments with their context. These models will then be used as the interoperability substrate for the creation of Chemistry semantic knowledge base (the eChemistry Web) that combines data from existing databases (e.g., crystallographic data), retrospective capture from

existing digital documents, authorship of new compound publications, and in-laboratory capture (e.g., electronic laboratory notebooks). In the latter phases of the project we hope to build innovative analysis tools that will extract new “scientometric” information and knowledge from the eChemistry Web.

Related work

Given the widespread use of aggregations in both the physical and the web world, it comes as no surprise that other efforts have investigated this domain. Prior work in the web realm can be grouped in two main categories depending on the party that introduces aggregations. In one case, that is the Web navigator (agent or reader), in the other case it is the administrator of a Web-based information system. In this section we look at a number of efforts in both categories, and evaluate their capabilities to identify aggregations, to enumerate the constituent resources of an aggregation, to express relationships among resources, and to accommodate resources that are distributed on the Web.

In the Web navigator case, either a user groups resources based on some intent, or a robot tries to infer the implicitly defined members of an aggregation. The robotic approaches range from heuristics [175, 344] to machine-learning [163, 164]. While these approaches are useful, they are imperfect and dependent on the perception of those encoding the heuristics or training set and they do not necessarily reflect the intention of the original authors of the Web resources. And, while these approaches may succeed at selecting the distributed resources that are part of an implicitly defined aggregation, they are not capable of inferring the relationships between those resources, nor do they propose a way to unambiguously describe the aggregation.

The approaches that involve an interactive user include tools such as GroupMe¹⁴³ and LinkBunch¹⁴⁴. LinkBunch lets users submit several URIs that are then assigned a new HTTP URI that, when dereferenced, returns an HTML page that lists and links to the originally submitted URIs. The “bunch” has a new HTTP URI identity, it enumerates its members, and it readily handles distributed Web resources. However, the identity of the bunch is the same as that of the HTML page that describes it, and expressing relationships between the bunched resources is not supported. GroupMe! is similar, with the addition of social tagging capabilities, but has the same problems as LinkBunch.

Some Web navigator approaches work in an opposite granular direction, supporting disaggregation of a single Web resource (i.e., an HTML page) into multiple resources. This can be done automatically, such as for segmented display on limited devices such as PDAs [114] or for recovering structured records from Web pages [176].

Decomposition can also be done manually, such as for reuse and sharing of parts of a Web page (e.g., ClipMarks¹⁴⁵). All these approaches, manually or automatically, can be thought of as adding (or inferring) HTML anchors where none exist. These approaches assign identities to the newly created resources (fragments of the original resource), but they provide no approach to describe the original resource as an aggregation of these new resources, nor do they allow expressing relationships among them.

¹⁴³ <http://groupme.org>

¹⁴⁴ <http://linkbun.ch>

¹⁴⁵ <http://clipmarks.com>

In approaches that have the administrator of a Web information system in the driver seat, several technologies exist to deal with resource aggregations. Sitemaps were briefly considered as a serialization option for Resource Maps. Google, Yahoo and Microsoft support the Sitemap Protocol¹⁴⁶, a simple XML file format that allows Web sites to list the URIs they want crawled by robots. Sitemaps provide for minimal metadata (e.g., last modification date, update frequency and crawl priority), but no attempt is made to provide semantic typing, and handling arbitrary distributed resources is not supported. Indeed, in the interest of trust, the Sitemap Protocol specifies a significant limitation on URI paths that can be listed in a Sitemap file. For example, a Sitemap at level `www.foo.com/a/b` can list URIs at level `a/b` and below, but it cannot list URIs at `www.foo.com/a/c`, `www.foo.com/d` or `www.bar.com/`.

We made a deliberate decision to avoid the many existing packaging formats, such as MPEG-21 DIDL [47], METS [367], FOXML [301], IMS-CP [4], and BagIt [86]. First, packaging base64-encoded content in a wrapper document does not resonate well with the Resource/URI/Representation paradigm of the Web Architecture. Still, most of these formats also support a by-reference mechanism to deliver content, in which URIs can be used. However, although these formats are prominent in their respective communities, they have not gained broader adoption. And while these approaches can address identification, and enumeration of distributed resources, they have uneven capabilities to express the graph-based OAI-ORE model, due to their hierarchical perspective.

In the course of the OAI-ORE effort, we also attempted to model aggregations as Atom feeds, not entries. We ultimately decided that was the wrong granularity,

¹⁴⁶ <http://sitemaps.org>

especially since common Web 2.0 reuse scenarios as well as Atom Publishing Protocol functionality are situated at the level of Atom entries. The Atom Syndication Format was preferred over the various RSS formats in anticipation of using the Atom Publishing Protocol [218].

The POWDER [23] specifications that were developed in the same timeframe as OAI-ORE address a problem space similar to that of OAI-ORE. However, POWDER approaches the problem from the opposite perspective, focusing on capabilities to assert (via "Description Resources") that a group of resources share certain properties (e.g. access rights), rather than asserting arbitrary properties about resources that, for some reason, are grouped into an aggregation. That is, in POWDER the notion of shared properties defines an aggregation, whereas in OAI-ORE an aggregation can be created for any reason deemed important by its creator. Also, while POWDER provides capabilities to describe a group of resources using a variety of approaches including regular expressions, it does not introduce an identity for the aggregation.

Conclusion

This paper has introduced the OAI-ORE solution to the resource aggregation problem, which we argue meets a critical need in the development of Scholarship 2.0.

Alignment of the solution with the Web Architecture and with the practices of the Semantic Web and Linked Data effort will integrate scholarly communication with the mainstream Web 2.0 environment. In this manner scholarly artifacts will be visible to common web tools and applications. This will benefit the broader community by making research materials more visible, verifiable, and by facilitating reuse in other domains such as teaching and learning.

While OAI-ORE was motivated by scholarly communication, we believe that the proposed solution has broader applicability. Aggregations, sets, and collections are as

common on the web as they are in the everyday physical world. There are many situations where agents and services on the web would benefit if aggregations were unambiguously enumerated and described.

Evaluation of the OAI-ORE work depends on its adoption and evolution over time. The work has so far benefited from significant community involvement throughout the specification process, and the international team that developed the solution includes representatives with backgrounds in scholarly publishing, eScience, repository infrastructure, digital libraries, Web search engines, linked data, and information interoperability. Work by early adopters, such as the Foresite project and Johns Hopkins publication workflow project, are promising indicators that these community contributions have led to a solution that stands realistic chances for significant adoption.

Chapter 13

Lessons for Cyberinfrastructure Projects¹⁴⁷

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:

¹⁴⁷ 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 arXiv¹⁴⁸ 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

¹⁴⁸ <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 chemistry¹⁴⁹, 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

¹⁴⁹ 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 sailing¹⁵⁰, suggested the notion of “20 questions” for requirements gathering¹⁵¹ 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

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

¹⁵¹ <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@home¹⁵², 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

¹⁵² <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@home¹⁵³ and Project FeederWatch¹⁵⁴. 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.

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

¹⁵⁴ <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 Observatory¹⁵⁵. 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

¹⁵⁵ <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.

Concluding Remarks and Observations

“May all your problems be technical” [135]

Jim Gray, 1998 ACM Turing Award Winner

When the NSF, DARPA, and other US and international funding agencies began large-scale funding of digital library research in the early 1990’s, they were motivated by a number of goals. As I have described, the work on the first goal, stimulating basic research in networking, security, databases, information retrieval, and other areas, was quite successful. These largely *technical* endeavors produced a number of results that rapidly evolved from initial research prototypes to technology that was deployed and used on a global scale. Work on another largely technical goal (which arose largely in the context of DLI-2), the deployment online of a number of culturally, intellectually, and historically important digital collections, was also quite successful.

However, another highly-promoted goal, the development and large-scale deployment of new network-based information infrastructure, met with a number of obstacles, and the results were significantly short of success. The obstacles to success were not technical. In fact, many of the proposed solutions were technically sound and provided rich functionality (often greater than that which we see on the web). The actual barrier to success was the reality that infrastructure, and most notably information infrastructure, is a deeply *sociotechnical* phenomenon. As scholarship in

Science, Technology, and Society (STS) and numerous infrastructure studies have shown, the development and spread of infrastructure entails complex interactions among the technology, the people and organizations who will use and be impacted by the technology, the social norms that govern the interaction of people, organizations, and technical artifacts, and co-existing technologies.

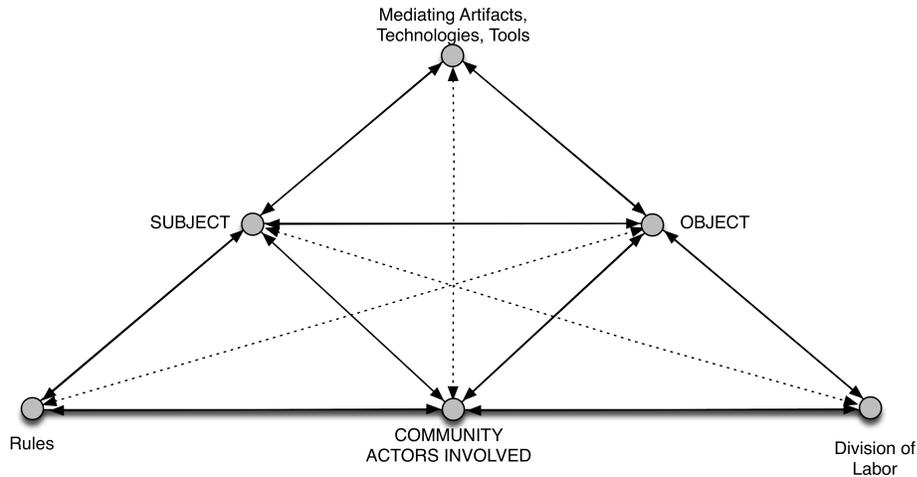


Figure 64 - Activity system

This complexity is the subject of the STS theories and frameworks described in Chapter 4. The notion of an activity system, illustrated in that chapter and repeated here in Figure 64, shows the mediated nature of technical change, especially one as broad reaching as information infrastructure. Activity Theory describes how technology is just one mediating factor in activity transactions and how a technology that does not mesh well with other mediating factors will be *disrupted*; modified or discarded. This is especially true for infrastructure technology, which should be non-intrusive and effectively invisible.

Because of a variety of factors described in Chapter 2, most digital library research effectively ignored this reality. Focusing mainly on technical goals, these researchers

assumed that existing, well-established norms of institutionally-based information organization and management would persist as the context for deployment of new search engines, scanning and display technologies, rights management mechanisms, and other technical advances. The digital library community was, of course, not alone in this myopia. Other examples include the entertainment industry, newspaper, publishers, and even the computer industry (e.g., Microsoft), all of which expected that digital information technology would somehow seamlessly mesh into their existing ways of doing business.

In reality, the act of putting information online and giving people almost universal access to that information dramatically disrupts every part of the activity network for virtually all information-related activities (i.e. every thing) that people do. As Yochai Benkler [48], Lawrence Lessig [335] and others have noted, it has a dramatic “democratizing effect”. The history of the web and the increasingly profound changes in Web 2.0 demonstrate the impact of this democratization. Based on a few relatively simple protocols and standards and constrained by virtually no rules, the web we use today is essentially an organic development – a creation from within the web itself, rather than defined by an institution, standards board, or funding agency. In the terminology of Activity Theory and Actor-Network Theory it represents a self-stabilization (and perhaps self-optimization) of the activity network in response to the multiple interactions of networked information technology with other components of the network.

Notably, these profound changes were enabled rather than determined by the web technology invented by Tim Berners-Lee in 1989. Those simple architectural components - resources, URIs, HTML, and HTTP - provided the basis for the rather simple Web 1.0 "document Web and, relatively unchanged, were later used as the

basis of the Web 2.0 "social web" that has so transformed the way we live, learn, communicate, and consume. Indeed, this basic web technology has not determined the nature of the information environments built upon it. Instead, the multiple applications and evolving forms demonstrate both the flexibility of the underlying technology and the shaping influences of different social contexts. We can expect in the future that these shaping factors will produce further unexpected changes in the way we interact with information.

In some ideal world, an analysis such as this would lead to a set of prescriptions for future efforts. In reality, though, the self-stabilization process describe above is probably non-deterministic and not attainable by tweaking some variables.

Nevertheless, there is certainly room for future research on heuristics on how to improve our approaches to infrastructure. Hopefully, the Data Conservancy project I describe in Chapter 13 will provide some progress towards developing these heuristics. In the meantime, we can expect to observe a continuation of the unpredictable and dramatic transformations in technology and information that we have experienced over the past two decades. As we work to find our way through these changes and choose alternatives, we may have to rely on the words of the enigmatic Yogi Berra who said "*when you get to a fork in the road, take it.*"¹⁵⁶.

¹⁵⁶ http://en.wikiquote.org/wiki/Yogi_Berra

REFERENCES

1. A Gentle Introduction to SGML. in Sperberg-McQueen, C.M. and Burnard, L. eds. *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*, Text Encoding Initiative, 1994.
2. Dublin Core Metadata Element Set, Version 1.1. Dublin Core Metadata Initiative. January 14, 2008. Available at <http://www.dublincore.org/documents/dces/>.
3. Functional requirements for bibliographic records: final report. IFLA Study Group on the Functional Requirements for Bibliographic Records, International Federation of Library Associations and Institutions. Section on Cataloguing. Standing Committee, 2009. Available at http://www.ifla.org/files/cataloguing/frbr/frbr_2008.pdf.
4. IMS Learning Resource Meta-data Best Practices and Implementation Guide. Version 1.2.1 - Final Specification. IMS Global Learning Consortium, Inc., September 28, 2001. Available at http://www.imsglobal.org/metadata/imsmdv1p2p1/imsmd_bestv1p2p1.html.
5. Information Retrieval (Z39.50): Application Service Definition and Protocol Specification. ANSI/NISO, 1995. Available at <http://www.loc.gov/z3950/agency/Z39-50-2003.pdf>.
6. Monty Python-The Black Knight, Retrieved February 17, 2009, from YouTube: <http://www.youtube.com/watch?v=zKhEw7nD9C4>.
7. No Brief Candle: Reconceiving Research Libraries for the 21st Century. Council on Library and Information Resources, August, 2008.
8. Sony Corp. v. Universal City Studios. U.S. Supreme Court ed. *464 U.S. 417*, Washington, DC, 1984.

