

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