9. Sustainable Digital Data Preservation and Access Network Partners (DataNet). NSF-07-601. National Science Foundation, Office of Cyberinfrastructure, Directorate for Computer & Information Science & Engineering, 2007. Available at <http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm>.
10. The Data Conservancy: A Digital Research and Curation Virtual Organization. Proposal to the National Science Foundation, Solicitation NSF 07-601. Johns Hopkins University, 2008.
11. The Mercury Project and Library Information System II: The First Three Years. Carnegie Mellon University, 1992.
12. The OpenURL Framework for Context-Sensitive Services. Z39.88. NISO, 2004.
13. *Understanding Metadata*. NISO Press, Bethesda, MD, 2004.
14. Abbas, J., Norris, C. and Soloway, E., Middle School Children's Use of the ARTEMIS Digital Library, in *ACM/IEEE Joint Conference on Digital Libraries (JCDL '02)*, (Portland, OR, 2002), ACM Press, 98-105.
15. Adams, N. Semantic Chemistry, 2009. Retrieved April 8, 2009, from <http://www.semanticuniverse.com/articles-semantic-chemistry.html>.
16. Adler, S., Lamersdorf, W., Munke, M., Rucker, S., Spahn, H., Berger, U., Bruggemann-Klein, A. and Haber, C. Grey Literature and Multiple Collections in NCSTRL. in Barth, A., Breu, M., Endres, A. and de Kemp, A. eds. *Digital Libraries in Computer Science: The MeDoc Approach*, 1998.
17. Agre, P.E. Information and Institutional Change: The Case of Digital Libraries. in Bishop, A.P., Van House, N.A. and Battenfield, B.P. eds. *Digital Library Use*, MIT Press, Cambridge, 2003.
18. Allemang, D. and Hendler, J.A. *Semantic web for the working ontologist: modeling in RDF, RDFS and OWL*. Morgan Kaufmann Publishers/Elsevier, Amsterdam, Boston, 2008.

19. Altman, M. and King, G. A Proposed Standard for the Scholarly Citation of Quantitative Data. *D-Lib Magazine*, 13 (3/4).
20. Andersen, D.G., Balakrishnan, H. and Kaashoek, M.F., Resilient Overlay Networks, in *18th ACM SOSP*, (Banff, Canada, 2001).
21. Anderson, C. The Long Tail. *Wired*, 2004. Available at <http://www.wired.com/wired/archive/12.10/tail.html>.
22. Antoniou, G. and Van Harmelen, F. *A Semantic Web primer*. MIT Press, Cambridge, Mass., 2004.
23. Archer, P. POWDER: Use Cases and Requirements. World Wide Web Consortium, 2007. Available at <http://www.w3.org/TR/2007/NOTE-powder-use-cases-20071031/>.
24. ARL/OCLC Strategic Issues Forum. The Keystone Principles. Bimonthly Report. 207. Association of Research Libraries, 1999. Available at <http://www.arl.org/newsltr/207/keystone.html>.
25. Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R.H., Konwinski, A., Lee, G., Patterson, D.A., Rabkin, A., Stoica, I. and Zaharia, M. Above the Clouds: A Berkeley View of Cloud Computing. UCB/EECS-2009-28. EECS Department, University of California, Berkeley, 2009. Available at <http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html>.
26. Arms, W. A Viewpoint Analysis of the Digital Library. *D-Lib Magazine*, 11 (7/8). Available at <http://www.dlib.org/dlib/july05/arms/07arms.html>.
27. Arms, W. Digital Libraries (Online Edition), 2005. Retrieved April 24, 2009, from <http://www.cs.cornell.edu/wya/diglib/>.
28. Arms, W.Y. Automated Digital Libraries: How Effectively Can Computers Be Used for the Skilled Tasks of Professional Librarianship? *D-Lib Magazine*, 6 (7/9). Available at <http://www.dlib.org/dlib/july00/arms/07arms.html>.
29. Arms, W.Y. *Digital libraries*. MIT Press, Cambridge, MA, 2000.

30. Arms, W.Y., Blanchi, C. and Overly, E.A. An Architecture for Information in Digital Libraries. *D-Lib Magazine* (February). Available at <http://www.dlib.org/dlib/february97/cnri/02arms1.html>.
31. Arms, W.Y., Dushay, N., Fulker, D.W. and Lagoze, C. A Case Study in Metadata Harvesting: the NSDL. *Library Hi Tech*, 21 (2).
32. Arms, W.Y., Hillmann, D., Lagoze, C., Krafft, D., Marisa, R., Saylor, J., Terrizzi, C. and Van de Sompel, H. A Spectrum of Interoperability: The Site for Science Prototype for the NSDL. *D-Lib Magazine*, 8 (1). Available at <http://www.dlib.org/dlib/january02/arms/01arms.html>.
33. Atkins, D.E., Droegemeier, K.K., Feldman, S.I., Garcia-Molina, H., Klein, M.L., Messerschmitt, D.G., Messina, P., Ostriker, J.P. and Wright, M.H. Revolutionizing Science and Engineering Through Cyberinfrastructure. National Science Foundation Blue-Ribbon Panel on Cyberinfrastructure, 2003.
34. Atkinson, R. Library Functions, Scholarly Communication, and the Foundation of the Digital Library: Laying Claim to the Control Zone. *The Library Quarterly*, 1996 (July).
35. Aufderheide, P. Competition and Commons: The Public Interest in and after the AOL-Time Warner Merger. *Journal of Broadcasting and Electronic Media*, 46 (4).
36. Avancini, H., Candela, L. and Straccia, U. Recommenders in a personalized, collaborative digital library environment. *Journal of Intelligent Information Systems*, 28 (3).
37. Backstrom, L., Huttenlocher, D.P., Kleinberg, J. and Lan, X., Group Formation in Large Social Networks: Membership, Growth, and Evolution, in *KDD'06*, (Philadelphia, 2006), ACM.
38. Bade, D. The Perfect Bibliographic Record: Platonic Ideal, Rhetorical Strategy or Nonsense? *Cataloging & Classification Quarterly*, 46 (1). 109 - 133. Available at <http://www.informaworld.com/10.1080/01639370802183081>.

39. Baresi, L.G., Barzotto, F. and Paolini, P. From Web Sites to Web Applications: New Issues for Conceptual Modeling. *Lecture Notes in Computer Science* (1921). 89-100.
40. Bauman, M.L. The Evolution of Internet Genres. *Computers and Composition*, 16 (2). 269-282.
41. Bearman, D. and Sochats, K. Metadata Requirements for Evidence. Archives & Museum Informatics, University of Pittsburgh, School of Information Science, 1996. Available at <http://www.lis.pitt.edu/~nhprc/BACartic.html>.
42. Bearman, D. and Trant, J. Electronic Records Research Working Meeting May 28-30, 1997, A Report from the Archives Community. *D-Lib Magazine*. Available at <http://www.dlib.org/dlib/july97/07bearman.html>.
43. Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F. and Stein, L.A. OWL Web Ontology Language Reference. W3C Recommendation. World Web Consortium, February 10, 2004. Available at <http://www.w3.org/TR/owl-ref/>.
44. Beckett, D. and McBride, B. RDF/XML Syntax Specification (Revised), 2004. W3C: <http://www.w3.org/TR/rdf-syntax-grammar/>.
45. Beer, D. and Burrows, R. Sociology and, of and in Web 2.0: Some Initial Considerations. *Sociological Research online*, 12 (5).
46. Beghtol, C. A proposed ethical warrant for global knowledge representation and organization systems. *Journal of Documentation*, 58 (5). 507-532.
47. Bekaert, J., Hochstenbach, P. and Van de Sompel, H. Using MPEG-21 DIDL to Represent Complex Digital Objects in the Los Alamos National Laboratory Digital Library. *D-Lib Magazine*, 9 (11). Available at <http://www.dlib.org/dlib/november03/bekaert/11bekaert.html>.
48. Benkler, Y. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press, New Haven, CT, 2006.

49. Bergman, M. The Deep Web: Surfacing Hidden Value. *Journal of Electronic Publishing*, 7 (1). 150-159. Available at <http://www.press.umich.edu/jep/07-01/bergman.html>.
50. Berman, S.J., Abraham, S., Battino, B., Shipnuck, L. and Neus, A. Navigating the media divide: Innovating and enabling new business models. IBM Global Business Services, 2007.
51. Berners-Lee, T. Cool URIs don't change, 1998. Retrieved February 11, 2009, from W3C: <http://www.w3.org/Provider/Style/URI>.
52. Berners-Lee, T. Information Management: A Proposal, 1989. Retrieved 2009, February 23, from <http://www.w3.org/History/1989/proposal.html>.
53. Berners-Lee, T., Fielding, R. and Masinter, L. Uniform Resource Identifiers (URI): Generic Syntax. RFC. 2396. IETF, August, 1998. Available at <http://www.ietf.org/rfc/rfc2396.txt>.
54. Berners-Lee, T., Fielding, R. and Masinter, L. Uniform Resource Identifiers (URI): Generic Syntax. RFC 3986. IETF, August, 2005. Available at <http://www.ietf.org/rfc/rfc3986.txt>.
55. Berners-Lee, T. and Fischetti, M. *Weaving the Web: the original design and ultimate destiny of the World Wide Web by its inventor*. Harpers, San Francisco, 1999.
56. Berners-Lee, T., Hendler, J. and Lassila, O. The Semantic Web. *Scientific American*, 2001 (50). Available at <http://www.sciam.com/2001/0501issue/0501berners-lee.html>.
57. Bijker, W.E. *Of bicycles, bakelites, and bulbs: toward a theory of sociotechnical change*. MIT Press, Cambridge, Mass., 1995.
58. Bijker, W.E., Hughes, T.P. and Pinch, T.J. *The Social construction of technological systems: new directions in the sociology and history of technology*. MIT Press, Cambridge, Mass., 1987.

59. Birman, K.P. *Building secure and reliable network applications*. Manning, Greenwich, CT, 1996.
60. Birmingham, W.P., Durfee, E.H., Mullen, T. and Wellman, M.P., The Distributed Agent Architecture of the University of Michigan Digital Library, in *AAAI Spring Symposium on Information Gathering in Heterogeneous, Distributed Environments*, (Stanford, CA, 1995), 98-108.
61. Bishop, A.P. and Star, S.L. Social informatics of digital library use and infrastructure. in Williams, M.E. ed. *Annual review of information science and technology: Vol. 31*, Information Today, Medford, NJ, 1996, 301-401.
62. Bishop, A.P., Van House, N.A. and Battenfield, B.P. *Digital Library Use*. MIT press, Cambridge, 2003.
63. Bishop, A.P., Van House, N.A. and Battenfield, B.P. *Digital library use: social practice in design and evaluation*. MIT Press, Cambridge, Mass., 2003.
64. Bizer, C., Cyganiak, R. and Heath, T. How to Publish Linked Data on the Web. Free University of Berlin, 2007. Available at <http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/>.
65. Black, R.W. Access and Affiliation: The Literacy and Composition Practices of English Language Learners in an Online Fanfiction Community. *Journal of Adolescent & Adult Literacy*, 49 (2). 118-128.
66. Black, R.W. Online fanfiction: What technology and Popular Culture Can Teach Us about Writing and Literacy Instruction. *New Horizons for Learning Online Journal*, 11 (2).
67. Blacker, F., Crump, N. and McDonald, S. Managing Experts and Competing through Innovation: An Activity Theoretical Analysis. *Organization*, 6 (5).
68. Blacker, F., Crump, N. and McDonald, S. Organizing Processes in Complex Activity Networks. *Organization*, 7 (277).
69. Blackmore, S.J. *The meme machine*. Oxford University Press, Oxford ; New York, 1999.

70. Boer, N.I., van Baalen, P. and Kumar, K., An Activity Theory Approach for Studying the Situatedness of Knowledge Sharing, in *35th Hawaii International Conference on System Sciences*, (Honolulu, 2002).
71. Bojars, U., Breslin, J.G., Finn, A. and Decker, S., Using the Semantic Web for linking and reusing data across Web 2.0 communities, in *Web Semantics: Science, Services and Agents on the World Wide Web*, (Karlsruhe, 2008).
72. Boler, M. *Digital media and democracy : tactics in hard times*. MIT Press, Cambridge, Mass., 2008.
73. Bollen, J. and Nelson, M., Adaptive Networks of Smart Objects, in *International Conference on Parallel Processing Workshops (ICPPW'02)*, (2002).
74. Borges, J.L. *Other inquisitions, 1937-1952*. Translated by Ruth L.C. Simms. Introd. by James E. Irby. University of Texas Press, Austin, 1964.
75. Borgman, C.L. Designing Digital Libraries for Usability. in Bishop, A.P., Van House, N.A. and Bittenfield, B.P. eds. *Digital Library Use*, MIT Press, Cambridge, 2003.
76. Borgman, C.L. Digital libraries and the continuum of scholarly communication. *Journal of Documentation*, 56 (4). 412-430.
77. Borgman, C.L., Digital Libraries: Now here, or nowhere? (Keynote), in *Joint Conference on Digital Libraries*, (Austin, TX, 2009). Available at <http://works.bepress.com/borgman/213>.
78. Borgman, C.L. From acting locally to thinking globally: A brief history of library automation. *Library Quarterly*, 67 (3). 215-249.
79. Borgman, C.L. *From Gutenberg to the global information infrastructure: access to information in the networked world*. MIT Press, Cambridge, Mass., 2000.
80. Borgman, C.L. *Scholarship in the digital age information, infrastructure, and the Internet*. MIT Press, Cambridge, Mass., 2007.

81. Borgman, C.L. The invisible library: Paradox of the global information infrastructure. *Library Trends*, 51 (4). 652.
82. Borgman, C.L. What are digital libraries? Competing visions. *Information Processing & Management*, 1999 (35). 227-243. Available at <http://yunus.hacettepe.edu.tr/~tonta/courses/fall2002/kut780/Borgman2.pdf>.
83. Borgman, C.L., Bates, M., Cloonana, M.V., Efthimiadis, E.N., Gilliland-Swetland, A., Kafai, Y., Leazer, G.L. and Maddox, A. Social aspects of digital libraries. National Science Foundation, 1996.
84. Boswell, D. Distributed high-performance web crawlers: A survey of the state of the art. UCSD, December 10, 2003. Available at <http://www.cs.ucsd.edu/~dboswell/PastWork/WebCrawlingSurvey.pdf>.
85. Bowker, G.C. and Star, S.L. *Sorting things out: classification and its consequences*. MIT Press, Cambridge, Mass., 1999.
86. Boyko, A., Kunze, J., Littman, J. and Madden, L. The BagIt Package Format, 2008. Retrieved April 3, 2009, from <http://www.cdlib.org/inside/diglib/bagit/bagitspec.html>.
87. Boyle, J. *The public domain : enclosing the commons of the mind*. Yale University Press, New Haven, CT, 2008.
88. Bradley, J.-C., Enhancing Scientific Communication through Open Notebook Science, in *Scholar2Scholar*, (Philadelphia, 2008), Drexel University Library. Available at <http://scholar2scholar.wikispaces.com/Presentation+Abstract+-+Enhancing+Scientific+Communication+through+Open+Notebook+Science>.
89. Braun, S., Schmidt, A. and Walter, A., Ontology Maturing: a Collaborative Web 2.0 Approach to Ontology Engineering, in *WWW2007*, (Banff, Canada, 2007).
90. Bray, T., Paoli, J. and Sperberg-McQueen, C.M. Extensible Markup Language (XML) 1.0. W3C Recommendation. REC-xml-19980201. World Wide Web Consortium, February, 1998. Available at <http://www.w3.org/TR/1998/REC-xml-19980210>.

91. Brickley, D. and Guha, R.V. RDF Vocabulary Description Language 1.0: RDF Schema. Recommendation. W3C, February 10, 2004. Available at <http://www.w3.org/TR/rdf-schema/>.
92. Brickley, D. and Guha, R.V. Resource Description Framework (RDF) Schema Specification. W3C Candidate Recommendation. CR-rdf-schema-20000327. World Wide Web Consortium, March 27, 2000. Available at <http://www.w3.org/TR/rdf-schema>.
93. Brickley, D., Hunter, J. and Lagoze, C. ABC: A Logical Model for Metadata Interoperability. Working Paper. Harmony Project, 1999. Available at http://www.ilrt.bris.ac.uk/discovery/harmony/docs/abc/abc_draft.html.
94. Briet, S. *Qu'est-ce que la documentation*. EDIT, Paris, 1951.
95. Brin, S., L. Page, The anatomy of a large-Scale hypertextual Web search engine, in *Proc. 7th International World Wide Web Conference*, (1998).
96. Brin, S. and Page, L. Anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30 (1-7). 107-117. Available at <http://www-db.stanford.edu/pub/papers/google.pdf>.
97. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J., Graph structure in the Web, in *Proceedings of the 9th International World Wide Web Conference: The web: The next generation*, (Amsterdam, 2000), Elsevier. Available at <http://www9.org/w9cdrom/160/160.html>.
98. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J., Graph structure in the web: Experiments and models, in *9th WWW Conference*, (2003), 309-320.
99. Broder, A.Z., Glassman, S.C., Manasse, M.S. and Zweig, G. Syntactic Clustering on the Web. Technical Note. 1997-015. DEC Systems Research Center, 1997. Available at <http://gatekeeper.dec.com/pub/DEC/SRC/technical-notes/SRC-1997-015-html/>.

100. Bruce, T.R. and Hillmann, D. The Continuum of Metadata Quality: Defining, Expressing, Exploiting. in Hillmann, D. ed. *Metadata in Practice*, American Library Association, Chicago, 2004.
101. Buchanan, G. and Hinze, A., A Generic Alerting Service for Digital Libraries, in *JCDL 2005*, (Tucson, AZ, 2005), ACM/IEEE, 131-140.
102. Buckland, M.K. *Library services in theory and context*. Pergamon Press, Oxford England ; New York, 1988.
103. Buckland, M.K. What is a "digital document"? *Document Numerique*, 2 (2). 221-230.
104. Buckland, M.K. What is a "document"? *Journal of the American Society of Information Science*, 48 (9).
105. Bush, V.F. As We May Think. *Atlantic Monthly*, 1945 (July). Available at <http://www.isg.sfu.ca/~duchier/misc/vbush/vbush-all.shtml>.
106. Calhoun, K. The Changing Nature of the Catalog and its Integration with Other Discovery Tools. Library of Congress, March 17, 2006. Available at <http://www.loc.gov/catdir/calhoun-report-final.pdf>.
107. Callan, J.P., Smeaton, A.F., Beaulieu, M., Borlund, P., Brusilovsky, P., Chalmers, M., Lynch, C., Riedl, J., Smyth, B., Straccia, U. and Toms, E. Personalisation and Recommender Systems in Digital Libraries. NSF-EU, May, 2003. Available at <http://www.ercim.org/publication/ws-proceedings/Delos-NSF/Personalisation.pdf>.
108. Callon, M., Law, J. and Rip, A. *Mapping the dynamics of science and technology: sociology of science in the real world*. Macmillan, Basingstoke, Hampshire, 1986.
109. Candela, L., Castelli, D., Pagano, P., Thanos, C., Ioannidis, Y., Koutrika, G., Ross, S., Schek, H.-J. and Schuldy, H. Setting the Foundations of Digital Libraries. *D-Lib Magazine*, 13 (3/4).

110. Carroll, J.J., Bizer, C., Hayes, P. and Stickler, P. Named Graphs. 2005. Available at <http://sites.wiwiss.fu-berlin.de/suhl/bizer/pub/NamedGraphs-WebSemanticsJournal.pdf>.
111. Carroll, J.J., Bizer, C., Hayes, P. and Stickler, P., Named Graphs, Provenance and Trust, in *WWW 2005*, (Chiba, Japan, 2005), ACM. Available at <http://www2005.org/cdrom/docs/p613.pdf>.
112. Castelli, D., Pagano, P. and Thanos, C. OpenDLib: an infrastructure for new generation digital libraries. *International Journal of Digital Libraries*, 4 (1). 45-47.
113. Chad, K. and Miller, P. Do Libraries Matter? The rise of Library 2.0, 2005. Talis: http://www.talis.com/downloads/white_papers/DoLibrariesMatter.pdf.
114. Chakrabarti, D., Kumar, R. and Punera, K., A graph-theoretic approach to webpage segmentation, in *17th International Conference on the World Wide Web*, (Beigin, China, 2008).
115. Chan, L.M., Comaromi, J.P., Mitchell, J.S. and Satija, M.P. *Dewey Decimal Classification: A Practical Guide*. Forest Press, Albany, 1996.
116. Chang, C.-C. and Garcia-Molina, H., Evaluating the Cost of Boolean Query Mapping, in *Second ACM International Conference on Digital Libraries*, (1997), ACM Press.
117. Chartier, R. *The order of books: readers, authors, and libraries in Europe between the fourteenth and eighteenth centuries*. Stanford University Press, Stanford, Calif., 1994.
118. Cheung, K., Hunter, J., Lashtabeg, A. and Drennan, J., SCOPE - A Scientific Compound Object Publishing and Editing System, in *3rd International Digital Curation Conference*, (Washington, D.C., 2007).
119. Chien, Y.T., Disruptive Technologies, Innovation, and Digital Libraries Research – The Case of a Billion-Dollar Business, in *DLKC'04*, (Tsukuba, Japan, 2004).

120. Choudhury, S., The Data Conservancy, in *CNI Spring Forum*, (Minneapolis, 2009).
121. Choudhury, S., DiLauro, T., Szalay, A., Vishniack, E. and Plante, R. Digital data preservation for scholarly publications in astronomy. *International Journal of Digital Curation*, 2 (2).
122. Christensen, C.M. *The innovator's dilemma: when new technologies cause great firms to fail*. Harvard Business School Press, Boston, Mass., 1997.
123. Christensen, C.M., Horn, M.B. and Johnson, C.W. *Disrupting class: how disruptive innovation will change the way the world learns*. McGraw-Hill, New York, 2008.
124. Christensen, E., Curbera, F., Meredith, G. and Weerawarana, S. Web Services Description Language (WSDL) 1.1. W3C, 2001. Available at <http://www.w3.org/TR/wsdl>.
125. Chu, W. Optimal File Allocation in Multiple Computer Systems. *IEEE Transactions on Computers*, 1969 (October).
126. Claburn, T. Web 2.0 Summit: President Elect Obama Typifies World 2.0. *InformationWeek*, 2008.
127. Clark, K.G., Feigenbaum, L. and Torres, E. SPARQL Protocol for RDF. W3C Recommendation. W3C, January 15, 2008. Available at <http://www.w3.org/TR/rdf-sparql-protocol/>.
128. Cleverdon, C.W. and Keen, E.M. Factors Determining the Performance of Indexing Systems, Vol. 1: Design. Aslib Cranfield Research Project, 1966.
129. Cleverdon, C.W. and Keen, E.M. Factors Determining the Performance of Indexing Systems, Vol. 2: Test Results. Aslib Cranfield Research Project, 1966.
130. Coffey, M. The future is smart machines (and soup), Retrieved August 21, 2009, from Making innovation flourish: <http://blogs.nesta.org.uk/innovation/2007/07/the-future-is-s.html>.

131. Cole, T.W. and Shreeves, S. Lessons Learned from the Illinois OAI Metadata Harvesting Project. in Hillmann, D. ed. *Metadata in Practice*, American Library Association, Chicago, 2004.
132. Collins, P., Shukla, S. and Redmiles, D. Activity Theory and System Design: A View from the Trenches. *Computer Supported Cooperative Work*, 11.
133. Collis, B. and Strijker, A. Technology and Human Issues in Reusing Learning. *Journal of Interactive Media in Education*, 4 (Special Issue on the Educational Semantic Web). Available at <http://www.jime.open.ac.uk/2004/4>.
134. Committee on Information Strategy for the Library of Congress *LC21: A Digital Strategy for the Library of Congress (2000)*. National Academy Press, Washington, DC, 2000.
135. Committee on the Fundamentals of Computer Science: Challenges and Opportunities *Computer Science: Reflections on the Field, Reflections from the Field*. Computer Science and Telecommunications Board, The National Academies Press, Washington, DC, 2004.
136. Conklin, J. Hypertext: an introduction and survey. *Computer* (20). 17-41.
137. Constantopoulos, P., Doerr, M., Theodoridou, M. and Tzobanakis, M. On Information Organization in Annotation Systems. in Grieser, G. and Tanaka, Y. eds. *Intuitive Human Interface 2004, LNAI3359*, Springer-Verlag, Berlin, 2004, 189-200.
138. Cormode, G. and Krishnamurthy, B. Key differences between Web 1.0 and Web 2.0. *first monday*, 13 (6). Available at <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2125/1972>.
139. Corporation for National Research Initiatives. CS-TR Project Introduction, 1995. Retrieved February 11, 2009, from <http://www.cnri.reston.va.us/describe.html>.
140. Crawford, W. and Gorman, M. *Future libraries: dreams, madness & reality*. American Library Association, Chicago, 1995.

141. Crockford, D. The application/json Media Type for JavaScript Object Notation (JSON). IETF, July, 2006. Available at <http://tools.ietf.org/html/rfc4627>.
142. Crofts, N., Doerr, M., Gill, T., Stead, S. and Stiff, M. Definition of the CIDOC object-oriented Conceptual Reference Model and Crossreference Manual. ICOM/CIDOC Documentation Standards Group, 2002. Available at http://cidoc.ics.forth.gr/crm_definition_documents/cidoc_crm_3.3.2/cidoc_crm_title.htm.
143. Crow, R. The Case for Institutional Repositories: A SPARC Position Paper, 2002. The Scholarly Publishing & Academic Resources Coalition: <http://www.arl.org/sparc/IR/ir.html>.
144. Csányi, V. *Evolutionary systems and society: a general theory of life, mind, and culture*. Duke University Press, Durham, 1989.
145. Csaey, M., Retrieved April 26, 2009, from LibraryCrunch: <http://librarycrunch.com/>.
146. Cutter, C.A. *Rules for a printed dictionary catalogue*. U.S. Government Printing Office, Washington., 1876.
147. Daniel Jr., R. and Lagoze, C., Distributed Active Relationships in the Warwick Framework, in *IEEE Metadata Conference*, (Bethesda, 1997).
148. Daniel Jr., R. and Lagoze, C. Extending the Warwick Framework: From Metadata Containers to Active Digital Objects. *D-Lib Magazine*, 1997 (November). Available at <http://www.dlib.org/dlib/november97/daniel/11daniel.html>.
149. David Hawking, J.Z. Does topic metadata help with Web search? *Journal of the American Society for Information Science and Technology*, 58 (5). 613-628.
150. David, S. *The Open World*. PhD thesis, Science, Technology, and Society, Cornell University. 2007.

151. Davis, J. and Lagoze, C. Dienst protocol version 5.0. 1997. Available at <http://www.cs.cornell.edu/lagoze/dienst/protocol5.htm>.
152. Davis, J.R., Krafft, D.B. and Lagoze, C. Dienst: Building a Production Technical Report Server. in *Advances in Digital Libraries*, Springer-Verlag, 1995, Chapter 15.
153. Davis, J.R. and Lagoze, C. NCSTRL: Design and Deployment of a Globally Distributed Digital Library. *Journal of the American Society for Information Science*, 51 (3). 273-280.
154. Dawkins, R. *The selfish gene*. Oxford University Press, Oxford ; New York, 2006.
155. Dean, M., Connolly, D., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F. and Stein, L.A. OWL Web Ontology Language 1.0 Reference. W3C Working Draft. 20020729. World Web Consortium, July 29, 2002. Available at <http://www.w3.org/TR/2002/WD-owl-ref-20020729/>.
156. Dempsey, L. Libraries and the Long Tail - Some Thoughts about Libraries in the Network Age. *D-Lib Magazine*, 12 (4). Available at <http://www.dlib.org/dlib/april06/dempsey/04dempsey.html#6>.
157. Dempsey, L. The Library Catalogue in the New Discovery Environment: Some Thoughts. *Ariadne*, 48 (July).
158. Dempsey, L. The library, the catalogue, the broker brokering access to information in the hybrid library. in Criddle, S., Dempsey, L. and Heseltine, R. eds. *Information Landscapes for a Learning Society*, Library Association, London, 1999.
159. Dempsey, L. and Weibel, S. The Warwick Metadata Workshop. *D-Lib Magazine*. Available at <http://www.dlib.org/dlib/july96/07weibel.html>.
160. Diekema, A. and Chen, J., Experimenting with the Automatic Assignment of Educational Standards to Digital Library Content, in *Joint Conference of Digital Libraries (JCDL)*, (Denver, 2005).

161. Digital Library Federation (DLF). The Making of America II Testbed Project White Paper. White Paper. Version 2.0. Digital Library Federation, September 15, 1998. Available at <http://sunsite.berkeley.edu/moa2>.
162. DiLauro, T., OAI-ORE for publishing workflows: Data archiving for journals of the American Astronomical Society, in *Open Repositories 2008*, (Southampton, UK, 2008).
163. Dmitriev, P. and Lagoze, C., Automatically Constructing Descriptive Site Maps, in *Eighth Asia Pacific Web Conference*, (Harbin, China, 2006).
164. Dmitriev, P., Lagoze, C. and Suchkov, B., As We May Perceive: Inferring Logical Documents from Hypertext, in *HT 2005 - Sixteenth ACM Conference on Hypertext and Hypermedia*, (Salzburg, Austria, 2005).
165. Doctorow, C. Metacrap: Putting the torch to seven straw-men of the meta-utopia, 2001. Retrieved February 20, 2009, from <http://www.well.com/~doctorow/metacrap.htm>.
166. Doerr, M., Hunter, J. and Lagoze, C. Towards a Core Ontology for Information Integration. *Journal of Digital Information*, 4 (1).
167. Downes, S. E-Learning 2.0. *eLearn*. Available at <http://www.elearnmag.org/subpage.cfm?section=articles&article=29-1>.
168. Drost, I. and Scheffer, T., Thwarting the Nigritude Ultramarine: Learning to Identify Link Spam, in *16th European Conference on Machine Learning*, (Porto, 2005).
169. Dushay, N. and French, J.C., Using Query Mediators for Distributed Searching in Federated Digital Libraries, in *Digital Libraries '99: The Fourth ACM Conference on Digital Libraries*, (Berkeley, California, 1999).
170. Dushay, N., French, J.C. and Lagoze, C. A Characterization Study of NCSTRL Distributed Searching. Technical Report. TR99-1725. Cornell University Computer Science, January, 1999. Available at <http://ecommons.library.cornell.edu/handle/1813/7379?mode=full>.

171. Dushay, N., French, J.C. and Lagoze, C., Predicting Indexer Performance in a Distributed Digital Library, in *Third European Conference on Research and Advanced Technology for Digital Libraries*, (Paris, France, 1999).
172. Dushay, N. and Hillmann, D., Analyzing Metadata for Effective Use and Re-Use, in *DCMI Metadata Conference and Workshop*, (Seattle, 2003).
173. Dushay, N. and Hillmann, D. NSDL Metadata Primer, 2005.
<http://metamanagement.comm.nsd.gov/outline.html>.
174. Edwards, P.N., Jackson, S.J., Bowker, G.C. and Knobel, C.P. Understanding Infrastructure: Dynamics, Tensions, and Design. National Science Foundation, January, 2007.
175. Eiron, N. and McCurley, K., Untangling compound documents on the web, in *Fourteenth ACM conference on Hypertext and Hypermedia*, (Nottingham, UK, 2003).
176. Embley, D., Jiang, Y. and Ng, Y., Record-boundary discovery in Web documents, in *1999 ACM SIGMOD International Conference on Management of Data*, (Philadelphia, PA, 1999).
177. Engelbart, D.C. A conceptual framework for the augmentation of man's intellect. in *Computer-supported cooperative work: a book of readings*, Morgan Kaufmann Publishers Inc., San Francisco, 1988, 35-65.
178. Engelbart, D.C., Knowledge-domain interoperability and an open hyperdocument system, in *Proceedings of the Conference on Computer-Supported Work*, (Los Angeles, 1990), 143-156.
179. Engeström, J. Activity Theory and Individual and Social Transformation. in Engeström, J., Miettinen, R. and Punamäki-Gitai, R.-L. eds. *Perspectives on Activity Theory*, Cambridge University Press, Cambridge, 1999.
180. Engeström, J. Activity theory as a framework for analyzing and redesigning work. *Ergonomics*, 43 (7). 960-974.

181. Engestrom, J. What makes a good social object?, Retrieved January 26, 2009, from Zengestrom: http://www.zengestrom.com/blog/objectcentered_sociality/.
182. Engestrom, J. Why some social network services work and others don't — Or: the case for object-centered sociality, Retrieved January 24, 2009, from Zengestrom: http://www.zengestrom.com/blog/2005/04/why_some_social.html.
183. Entlich, R., Garson, L., Lesk, M., Normore, L., Olsen, J. and Weibel, S. Making a Digital Library: The Contents of the CORE Project. *Transactions on Information Systems*, 15 (2). 103-123.
184. Faaborg, A. and Lagoze, C. Semantic Browsing. in *Lecture Notes in Computer Science*, Springer-Verlag, Trondheim, Norway, 2003, 70-81.
185. Fensel, D. *Spinning the semantic Web: bringing the World Wide Web to its full potential*. MIT Press, Cambridge, Mass., 2003.
186. Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P. and Berners-Lee, T. Hypertext Transfer Protocol -- HTTP/1.1. RFC. 2616. The Internet Society, June, 1999. Available at <http://www.ietf.org/rfc/rfc2616.txt>.
187. Fisher, R., Lugg, R. and Boese, K.C. Cataloging: how to take a business approach. *The Bottom Line: Managing Library Finances*, 17 (2). 50-54.
188. Flanagan, D. *JavaScript : the definitive guide*. O'Reilly, Sebastopol, CA, 2002.
189. Foster, I., Kesselman, C. and Tuecke, S. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. *International Journal of Supercomputer Applications*, 15 (3).
190. Foulonneau, M., Cole, T.W., Habing, T.G. and Shreeves, S., Using Collection Descriptions to Enhance an Aggregation of Harvested Item-Level Metadata, in *Joint Conference on Digital Libraries (JCDL 2005)*, (Denver, 2005), ACM.
191. Fox, E. (ed.), *Source Book on Digital Libraries*. National Science Foundation, 1993.

192. Fox, E., Akscyn, R.M., Furuta, R.K. and Leggett, J.J. Digital libraries. *Communications of the ACM*, 38 (4). 22-28.
193. Fox, G., Web 2.0 and Grids, in *Open Grid Forum*, (Chapel Hill, NC, 2007).
194. French, J.C., Powell, A.L. and Creighton III, W.R., Efficient Searching in Distributed Digital Libraries, in *ACM Digital Libraries '98*, (Pittsburgh, 1998), 283-284.
195. Friedlander, A. *Emerging infrastructure : the growth of railroads*. Corporation for National Research Initiatives, Reston, Va., 1995.
196. Friedlander, A. *"In God We Trust": All others pay Cash: Banking as an American infrastructure, 1800 to 1935*. Corporation for National Research Initiatives, Reston, VA, 1996.
197. Friedlander, A. *Natural monopoly and universal service : telephones and telegraphs in the U.S. communications infrastructure, 1837-1940*. Corporation for National Research Initiatives, Reston, Va., 1995.
198. Friedlander, A. *Power and light : electricity in the U.S. energy infrastructure, 1870-1940*. Corporation for National Research Initiatives, Reston, Va., 1996.
199. Fuchs, C. *Internet and Society. Social Theory in the Information Age*. Routledge, New York, 2008.
200. Fuchs, C. and Gasse, S.H. The Self-Organization of Virtual Communities. *Journal of New Communications Research*.
201. Furie, B. *Understanding MARC Bibliographic: Machine-Readable Cataloging*. Cataloging Distribution Office, Library of Congress, Washington DC, 1998.
202. Garfield, E. *Citation indexing: Its theory and application in science, technology, and humanities*. John Wiley, New York, NY, 1979.

203. Garrett, J.J. Ajax: A New Approach to Web Applications, 2005. Retrieved March 6, 2009, from <http://www.adaptivepath.com/publications/essays/archives/000385.php>.
204. Garson, G.D. Actor-Network Theory, 2008. Retrieved May 26, 2009, from <http://faculty.chass.ncsu.edu/garson/PA765/actornetwork.htm>.
205. Gay, G. and Hembrooke, H. *Activity Centered Design*. MIT Press, Cambridge, MA, 2004.
206. Gee, J.P. *Situated Language and Learning: A Critique of Traditional Schooling*. Routledge, New York, 2004.
207. Gee, J.P. *What Video Games Can Teach Us About Literacy and Learning*. Palgrave-McMillan, New York, 2003.
208. Gerber, A. and Hunter, J. LORE: A Compound Object Authoring and Publishing Tool for the Australian Literature Studies Community. in *Digital Libraries: Universal and Ubiquitous Access to Information*, Springer Berlin, Heidelberg, 2008.
209. Giles, J. Special Report: Internet encyclopaedias go head to head. *Nature*, 438.
210. Ginsparg, P. Next-Generation Implications of Open Access. *CTWatch Quarterly*, 3 (3).
211. Gladney, H.M., Fox, E., Ahmed, Z., Ashany, R., Belkin, N. and Zemankova, M., Digital Library: Gross Structure and Requirements: Report from a March 1994 Workshop, in *DL'94*, (College Station, 1994), IEEE.
212. Gordon, M. and Pathak, P. Finding information on the World Wide Web: the retrieval effectiveness of search engines. *Information Processing & Management*, 35 (2). 141-180.
213. Gorman, M. *The concise AACR2, 1988 revision*. American Library Association, Chicago, 1989.

214. Gravano, L., Chang, C.-C., Garcia-Molina, H. and Paepcke, A., STARTS: Stanford Proposal for Internet Meta-Searching, in *ACM SIGMOD International Conference on Management of Data*, (1997).
215. Gravano, L., Chang, K., Garcia-Molina, H., Lagoze, C. and Paepcke, A. STARTS: Stanford Protocol Proposal for Internet Retrieval and Search. Digital Library Project Working Paper. Stanford University, January, 1997.
216. Gravano, L., Garcia-Molina, H. and Tomasic, A., The Effectiveness of GLOSS for the Text-Database Discovery Problem, in *ACM SIGMOD International Conference on The Management of Data*, (1994), ACM Press.
217. Greenberg, D. Camel drivers and gatecrashers: quality control in the digital research library. in Hawkins, B.L. and Battin, P. eds. *The mirage of continuity: reconfiguring academic information resources for the 21st century*, Council on Library and Information Resources and the Association of American Universities, Washington, DC, 1998, 105-116.
218. Gregorio, J. and de Hora, B. The Atom Publishing Protocol. RFC 5023. IETF, 2007. Available at <http://www.ietf.org/rfc/rfc5023.txt>.
219. Griffin, S.M. Funding for Digital Libraries Research: Past and Present. *D-Lib Magazine*, 11 (7/8).
220. Griffiths, J.M. Why the Web is not a library. in Hawkins, B.L. and Battin, P. eds. *The mirage of continuity: reconfiguring academic information resources for the 21st century*, Council on Library and Information Resources and the Association of American Universities, Washington, DC, 1998, 229-246.
221. Guy, M. and Tonkin, E. Folksonomies: Tidying up Tags? *D-Lib Magazine*, 12 (1). Available at <http://dlib.org/dlib/january06/guy/01guy.html>.
222. Gyongyi, Z. and Garcia-Molina, H., Web Spam Taxonomy, in *First International Workshop on Adversarial Information Retrieval on the Web*, (Chiba, Japan, 2005).
223. Haase, P., Broekstra, Egerhart, A. and Volz, R., A comparison of RDF query languages, in *Third International Semantic Web Conference*, (Hiroshima, Japan, 2004).

224. Hagedorn, K. OAIster: a "no dead ends" OAI service provider. *Library Hi Tech*, 21 (2). 170-181.
225. Halbert, M., Kaczmarek, J. and Hagedorn, K., Findings from the Mellon Metadata Harvesting Initiative, in *Research and Advanced Technology for Digital Libraries: 7th European Conference, ECDL 2003*, (Trondheim, Norway, 2003), Springer-Verlag.
226. Halpern, J.Y. and Lagoze, C., The Computing Research Repository: Promoting the Rapid Dissemination and Archiving of Computer Science Research, in *Digital Libraries '99, The Fourth ACM Conference on Digital Libraries*, (Berkeley, CA, 1999).
227. Hannay, T. Web 2.0 in Science. *CTWatch Quarterly*, 2007 (August).
228. Hanseth, O. and Monteiro, E. Understanding Information Infrastructure, 1998. Retrieved May 26, 2009, from <http://heim.ifi.uio.no/~oleha/Publications/bok.html>.
229. Harnad, S. The PostGutenberg Galaxy: How to Get There from Here. *The Information Society*, 11 (4). 285-291.
230. Harrison, T.L., Elango, A., Bollen, J. and Nelson, M. Initial Experiences Re-exporting Duplicate and Similarity Computations with an OAI-PMH Aggregator. arXiv Report. cs.DL/0401001. Computer Science Department, Old Dominion University, January 5, 2004.
231. Heath, C., Knoblauch, H. and Luff, P. Technology and social interaction: the emergence of "workplace studies". *British Journal of Sociology*, 51 (2).
232. Hendler, J., Shadbot, N., Hall, W., Berners-Lee, T. and Wietzner, D. Web science: an interdisciplinary approach to understanding the web. *Communications of the ACM*, 51 (7). 60-69.
233. Hey, T. and Trefethen, A.E. Cyberinfrastructure for e-Science. *Science*, 308 (5723). 817-821. Available at <http://www.sciencemag.org/cgi/content/abstract/308/5723/817>.

234. Hey, T., Waters, D.J., Lynch, C.A., Van de Sompel, H. and Lagoze, C., Augmenting Interoperability Across Scholarly Repositories, in *JCDL 2006*, (Chapel Hill, NC, 2006), ACM/IEEE.
235. Hillmann, D., Dushay, N. and Phipps, J., Improving Metadata Quality: Augmentation and Recombination, in *DC-2004*, (Shanghai, China, 2004). Available at http://students.washington.edu/jtennis/dcconf/Paper_21.pdf.
236. Hitchcock, S., Bergmark, D., Brody, T., Gutteridge, C., Carr, L., Hall, W., Lagoze, C. and Harnad, S. Open Citation Linking: The Way Forward. *D-Lib Magazine*, 8 (10).
237. Hoffman, M.M., O'Gorman, L., Story, G.A., Arnold, J.Q. and Macdonald, N.H. The RightPages Service: An image-based electronic library. *Journal of the American Society for Information Science*, 44 (8). 446-452. Available at [http://dx.doi.org/10.1002/\(SICI\)1097-4571\(199309\)44:8<446::AID-ASI3>3.0.CO;2-9](http://dx.doi.org/10.1002/(SICI)1097-4571(199309)44:8<446::AID-ASI3>3.0.CO;2-9).
238. Hoogerwerf, M., Durable enhanced publication, in *African Digital Scholarship & Curation*, (2009).
239. Hooker, B. The Future of Science is Open, Part 1: Open Access, Retrieved April 6, 2009, from 3quarksdaily: http://3quarksdaily.blogspot.com/3quarksdaily/2006/10/the_future_of_s_1.html.
240. Hooker, B. The Future of Science is Open, Part 2: Open Science, Retrieved April 6, 2009, from 3quarksdaily: http://3quarksdaily.blogspot.com/3quarksdaily/2006/11/the_future_of_s.html.
241. Hooker, B. The Future of Science is Open, Part 3: An Open Science World, Retrieved April 6, 2009, from 3quarksdaily: http://3quarksdaily.blogspot.com/3quarksdaily/2007/03/the_future_of_s.html.
242. Hunter, J. and Cheung, K. Provenance explorer - a graphical interface for constructing scientific publication packages from provenance trails. *International Journal of Digital Libraries*, 7 (1-2). 99-107.

243. Huynh, D., Mazzocchi, S. and Karger, D., Piggy Bank: Experience the Semantic Web Inside Your Web Browser, in *International Semantic Web Conference (ISWC)*, (2005).
244. Infoseek, Document retrieval over networks wherein ranking and relevance scores are computed at the client for multiple database documents, Patent Issued by United States, Number 5659732.
245. Iverson, V., Song, Y.-W., Van de Walle, R., Rowe, M., Doim Chang, Santos, E. and Schwartz, T. MPEG-21 Digital Item Declaration. ISO/IEC JTC 1/SC 29/WG 11 N3971. International Organization for Standardization, 2000. Available at <http://xml.coverpages.org/MPEG21-WG-11-N3971-200103.pdf>.
246. Jacobs, I. and Walsh, N. Architecture of the World Wide Web. Proposed Recommendation. W3C, April, 2004. Available at <http://www.w3.org/TR/2004/PR-webarch-20041105/>.
247. Jacobs, N. *Open access: key strategic, technical and economic aspects*. Chandos, Oxford, 2006.
248. Jantz, R. and Giarlo, M.J. Digital Preservation: Architecture and Technology for Trusted Digital Repositories. *D-Lib Magazine*, 11 (6). Available at <http://www.dlib.org/dlib/june05/jantz/06jantz.html>.
249. Jenkins, H. *Convergence culture: where old and new media collide*. New York University Press, New York, 2006.
250. Jenkins, H. Fans, bloggers, and gamers exploring participatory culture, New York University Press, New York, 2006, vi, 279 p.
251. Jenkins, H., Clinton, K., Purushotma, R., Robinson, A.J. and Weigel, M. Confronting the challenges of participatory culture: Media education for the 21st century. MacArthur Foundation, 2006. Available at http://digitalllearning.macfound.org/atf/cf/%7B7E45C7E0-A3E0-4B89-AC9C-E807E1B0AE4E%7D/JENKINS_WHITE_PAPER.PDF.
252. Johansen, R. *Teleconferencing and beyond: communications in the office of the future*. McGraw-Hill, New York, N.Y., 1984.

253. Johnson, R.K. Institutional Repositories: Partnering with Faculty to Enhance Scholarly Communication. *D-Lib Magazine*, 8 (11). Available at <http://www.dlib.org/dlib/november02/johnson/11johnson.html>.
254. Kahan, J., Koivunen, M.-R., Prud'Hommeaux, E. and Swick, R.R., Annotea: An Open RDF Infrastructure for Shared Web Annotations, in *WWW10*, (Hong Kong, 2001). Available at <http://www10.org/cdrom/papers/488/index.html>.
255. Kahn, R. and Wilensky, R. A Framework for Distributed Digital Object Services. Corporation for National Research Initiatives, 1995. Available at <http://www.cnri.reston.va.us/k-w.html>.
256. Kahn, R.E. and Cerf, V.G. The Digital Library Project, volume I: The world of Knowbots. Corporation for National Research Initiatives, 1988.
257. Katz, R. and Chien, Y.T. Information Infrastructure Technology And Applications. HPCC National Coordination Office, IITA Interagency Task Group, 1994.
258. Katz, R., editor *The tower and the cloud: higher education in the age of cloud computing*. EDUCAUSE, Boulder, 2008.
259. Katz, R.N. and Gandel, P.B. The Tower, the Cloud, and Posterity. in Katz, R.N. ed. *The Tower and The Cloud*, EDUCAUSE, Inc., Boulder, CO, 2008.
260. Keller, M.A., Reich, V. and Herkovic, A.C. What is a library anymore, anyway? *first monday*, 2003 (May 5). Available at http://www.firstmonday.org/issues/issue8_5/keller/index.html.
261. Kelly, K. Scan This Book! *New York Times Magazine*. May 14, 2006. Available at http://www.kk.org/writings/scan_this_book.php.
262. Khoo, M., A Sociotechnical Framework for Evaluating a Large-Scale Distributed Educational Digital Library, in *ECDL 2006*, (Alicante, Spain, 2006), 449-452.

263. Kilker, J.A. and Gay, G. The Social Construction of a Digital Library: A Case Study Examining Implications for Evaluation. *Information Technology and Libraries*, 17 (2). 60-70.
264. Kim, K.S. Recent Work in Cataloging and Classification, 2000-2002. *Library Resources and Technical Services*, 47 (3).
265. King, S. Bookstores fight to survive latest plot twists. *Kansas City Star*. May 18, 2009.
266. Kittur, A., Chi, E., Pendleton, B.A., Suh, B. and Mytkowicz, T., Power of the Few vs. Wisdom of the Crowd: Wikipedia and the Rise of the Bourgeoisie, in *25th Annual ACM Conference on Human Factors in Computing Systems (CHI 2007)*, (San Jose, CA, 2007).
267. Kleinberg, J., Authoritative sources in a hyperlinked environment, in *Proc. 9th ACM-SIAM Symposium on Discrete Algorithms*, (1998).
268. Kleinberg, J. The Convergence of Social and Technological Networks. *Communications of the ACM*, 51 (11).
269. Kling, R. Information Technologies and the Strategic Reconfiguration of Libraries in Communication Networks. WP-00-04. Center for Social Informatics, Indiana University, 2000.
270. Kling, R. Learning about information technologies and social change: The contributions of social informatics. *The Information Society*, 16. 217-232.
271. Kling, R. What is Social Informatics and Why Does it Matter? *D-Lib Magazine*, 5 (1). Available at <http://www.dlib.org/dlib/january99/kling/01kling.html>.
272. Kling, R., Spector, L.B. and Fortuna, J. The real stakes of virtual publishing: The transformation of E-Biomed into PubMed central. *Journal of the American Society for Information Science and Technology*, 55 (2). 127-148.

273. Klyne, G. and Carroll, J.J. Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation. 20040210. W3C, February, 2004. Available at <http://www.w3.org/TR/rdf-concepts/>.
274. Koehler, W. Digital libraries and World Wide Web sites and page persistence. *Information Research*, 4 (4).
275. Krafft, D., Birkland, A. and Cramer, E., NCore: architecture and implementation of a flexible, collaborative digital library., in *JCDL 2008*, (Pittsburgh, 2008), ACM.
276. Krichel, T. Access to Scientific Literature on the WWW: the RePEc concept, 1998. Retrieved June 17, 2009, from <http://openlib.org/home/krichel/osborne.html>.
277. Kruk, S.R., Decker, S. and Zieborak, L. JeromeDL - Adding Semantic Web Technologies to Digital Libraries. in *Database and Expert Systems Applications*, Springer Berlin, Heideberg, 2005.
278. Kruk, S.R. and McDaniel, B. *Semantic digital libraries*. Springer, Berlin, 2009.
279. Kumar, A., Saigal, R., Chavez, R. and Schwerner, N., Architecting an extensible digital repository, in *4th ACM/IEEE-CS joint conference on Digital libraries*, (Tucson, AZ, 2004), AZ. Available at <http://doi.acm.org/10.1145/996350.996354>.
280. Kurtz, M.J., Eichhorn, G., Accomazzi, A., Grant, C.S., Demleitner, M., Murray, S.S., Martimbeau, N. and Elwell, B. The NASA Astrophysics Data System: Sociology, Bibliometrics, and Impact. *Journal of the American Society for Information Science and Technology*, 2003 (March).
281. Kuutti, K., The concept of activity a basic unit of analysis for CSCW research, in *ECSCW '91*, (Amsterdam, 1991).
282. Lagoze, C., Accommodating Simplicity and Complexity in Metadata: Lessons from the Dublin Core Experience, in *Seminar on Metadata*, (Archiefschool, Netherlands Institute for Archival Education and Research, The Hague, 2000).

283. Lagoze, C., Business Unusual; How "event awareness" may breathe life into the catalog, in *Bicentennial Conference on Bibliographic Control in the New Millennium*, (Library of Congress, Washington DC, 2000). Available at http://lcweb.loc.gov/catdir/bibcontrol/lagoze_paper.html.
284. Lagoze, C. From Static to Dynamic Surrogates: Resource Discovery in the Digital Age. *D-Lib Magazine*. Available at <http://www.dlib.org/dlib/june97/06lagoze.html>.
285. Lagoze, C., The oreChem Project: Integrating Chemistry Scholarship with the Semantic Web, in *WebSci'09: Society On-Line*, (Athens, 2009).
286. Lagoze, C. The Warwick Framework: A Container Architecture for Diverse Sets of Metadata. *D-Lib Magazine*, 2 (7/8). Available at <http://www.dlib.org/dlib/july96/lagoze/07lagoze.html>.
287. Lagoze, C., Arms, W., Gan, S., Hillmann, D., Ingram, C., Krafft, D., Marisa, R., Phipps, J., Saylor, J., Terrizzi, C., Hoehn, W., Millman, D., Allan, J., Guzman-Lara, S. and Kalt, T., Core Services in the Architecture of the National Digital Library for Science Education (NSDL), in *Joint Conference on Digital Libraries*, (Portland, Oregon, 2002), ACM/IEEE. Available at <http://arxiv.org/abs/cs.DL/0201025>.
288. Lagoze, C., Arms, W., Gan, S., Hillmann, D., Ingram, C., Krafft, D., Marisa, R., Phipps, J., Saylor, J., Terrizzi, C., Hoehn, W., Millman, D., Allan, J., Guzman-Lara, S. and Kalt, T. Core Services in the Architecture of the National Digital Library for Science Education (NSDL). arXiv Report. cs.DL/0201025. Cornell University, January 29, 2002. Available at <http://arxiv.org/abs/cs.DL/0201025>.
289. Lagoze, C. and Davis, J.R. Dienst - An Architecture for Distributed Document Libraries. *Communications of the ACM*, 38 (4). 47.
290. Lagoze, C. and Ely, D. Implementation Issues in an Open Architectural Framework for Digital Object Services. Computer Science Technical Report. TR95-1540. Cornell University, September, 1995.

291. Lagoze, C., Fielding, D. and Payette, S., Making Global Digital Libraries Work: Collection Service, Connectivity Regions, and Collection Views, in *ACM Digital Libraries '98*, (Pittsburgh, 1998).
292. Lagoze, C. and Hunter, J. The ABC Ontology and Model. *Journal of Digital Information*, 2 (2). Available at <http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Lagoze/>.
293. Lagoze, C., Hunter, J. and Brickley, D. An Event-Aware Model for Metadata Interoperability. Cornell Computer Science Technical Report. TR2000-1801. Cornell University, June 30, 2000. Available at http://archive.dstc.edu.au/RDU/staff/jane-hunter/harmony/harmony_ECDL2000.zip.
294. Lagoze, C., Hunter, J. and Brickley, D., An Event-Aware Model for Metadata Interoperability, in *ECDL 2000*, (Lisbon, 2000). Available at http://archive.dstc.edu.au/RDU/staff/jane-hunter/harmony/harmony_ECDL2000.zip.
295. Lagoze, C. and Kenney, A. The Prism Project: Vision and Focus. Prism Working Paper. Cornell University, February 7., 2000. Available at <http://www.cs.cornell.edu/prism/Publications/WorkingPapers/Visions.htm>.
296. Lagoze, C., Krafft, D., Cornwell, T., Dushay, N., Eckstrom, D. and Saylor, J., Metadata aggregation and "automated digital libraries": A retrospective on the NSDL experience, in *Joint Conference on Digital Libraries*, (Chapel Hill, NC, 2006), ACM. Available at <http://arxiv.org/abs/cs.DL/0601125>.
297. Lagoze, C., Krafft, D., Cornwell, T., Eckstrom, D., Jesuroga, S. and Wilper, C., Representing Contextualized Information in the NSDL, in *ECDL2006*, (Alicante, Spain, 2006), Springer.
298. Lagoze, C., Krafft, D., Jesuroga, S., Cornwell, T., Cramer, E. and Shin, E. An Information Network Overlay Architecture for the NSDL. arXiv/CoRR Report. cs.DL/0501080. Cornell University, 2005. Available at <http://arxiv.org/abs/cs.DL/0501080>.
299. Lagoze, C., Krafft, D.B., Payette, S. and Jesuroga, S. What Is a Digital Library Anymore, Anyway? Beyond Search and Access in the NSDL. *D-Lib*

- Magazine*, 11 (11). Available at <http://dx.doi.org/10.1045%2Fnovember2005-lagoze>.
300. Lagoze, C., Lynch, C.A. and Jr., R.D. The Warwick Framework: A Container Architecture for Aggregating Sets of Metadata. Technical Report. TR96-1593. Cornell University Computer Science, June, 1996. Available at <http://cs-tr.cs.cornell.edu:80/Dienst/UI/2.0/Describe/ncstrl.cornell/TR96-1593?abstract=>.
 301. Lagoze, C., Payette, S., Shin, E. and Wilper, C. Fedora: An Architecture for Complex Objects and their Relationships. *International Journal of Digital Libraries*, 6 (2). 124-138. Available at <http://arxiv.org/abs/cs.DL/0501012>.
 302. Lagoze, C., Shaw, E., Davis, J.R. and Krafft, D.B. Dienst Implementation Reference Manual. Technical Report. TR95-1514. Cornell University Computer Science, May, 1995. Available at <http://portal.acm.org/citation.cfm?id=866787>.
 303. Lagoze, C. and Van de Sompel, H., The Open Archives Initiative: Building a low-barrier interoperability framework, in *Joint Conference on Digital Libraries*, (Roanoke, VA, 2001). Available at <http://www.openarchives.org/documents/oai.pdf>.
 304. Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R. and Warner, S. Abstract Data Model. Open Archives Initiative, 2008. Available at <http://www.openarchives.org/ore/1.0/datamodel>.
 305. Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R. and Warner, S. HTTP Implementation. Open Archives Initiative, 2008. Available at <http://www.openarchives.org/ore/1.0/http>.
 306. Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R. and Warner, S. ORE Specification and User Guide - Table of Contents. Open Archives Initiative, 2008. Available at <http://www.openarchives.org/ore/1.0/toc>.
 307. Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R. and Warner, S. ORE User Guide - Primer. Open Archives Initiative, 2008. Available at <http://www.openarchives.org/ore/1.0/primer>.

308. Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R. and Warner, S. Resource Map Discovery. Open Archives Initiative, 2008. Available at <http://www.openarchives.org/ore/1.0/discovery>.
309. Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R. and Warner, S. Resource Map Implementation in Atom. Open Archives Initiative, 2008. Available at <http://www.openarchives.org/ore/1.0/atom>.
310. Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R. and Warner, S. Resource Map Implementation in RDF/XML. Open Archives Initiative, 2008. Available at <http://www.openarchives.org/ore/1.0/rdfxml>.
311. Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R. and Warner, S. Resource Map Implementation in RDFa. Open Archives Initiative, 2008. Available at <http://www.openarchives.org/ore/1.0/rdfa>.
312. Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R. and Warner, S. Vocabulary. Open Archives Initiative, 2008. Available at <http://www.openarchives.org/ore/1.0/vocabulary>.
313. Lagoze, C., Van de Sompel, H., Nelson, M. and Warner, S. Guidelines for Aggregators, Caches and Proxies. Open Archives Initiative, 2002. Available at <http://www.openarchives.org/OAI/2.0/guidelines-aggregator.htm>.
314. Lagoze, C., Van de Sompel, H., Nelson, M. and Warner, S. The Open Archives Initiative Protocol for Metadata Harvesting - Version 2.0. Open Archives Initiative, 2002. Available at http://www.openarchives.org/OAI_protocol/openarchivesprotocol.html.
315. Lagoze, C., Van de Sompel, H., Nelson, M., Warner, S., Sanderson, R. and Johnston, P. A Web-Based Resource Model for Scholarship 2.0: Object Reuse and Exchange. *Concurrency and Computation: Practice and Experience* (Special Issue - Success in Furthering Scientific Discovery).
316. Lagoze, C., Van de Sompel, H., Nelson, M., Warner, S., Sanderson, R. and Johnston, P. Object Re-Use & Exchange: A Resource-Centric Approach. arXiv. 2008. Available at <http://arxiv.org/abs/0804.2273>.

317. Lakoff, G. *Women, fire, and dangerous things : what categories reveal about the mind*. University of Chicago Press, Chicago, 1987.
318. Lamb, R., Sawyer, S. and Kling, R., A Social Informatics Perspective on Socio-Technical Networks, in *Americas Conference on Information Systems*, (Honolulu, 2000).
319. Landler, M. and Stelter, B. With a Hint to Twitter, Washington Taps Into a Potent New Force in Diplomacy. *New York Times*. June 16, 2009.
320. Large, D. Is the Internet a threat to our core revenue streams? *CTAM Quarterly Journal*, 46.
321. Larsen, L.L. Information Literacy: The Web is not an Encyclopedia. Office of Information Technology, University of Maryland, April, 2006. Available at <http://www.oit.umd.edu/units/web/literacy/>.
322. Lasher, R. and Cohen, D. A Format for Bibliographic Records. RFC. 1807. Internet Engineering Task Force, June, 1995. Available at <http://www.ietf.org/rfc/rfc1807.txt>.
323. Lassila, O. and Swick, R.R. Resource Description Framework: (RDF) Model and Syntax Specification. W3C Proposed Recommendation. PR-rdf-syntax-19990105. World Wide Web Consortium, January, 1999. Available at <http://www.w3.org/TR/PR-rdf-syntax/>.
324. Latour, B. *Pandora's Hope: Essays on the Reality of Science Studies*. Harvard University Press, Cambridge, 1999.
325. Latour, B. *Reassembling the social: an introduction to actor-network-theory*. Oxford University Press, Oxford ; New York, 2005.
326. Latour, B. *Science in action: how to follow scientists and engineers through society*. Harvard University Press, Cambridge, Mass., 1987.
327. Latour, B. Where are the Missing Masses? The Sociology of a Few Mundane Artifacts. in Bijker, W.E. and Law, J. eds. *Shaping Technology/Building*

- Society: Studies in Technological Change*, MIT Press, Cambridge, MA, 1992, 225-258.
328. Law, J. and Hassard, J. *Actor network theory and after*. Blackwell/Sociological Review, Oxford [England] ; Malden, MA, 1999.
329. Le Boeuf, P. *Functional requirements for bibliographic records (FRBR): hype or cure-all?* Haworth Information Press, Binghamton, NY, 2005.
330. Leiner, B.M. The NCSTRL Approach to Open Architecture for the Confederated Digital Library. *D-Lib Magazine*, December. Available at <http://www.dlib.org/dlib/december98/leiner/12leiner.html>.
331. Leiner, B.M., Cerf, V.G., Clark, D.D., Kahn, R.E., Kleinrock, L., Lynch, D.C., Postel, J., Roberts, L.G. and Wolff, S. A Brief History of the Internet. Internet Society, 2000. Available at <http://www.isoc.org/internet/history/brief.shtml>.
332. Lerner, F.A. *The story of libraries: from the invention of writing to the computer age*. Continuum, New York, 1998.
333. Lesk, M. *Practical digital libraries: books, bytes, and bucks*. Morgan Kaufmann Publishers, San Francisco, Calif., 1997.
334. Leskovec, J., Backstrom, L. and Kleinberg, J., Meme-tracking and the Dynamics of the News Cycle, in *15th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, (2009).
335. Lessig, L. *Free culture : how big media uses technology and the law to lock down culture and control creativity*. Penguin Press, New York, 2004.
336. Levy, D., Cataloging in the Digital Order, in *The Second Annual Conference on the Theory and Practice of Digital Libraries*, (1995). Available at <http://www.csdl.tamu.edu/DL95/papers/levy/levy.html>.
337. Levy, D. Digital Libraries and the Problem of Purpose. *Bulletin of the American Society for Information Science*, 26 (6).

338. Levy, D. Documents and Libraries: A Sociotechnical Perspective. in Bishop, A.P., Van House, N.A. and Bittenfield, B.P. eds. *Digital Library Use*, MIT Press, Cambridge, MA, 2003.
339. Levy, D.M., Fixed or Fluid? Document Stability and New Media, in *1994 European Conference on Hypermedia Technology*, (1994), ACM Press.
340. Levy, D.M. *Scrolling forward: making sense of documents in the digital age*. Arcade Publishers, New York, 2001.
341. Levy, D.M. and Marshall, C.C. Going Digital: A look at assumptions underlying digital libraries. *Communications of the ACM*, 38 (4). 77-84.
342. Levy, P. *Collective Intelligence*. Plenum, New York, 1997.
343. Li, C. and Bernoff, J. *Groundswell: winning in a world transformed by social technologies*. Harvard Business Press, Boston, Mass., 2008.
344. Li, W., Kolak, O., Vu, Q. and Tokano, H., Defining logical domains in a web site, in *Eleventh ACM Conference on Hypertext and Hypermedia*, (San Antonio, TX, 2003).
345. Library of Congress. METS: An Overview & Tutorial, 2004. <http://www.loc.gov/standards/mets/METSOverview.v2.html>.
346. Licklider, J.C.R., Council on Library Resources and Bolt Beranek and Newman inc. Cambridge Mass *Libraries of the future*. M.I.T. Press, Cambridge, MA, 1965.
347. Liu, X., Maly, K., Zubair, M. and Nelson, M., DP9: An OAI Gateway Service for Web Crawlers, in *Joint Conference on Digital Libraries*, (Portland, Oregon, 2002).
348. Liu, Y., Bai, K., Mitra, P. and Giles, C.L., TableSeer: automatic table metadata extraction and searching in digital libraries, in *7th ACM/IEEE-CS joint conference on Digital libraries*, (Vancouver, B.C., 2007).

349. Lyman, P. What is a Digital Library? Technology, Intellectual Property, and the Public Interest. in Graubard, S.R. and LeClerc, P. eds. *Books, bricks & bytes: libraries in the twenty-first century*, Transaction, New Brunswick, NJ, 1998.
350. Lynch, C.A. Colliding with the real world: Heresies and unexplored questions about audience, economics, and control of digital libraries. in Bishop, A.P., Van House, N.A. and Battenfield, B.P. eds. *Digital Library Use*, MIT Press, Cambridge, 2003.
351. Lynch, C.A. Institutional repositories: essential infrastructure for scholarship in the digital age. *ARL : A Bimonthly Report* (226).
352. Lynch, C.A. The Z39.50 Information Retrieval Standard; Part I: A Strategic View of Its Past, Present and Future. *D-Lib Magazine*, April. Available at <http://www.dlib.org/dlib/april97/04lynch.html>.
353. Lynch, C.A. Where Do We Go From Here? *D-Lib Magazine*, 11 (7/8). Available at <http://www.dlib.org/dlib/july05/lynch/07lynch.html>.
354. Lynch, C.A. and Garcia-Molina, H. Interoperability, Scaling, and the Digital Libraries Research Agenda: Information Infrastructure Technology and Applications Workshop on Digital Libraries. August 22, 1995.
355. MacKenzie, D.A. and Wajcman, J. *The social shaping of technology*. Open University Press, Buckingham England ; Philadelphia, 1999.
356. Maly, K., French, J., Selman, A. and Fox, E. Wide Area Technical Report Service. Technical Report. 94-13. Old Dominion University, June, 1994.
357. Maly, K., Nelson, M.L. and Zubair, M. Smart Objects, Dumb Archives: A User-Centric, Layered Digital Library Framework. *D-Lib Magazine*, 1999 (March). Available at <http://www.dlib.org/dlib/march99/maly/03maly.html>.
358. Maness, J.M. Library 2.0 Theory: Web 2.0 and its Implications for Libraries. *Webology*, 3 (2).

359. Markoff, J. A free and simple computer link. *New York Times*. December 8, 1993.
360. Markoff, J. Entrepreneurs See a Web Guided by Common Sense. *New York Times*. November 12, 2006.
361. Marshall, C.C., Annotation: from paper books to the digital library, in *Digital Libraries '97*, (1997), ACM Press.
362. Marshall, C.C. Finding the Boundaries of the Library without Walls. in Bishop, A.P., Bittenfield, B.P. and Van House, N.A. eds. *Digital Library Use: Social Practice in Design and Evaluation*, MIT Press, Cambridge, 2003.
363. Martin, K. Learning in Context. *Issues of Teaching and Learning*, 1998 (September). Available at <http://www.csd.uwa.edu.au/newsletter/issue0898/learning.html>.
364. Mason, J. and Sutton, S. Education Working Group: Draft Proposal. Dublin Core Metadata Initiative, 2000. Available at <http://dublincore.org/documents/2000/10/05/education-namespace/>.
365. Mathes, A. Folksonomies - Cooperative Classification and Communication Through Shared Metadata. Graduate School of Library and Information Science, University of Illinois Urbana-Champaign, 2004. Available at <http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html>.
366. McCalla, G. The Ecological Approach to the Design of E-Learning Environments: Purpose-based Capture and Use of the Information about Learners. *Journal of Interactive Media in Education*, 7 (Special Issue on the Educational Semantic Web). Available at <http://www.jime.open.ac.uk/2004/7>.
367. McDonough, J.P. METS: Standardized encoding for digital library objects. *International Journal of Digital Libraries*, 6 (2). 148-158.
368. McMartin, F. and Terada, Y., Digital Library Services for Authors of Learning Materials, in *ACM/IEEE Joint Conference on Digital Libraries (JCDL '02)*, (Portland, OR, 2002), ACM Press, 117-118.

369. McNee, S.M., Riedl, J. and Konstan, J.A., Making Recommendations Better: An Analytic Model for Human-Recommender Interaction, in *ACM SIGCHI Conference on Human Factors in Computing Systems*, (Montreal, Quebec, 2007), ACM, 1103-1108.
370. Mervis, J. NSF Rethinks Its Digital Library. *Science*, 323 (5910). 54-58.
371. Milgram, S. The small world problem. *Psychology Today*, 2. 60-67.
372. Miller, P. Coming Together around Library 2.0. *D-Lib Magazine*, 12 (4).
373. Miller, P. Web 2.0: Building the New Library. *Ariadne*, 45 (October). Available at <http://www.ariadne.ac.uk/issue45/miller/>.
374. Mitchell, S., Mooney, M., Mason, J., Paynter, G.W., Ruschinski, J., Kedzierski, A. and Humphreys, K. iVia Open Source Virtual Library System. *D-Lib Magazine*, 9 (1). Available at <http://www.dlib.org/dlib/january03/mitchell/01mitchell.html>.
375. Morris, C.W. *Foundations of the theory of signs*. The University of Chicago Press, Chicago, 1938.
376. Motta, E., Shum, S.B. and Domingue, J. ScholOnto: an ontology-based digital library server for research documents and discourse. *International Journal on Digital Libraries*, 3 (3).
377. Nardi, B.A. Some Reflections on the Application Of Activity Theory. in Nardi, B.A. ed. *Context and Consciousness: Activity and Human-Computer Interaction*, MIT press, Cambridge, 1996.
378. Nardi, B.A. and O'Day, V. *Information ecologies: using technology with heart*. MIT Press, Cambridge, Mass., 1999.
379. National Science Foundation. Digital Libraries Initiative Phase One, Retrieved February 10, 2009, from <http://www.dli2.nsf.gov/dlione/>.

380. National Science Foundation. DLI2 - Digital Libraries Initiative Phase 2, Retrieved February 10, 2009, from <http://dli2.nsf.gov/>.
381. National Science Foundation Cyberinfrastructure Panel. Cyberinfrastructure Vision for 21st Century Discovery. National Science Foundation, 2007. Available at http://www.nsf.gov/od/oci/CI_Vision_March07.pdf.
382. Nelson, M.L. *Buckets: Smart Objects for Digital Libraries*. PhD thesis, Old Dominion University. 2000.
383. Nelson, T., A File Structure for the Complex, the Changing, and Indeterminate, in *20th National Conference of the Association for Computing Machinery*, (New York, 1965), ACM.
384. Nelson, T.H. *Literary machines*. Mindful Press, Sausalito, CA, 1990.
385. Nottingham, M. HTTP Header Linking. draft-nottingham-http-link-header-00. IETF - Network Working Group, December 18, 2006. Available at <http://www.mnot.net/drafts/draft-nottingham-http-link-header-00.txt>.
386. Nottingham, M. and Sayre, R. The Atom Syndication Format. Request for Comments. 4287. Network Working Group, Internet Engineering Task Force, December, 2005. Available at <http://tools.ietf.org/html/rfc4287>.
387. Nunberg, G. (*Going nuclear*) : *language, politics, and culture in confrontational times*. PublicAffairs, New York, 2004.
388. O'Day, V. and Nardi, B.A. An Ecological Perspective on Digital Libraries. in Schatz, B., Bishop, A.P., Van House, N.A. and Bittenfield, B.P. eds. *Digital Library Use: Social Practice in Design and Evaluation (Digital Libraries and Electronic Publishing)*, MIT Press, Cambridge, MA, 2003.
389. O'Donnell, J.J. *Avatars of the word: from papyrus to cyberspace*. Harvard University Press, Cambridge, Mass., 1998.
390. O'Reilly, T. *What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software*. O'Reilly Media, Inc., 2005. Available at

<http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html>.

391. Olson, N.B. *Cataloging Internet Resources*. OCLC Online Computer Library Center, Inc., Dublin, OH, 1997.
392. Otlet, P. *Traité de documentation : le livre sur le livre, théorie et pratique*. Editions Mundaneum, Bruxelles, 1934.
393. Paepcke, A., Brandriff, R., Janee, G., Larson, R., Ludaeshcer, B., Melnik, S. and Raghavan, S. Search Middleware and the Simple Digital Library Interoperability Protocol. *D-Lib Magazine*, 5 (3). Available at <http://www.dlib.org/dlib/march00/paepcke/03paepcke.html>.
394. Paepcke, A., Chang, C.-C., Winograd, T. and Garcia-Molina, H. Interoperability for digital libraries worldwide. *Communications of the ACM*, 41 (4). 33-42.
395. Paepcke, A., Cousins, S. and Garcia-Molina, H. Towards Interoperability in Digital Libraries. Technical Report. CS-TR-97-1581. Stanford University Computer Science Department, June, 1997.
396. Paepcke, A., Cousins, S.B., Garcia-Molina, H., Hassan, S.W., Ketchpel, S.P., Röscheisen, M. and Winograd, T. Using Distributed Objects for Digital Library Interoperability. *Computer*, 29 (51). 61-68.
397. Paepcke, A., et al Search Middleware and the Simple Digital Library Interoperability Protocol. *D-Lib Magazine*, 2000. Available at <http://www.dlib.org/dlib/march00/paepcke/03paepcke.html>.
398. Paepcke, A., Garcia-Molina, H. and Wesley, R. Dewey Meets Turing: Librarians, Computer Scientists, and the Digital Libraries Initiative. *D-Lib Magazine*, 11 (7/8).
399. Parrish, P. The Trouble with Learning Objects. *Educational Technology Research and Development*, 52 (1). 49-67.

400. Paskin, N. and Rust, G. The Digital Object Identifier Initiative: Metadata Implications. Version 3. International DOI Foundation, February, 1999. Available at <http://dx.doi.org/10.1000/131>.
401. Payette, S., Blanchi, C., Lagoze, C. and Overly, E. Interoperability for Digital Objects and Repositories: The Cornell/CNRI Experiments. *D-Lib Magazine*, 1999 (May). Available at <http://www.dlib.org/dlib/may99/payette/05payette.html>.
402. Payette, S. and Lagoze, C., Flexible and Extensible Digital Object and Repository Architecture, in *In Research and Advanced Technology for Digital Libraries, Second European Conference, ECDL '98*, (Berlin, Heidelberg, 1998), Springer (Lecture notes in computer science). Available at <http://www.cs.cornell.edu/payette/papers/ECDL98/FEDORA.html>.
403. Payette, S. and Lagoze, C., Flexible and Extensible Digital Object and Repository Architecture (FEDORA), in *Second European Conference on Research and Advanced Technology for Digital Libraries*, (Heraklion, Crete, 1998).
404. Payette, S. and Lagoze, C., Policy-Enforcing, Policy-Carrying Digital Objects, in *Fourth European Conference on Research and Advanced Technology for Digital Libraries*, (Lisbon, Portugal, 2000). Available at <http://link.springer.de/link/service/series/0558/papers/1923/19230144.pdf>.
405. Payette, S. and Lagoze, C. Value-Added Surrogates for Distributed Content: Establishing a Virtual Control Zone. *D-Lib Magazine*, 6 (6). Available at <http://www.dlib.org/dlib/june00/payette/06payette.html>.
406. Payette, S. and Staples, T., The Mellon Fedora Project: Digital Library Architecture Meets XML and Web Services, in *European Conference on Research and Advanced Technology for Digital Libraries*, (Rome, 2002).
407. Pilgram, M. What is RSS?, 2002. Retrieved March 6, 2009, from <http://www.xml.com/pub/a/2002/12/18/dive-into-xml.html>.
408. Plutarch. Thesues, Retrieved May 20, 2009, from The Internet Classics Archive: <http://classics.mit.edu/Plutarch/theseus.html>.

409. President's Information Technology Advisory Committee: Panel on Digital Libraries. Digital Libraries: Universal Access to Human Knowledge. PITAC, February, 2001. Available at <http://www.itrd.gov/pubs/pitac/pitac-dl-9feb01.pdf>.
410. Prud'Hommeaux, E. and Seaborne, A. SPARQL Query Language for RDF. W3C, January 15, 2008. Available at <http://www.w3.org/TR/rdf-sparql-query/>.
411. Putz, S. Design and Implementation of the System 33 Document Service. Technical Report ISTL-NLTT-93-07-01. Xerox Palo Alto Research Center, 1993.
412. Raffl, C., Hofkirchner, W., Fuchs, C. and Schafranek, M. The Web as Techno-Social System. The Emergence of Web 3.0. in Trappl, R. ed. *Cybernetics and Systems 2008*, Austrian Society for Cybernetic Studies, Vienna, 2008, 204-209.
413. Raggett, D., Le Hors, A. and Jacobs, I. HTML 4.01 Specification. World Wide Web Consortium, December 24, 1999. Available at <http://www.w3.org/TR/1999/REC-html401-19991224/>.
414. Recker, M., Dorward, J., Dawson, D., Mao, X., Liu, Y., Palmer, B. and Park, J. Teaching, Designing, and Sharing: A Context for Learning Objects. *Interdisciplinary Journal of Knowledge and Learning Objects*, 1. 197-216.
415. Recker, M., Dorward, J. and Nelson, L.M. Discovery and Use of Online Learning Resources: Case Study Findings. *Educational Technology and Society*, 7 (2). 93-104.
416. Recker, M. and Walker, A. Collaboratively filtering learning objects. in Wiley, D.A. ed. *Designing Instruction with Learning Objects*, 2000.
417. Reeves, T.C. The Impact of Media and Technology in Schools: A Research Report prepared for The Bertelsmann Foundation. February 12, 1998. Available at http://www.ic.sunysb.edu/Stu/ashidele/The_Impact_of_Media_by_Bertelsmann_Fdtn.pdf.

418. Reich, L. and Sawyer, D. Reference Model for an Open Archival Model for an Open Archival Information Systems (OAIS): Information Systems (OAIS): Overview and Current Status Overview and Current Status, 2001. <http://www.dpconline.org/graphics/events/presentations/pdf/loureich.pdf>.
419. Reich, V. LOCKSS: A Permanent Web Publishing and Access System. *D-Lib Magazine*, 7 (6). Available at <http://www.dlib.org/dlib/june01/reich/06reich.html>.
420. Richardson, W. *Blogs, Wikis, Podcasts, and Other Powerful Web Tools for Classrooms*. Corwin Press, 2006.
421. Roosendaal, H.E. and Guerts, P.A.T.M., Forces and functions in scientific communities: an analysis of their interplay, in *CRISP 97: Cooperative Research Information Systems in Physics*, (Oldenburg, Germany, 1997).
422. Roscheisen, M., Baldonado, M.Q.W., Chang, K.C.-C., Gravano, L., Ketchpel, S.P. and Paepcke, A. The Stanford InfoBus and Its Service Layers: Augmenting the Internet with High-Level Information Management Protocols. in *Digital Libraries in Computer Science: The MeDoc Approach*, Springer-Verlag, London, UK, 1998, 213-230.
423. Roscheisen, M., Mogensen, C. and Winograd, T. Shared Web Annotations as a Platform for Third-Party Value-Added, Information Providers: Architecture, Protocols, and Usage Examples. Technical Report. CS-TR-97-1582. Stanford, 1997.
424. Rust, G. Metadata: The Right Approach. *D-Lib Magazine* (July/August 1998). Available at <http://www.dlib.org/dlib/july98/rust/07rust.html>.
425. Rust, G. and Bide, M. The <indec> Metadata Framework: Principles, model and data dictionary. WP1a-006-2.0. June, 2000. Available at http://www.doi.org/topics/indec/indec_framework_2000.pdf.
426. Ruusalepp, R. Infrastructure Planning And Data Curation: A Comparative Study Of International Approaches To Enabling The Sharing Of Research Data. JISC, November 30, 2008. Available at http://www.dcc.ac.uk/docs/publications/reports/Data_Sharing_Report.pdf.

427. Sairamesh, J., Kapidakis, S., Terzis, S. and Nikolaou, C. Performance Framework for QoS based Searching and Retrieval. TR97-0204. Institute of Computer Science - FORTH, 1997.
428. Salton, G. *Dynamic information and library processing*. Prentice-Hall, Englewood Cliffs, N.J., 1975.
429. Salton, G. and McGill, M.J. *Introduction to modern information retrieval*. McGraw-Hill, New York, 1983.
430. Sauermann, L., Cyganiak, R. and Volkel, M. Cool URIs for the Semantic Web. World Wide Web Consortium, 2007. Available at <http://www.w3.org/TR/cooluris/>.
431. Schatz, B., Telesophy: A System for Manipulating the Knowledge of a Community, in *GLOBECOM'87*, (Tokyo, 1987), IEEE.
432. Segaran, T. *Programming collective intelligence: building smart web 2.0 applications*, O'Reilly, Sebastapol, Calif., 2007.
433. Shirky, C. *Clay Shirky on institutions vs. collaboration*, 2005. Podcast, Accessed June 17, 2009 from TED: http://www.ted.com/index.php/talks/clay_shirky_on_institutions_versus_collaboration.html.
434. Shirky, C. *Ontology is Overrated: Categories, Links, and Tags*, Retrieved June 17, 2009, from Clay Shirky's Writings About the Internet: http://www.shirky.com/writings/ontology_overrated.html.
435. Shreeves, S., Knutson, E.M., Stvilia, B., Palmer, C.L., Twidale, M.B. and Cole, T.W., Is "Quality" Metadata "Shareable" Metadata? The Implications of Local Metadata Practices for Federated Collections, in *ACRL Twelfth National Conference*, (Minneapolis, 2005), ALA. Available at <http://www.ala.org/ala/acrl/acrlevents/shreeves05.pdf>.
436. Silverstein, C., Henzinger, M.R., Marais, H. and Moricz, M. Analysis of a very large Web search engine query log. *ACM SIGIR Forum*, 33 (3).

437. Smith, A., The research library in the 21st century: collecting, preserving, and making it accessible resources for scholarship, in *No brief candle: Reconceiving research libraries for the 21st century*, (Washington, DC, 2008), Council on Library and Information Resources.
438. Smith, M., DSpace: An institutional repository from the MIT libraries and Hewlett Packard Laboratories, in *6th European Conference on Research and Advanced Technology for Digital Libraries*, (2002), 542-549.
439. Smith, M., Bass, M., McClellan, G., Tansley, R., Barton, M., Branschofsky, M., Stuve, D. and Walker, J.H. DSpace: An Open Source Dynamic Digital Repository. *D-Lib Magazine*, 9 (1). Available at <http://www.dlib.org/dlib/january03/smith/01smith.html>.
440. Sowa, J.F. Conceptual Graphs, 2001. Retrieved May 5, 2009, from <http://www.jfsowa.com/cg/>.
441. Spasser, M.A., Realist Activity Theory for Digital Library Evaluation: Conceptual Framework and Case Study, in *Computer Supported Cooperative Work*, (2002), Springer.
442. Spasser, M.A. The Flora of North America Project. in Bishop, A.P., Van House, N.A. and Bittenfield, B.P. eds. *Digital Library Use*, MIT Press, Cambridge, Massachusetts, 2003.
443. Staples, T. and Wayland, R. Virginia Dons FEDORA: A Prototype for a Digital Object Repository. *D-Lib Magazine*, July. Available at <http://www.dlib.org/dlib/july00/staples/07staples.html>.
444. Staples, T., Wayland, R. and Payette, S. The Fedora Project. *D-Lib Magazine*, 9 (4).
445. Star, S.L. and Ruthleder, K. Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*, 7 (1). 111-134.
446. Stefik, M. Letting Loose the Light: Igniting Commerce in Electronic Publication. in Stefik, M. ed. *Internet Dreams*, MIT Press, 1996.

447. Stvilia, B., Twidale, M.B., Gasser, L. and Smith, L., Information quality in a community-based encyclopedia, in *Knowledge Management: Nurturing Culture, Innovation, and Technology - 2005 International Conference on Knowledge Management*, (2005), 101-113.
448. Suchman, L.A. *Plans and situated actions: the problem of human-machine communication*. Cambridge University Press, Cambridge [Cambridgeshire] ; New York, 1987.
449. Sun, S.X., Lammom, L. and Boesch, B. Handle System Overview. RFC. 3650. IETF, November, 2003. Available at <http://www.handle.net/rfc/rfc3650.html>.
450. Surowiecki, J. *The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. Doubleday New York, 2004.
451. Sutor, B. Open Standards vs. Open Source, 2006. Retrieved March 30, 2008, from <http://www.sutor.com/newsite/essays/e-OsVsOss.php>.
452. Svenonius, E. *The intellectual foundation of information organization*. MIT Press, Cambridge, Mass., 2000.
453. Tapscott, D. and Williams, A.D. *Wikinomics: how mass collaboration changes everything*. Portfolio, New York, 2008.
454. Taylor, A.G. *Understanding FRBR: what it is and how it will affect our retrieval tools*. Libraries Unlimited, Westport, Conn., 2007.
455. Teevan, J., Dumais, S.T. and Horvitz, E., Personalizing search via automated analysis of interests and activities, in *Annual ACM Conference on Research and Development in Information Retrieval*, (Salvador, Brazil, 2005), ACM.
456. Till, J. Predecessors of preprint servers. *Learned Publishing*, 14 (1). 7-13.
457. Treloar, A., Building an Institutional Research Repository from the Ground Up: The ARROW Experience, in *AusWeb04*, (Gold Coast, Australia, 2003). Available at

http://andrew.treloar.net/research/publications/ausweb04/ARROW_Architecture.shtml.

458. Ullrich, C., Borau, K., Luo, H., Tan, X., Shen, L. and Shen, R., Why web 2.0 is good for learning and for research: principles and prototypes, in *17th international Conference on World Wide Web*, (Beijing, 2008), ACM.
459. Unmil, P.K., Francisco-Revilla, L., Furuta, R.K., Hsieh, H. and Shipman III, F., M., Evolution of the Walden's Paths Authoring Tools, in *Webnet 2000*, (San Antonio, TX, 2000).
460. Van Damme, C., Hepp, M. and Siorpaes, K., FolksOntology: An Integrated Approach for Turning Folksonomies into Ontologies, in *Bridging the Gap between Semantic Web and Web 2.0*, (Innsbruck, Austria, 2007), June 7.
461. Van de Sompel, H., Bekaert, J., Liu, X., Balakireva, L. and Schwander, T. aDORe: a modular, standard-based Digital Object Repository. 2005. Available at <http://www.arxiv.org/abs/cs.DL/0502028>.
462. Van de Sompel, H., Hammond, T.H., Neylon, E. and Weibel, S. The "info" URI Scheme for Information Assets with Identifiers in Public Namespaces. RFC 4552. IETF, 2006.
463. Van de Sompel, H. and Lagoze, C. Interoperability for the Discovery, Use, and Re-Use of Units of Scholarly Communication. *CTWatch Quarterly*, 3 (3). Available at <http://www.ctwatch.org/quarterly/articles/2007/08/interoperability-for-the-discovery-use-and-re-use-of-units-of-scholarly-communication/8/>.
464. Van de Sompel, H., Lagoze, C., Nelson, C.E., Warner, S., Sanderson, R. and Johnston, P., Adding eScience Publications to the Data Web, in *Linked Data on the Web 2009*, (Madrid, 2009).
465. Van de Sompel, H., Nelson, M., Lagoze, C. and Warner, S. Resource Harvesting within the OAI-PMH Framework. *D-Lib Magazine*, 10 (12). Available at <http://www.dlib.org/dlib/december04/vandesompel/12vandesompel.html>.

466. Van de Sompel, H., Payette, S., Erickson, J., Lagoze, C. and Warner, S. Rethinking Scholarly Communication: Building the System that Scholars Deserve. *D-Lib Magazine*, 10 (9). Available at <http://www.dlib.org/dlib/september04/vandesompel/09vandesompel.html>.
467. Van House, N.A. Digital Libraries and Collaborative Knowledge Construction. in Van House, N.A. and Bishop, A.P. eds. *Digital Library Use: Social Practice in Design and Evaluation*, MIT Press, Cambridge, 2003.
468. Van House, N.A. Science and Technology Studies and Information Studies. *Annual Review of Information Science and Technology*, 38. 3-36.
469. Van House, N.A., Bishop, A.P. and Battenfield, B.P. Introduction: Digital Libraries as Sociotechnical Systems. in Bishop, A.P., Van House, N.A. and Battenfield, B.P. eds. *Digital Library Use*, MIT Press, Cambridge, MA, 2003.
470. Van House, N.A., Butler, M. and Schiff, L., Cooperative knowledge work and practices of trust: sharing and environmental planning data sets, in *ACM conference on Computer Supported Cooperative Work*, (1998), 335-343.
471. VanHeyningen, The Unified Computer Science Technical Report Index: Lessons in indexing diverse resources, in *WWW94*, (Chicago, 1994). Available at <http://www.cs.indiana.edu/ucstri/paper/paper.html>.
472. Velden, T. and Lagoze, C., Patterns of Collaboration in Co-authorship Networks in Chemistry - Mesoscopic Analysis and Interpretation, in *12th International Conference on Scientometrics and Informetrics*, (Rio de Janeiro, 2009).
473. Velden, T. and Lagoze, C., The transformation of scientific communication systems in the digital age: towards a methodology for comparing scientific communication cultures, in *Oxford e-Research 08*, (Oxford, UK, 2008).
474. W. Boyd Rayward Visions of Xanadu: Paul Otlet (1868-1944) and hypertext. *Journal of the American Society for Information Science*, 45 (4). 235-250. Available at [http://dx.doi.org/10.1002/\(SICI\)1097-4571\(199405\)45:4<235::AID-ASI2>3.0.CO;2-Y](http://dx.doi.org/10.1002/(SICI)1097-4571(199405)45:4<235::AID-ASI2>3.0.CO;2-Y).

475. Waldrop, M.M. Science 2.0 -- Is Open Access Science the Future? *Scientific American*, 2008 (May).
476. Warner, S. The OAI Data-Provider Registration and Validation Service. arXiv. cs.DL/0506010. June 3, 2005. Available at <http://arxiv.org/abs/cs.DL/0506010>.
477. Warner, S., Bekaert, J., Lagoze, C., Liu, X., Payette, S. and Van de Sompel, H. Pathways: Augmenting interoperability across scholarly repositories. *International Journal on Digital Libraries special issue on Digital Libraries and eScience*, 7 (1). 32-52.
478. Waters, D.J. What are digital libraries? *CLIR Issues*, 4. Available at <http://www.clir.org/pubs/issues/issues04.html>.
479. Wattenberg, F. A National Digital Libraries for Science, Mathematics, Engineering, and Technology Education. *D-Lib Magazine*, 1998 (October). Available at <http://www.dlib.org/dlib/october98/wattenberg/10wattenberg.html>.
480. Weibel, S. Metadata: The Foundations of Resource Description. *D-Lib Magazine*, July. Available at <http://www.dlib.org/dlib/July95/07weibel.html>.
481. Weibel, S., Iannella, R. and Cathro, W. The 4th Dublin Core Metadata Workshop Report: DC-4, March 3-5, 1997, National Library of Australia, Canberra. *D-Lib Magazine*, 1997 (June). Available at <http://www.dlib.org/dlib/june97/metadata/06weibel.html>.
482. Weibel, S., Kunze, J., Lagoze, C. and Wolf, M. Dublin Core Metadata for Resource Discovery. Request for Comments. 2413. Internet Engineering Task Force, September, 1998. Available at <ftp://ftp.isi.edu/in-notes/rfc2413.txt>.
483. Weibel, S.L. and Lagoze, C. An Element Set to Support Resource Discovery: The State of the Dublin Core. *International Journal of Digital Libraries*, 1 (1).
484. Weinberger, D. Why Dewey's Decimal System is prejudiced, 2004. <http://www.hyperorg.com/backissues/joho-sep03-04.html#dewey>.

485. Wheeler, B. E-research is a fad: scholarship 2.0, cyberinfrastructure, and IT governance. in Katz, R.N. ed. *The Tower and the Cloud*, EDUCAUSE. Inc., Washington DC, 2008.
486. Whyman, B. AOL's business model: Carnivore or herbivore? Precursor Group, 2000.
487. Wiens, J.A. *The ecology of bird communities*. Cambridge University Press, Cambridge [England] ; New York, 1989.
488. Wilczek, E. and Glick, K. Fedora and the Preservation of University Records. Tufts University and Yale University, 2004. Available at <http://dca.tufts.edu/features/nhprc/>.
489. Wilensky, R. Digital library resources as a basis for collaborative work. *Journal of the American Society for Information Science*, 51 (3). 228-245. Available at [http://dx.doi.org/10.1002/\(SICI\)1097-4571\(2000\)51:3<228::AID-ASI3>3.0.CO;2-5](http://dx.doi.org/10.1002/(SICI)1097-4571(2000)51:3<228::AID-ASI3>3.0.CO;2-5).
490. Williams, R. and Edge, D. The Social Shaping of Technology. *Research Policy*, 25. 856-899.
491. Willinsky, J. *The access principle: the case for open access to research and scholarship*. MIT Press, Cambridge, Mass., 2006.
492. Wilson, B. and Powell, A. A Tenth Anniversary for D-Lib Magazine. *D-Lib Magazine*, 11 (7/8).
493. Winograd, M. and Hais, M.D. *Millennial makeover : MySpace, YouTube, and the future of American politics*. Rutgers University Press, New Brunswick, N.J., 2008.
494. Witten, I.H., Boddie, S.J., Bainbridge, D.I. and McNab, R.J., Greenstone: a comprehensive open-source digital library software system, in *fifth ACM conference on Digital libraries*, (San Antonio, 2000), ACM, 113-121.
495. Wolfe, J.L., Effects of Annotations on Student Readers and Writers, in *Fifth ACM International Conference on Digital Libraries*, (San Antonio, TX, 2000).

496. Wood, D., Gearon, P. and Adams, T., Kowari: A Platform for Semantic Web Storage and Analysis, in *XTech2005: XML, the Web and beyond*, (Amsterdam, 2005). Available at <http://idealliance.org/proceedings/xtech05/papers/04-02-04/>.
497. World Wide Web Consortium. XHTML 1.0: The Extensible HyperText Markup Language (Second Edition), 2001. <http://www.w3.org/TR/2001/WD-xhtml1-20011004/>.
498. Zia, L.L. The NSF National Science, Technology, Engineering, and Mathematics Education Digital Library (NSDL) Program. *D-Lib Magazine*, 8 (11). Available at <http://dlib.org/dlib/november02/zia/11zia.html>.