

UNIVERSITY OF CALGARY

The Gap between Personal vs Institutional Digital Archives of Researchers

by

Timothy Chung Yin Au Yeung

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES  
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE  
DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

CALGARY, ALBERTA

DECEMBER 2010

© Tim Au Yeung, 2010

UNIVERSITY OF CALGARY  
FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled "The Gap between Personal vs Institutional Digital Archives of Researchers" submitted by Tim Au Yeung in partial fulfilment of the requirements of the degree of Master of Science.

---

*Supervisor, Saul Green, Department of Computer Science*

---

*Sheelagh Carpendale, Department of Computer Science*

---

*Murray McGillivray, Department of English*

---

*Date*

# Abstract

---

Preservation of a researcher's digital material is a challenging problem for cultural heritage institutions mandated with archiving such materials. Archivists and curators have to do a significant amount of work to organize and describe the material. Archivists also have to act quickly on new material, as it is time-sensitive due to technological obsolescence and the inherent instability of its evolving software environment. One possibility is to have the institution incorporate digital materials created, maintained and archived by researchers via their personal websites. The problem is that personal websites are quite different from institutional archives. In this thesis, I explore these differences. I investigate researcher websites as a potential digital personal archive that could assist archivists in preserving the researcher's work. I survey websites of senior researchers within the domain of Human-Computer Interaction to see what they are currently creating. I interview researchers to understand their motivation. The results articulate the mismatch between the needs of archivists and the goals of researchers. From the results I derive a set of design guidelines for building digital archiving systems that could bring researchers and archivists closer together.

# Acknowledgements

---

As with any undertaking of substance, this thesis could not have happened without the support of a group of supporters I'd like to acknowledge here.

To Saul Greenberg, thank you for your patience and commitment as my supervisor. A lesser person would have let me flounder during the rough patches that inevitably accompany a process like this. Thank you for pushing me through instead. Your insights really helped to sharpen my thinking and keep me focused on the end goal.

To Mary Westell, thank you for your forbearance when this process distracted and pulled me away from my work duties and for letting me take the time I needed.

To Jackie Solar, thank you for your assistance in editing my thesis. It is well appreciated.

To my colleagues in the Interactions Lab, thank you for welcoming me and for your feedback.

To my friends and family, thank you for your understanding while I've been stressed, distracted, busy and otherwise occupied over these past few years.

And especially to my wife and son. Thank you. There are no words to express the depth of my appreciation for all you've had to put up with through this entire time.

I would also like to acknowledge the Natural Sciences and Engineering Research Council of Canada (NSERC) and Alberta's Informatics Circle of Research Excellence (iCORE) for their financial support.



*To my wife June*

*For her unshakeable confidence in me*

# Table of Contents

---

Approval Page.....	ii
Abstract .....	iii
Acknowledgements .....	iv
Table of Contents.....	vi
List of Tables .....	x
List of Figures and Illustrations .....	xi
 CHAPTER ONE: INTRODUCTION.....	 1
1.1 Background and Motivation .....	1
1.1.1 A Tale of Two Digital Lives .....	1
1.1.2 The Current State of Digital Preservation .....	5
1.1.3 Preserving the Individual .....	6
1.2 Research Problems .....	8
1.3 Thesis Goals .....	9
1.4 Definitions .....	10
1.5 Thesis Organization.....	11
1.5.1 Chapter Two: Literature review .....	11
1.5.2 Chapter Three: Survey of Researcher Websites.....	11
1.5.3 Chapter Four: Interviews with Researchers .....	12
1.5.4 Chapter Five: Design Recommendations and Analysis of Repository System .....	12
1.5.5 Chapter Six: Conclusion .....	12
1.6 Supplementary Content .....	12
1.6.1 Appendix A .....	12
1.6.2 Appendix B .....	12
1.6.3 Appendix C .....	12
1.6.4 Appendix D .....	12
1.6.5 Appendix E.....	12
1.6.6 Appendix F .....	12
 CHAPTER TWO: LITERATURE REVIEW .....	 13
2.1 Overview .....	13
2.2 Digital Preservation .....	13
2.3 Defining Digital Preservation.....	14
2.4 Key Concepts in Digital Preservation .....	16
2.4.1 The Information Object.....	16
2.4.2 Born Digital versus Digitized Content.....	16
2.4.3 Digital Object Lifecycle.....	17
2.5 Issues in Digital Preservation .....	18
2.6 Approaches in Digital Preservation.....	20
2.6.1 Open Archival Information System Model.....	20
2.6.2 Preservation Metadata .....	23
2.6.3 Trusted Digital Repositories .....	24

2.6.4 Preservation Activities .....	25
2.7 How Digital Preservation Relates to the Researcher.....	26
2.8 Analysis of Websites .....	27
2.8.1 Research on Websites .....	27
2.8.1.1 Identity Construction .....	27
2.8.1.2 Content Analysis .....	28
2.8.1.3 Academia and Websites .....	29
2.8.2 Research on Academic Websites .....	29
2.8.2.1 Dumont and Frindte.....	29
2.8.2.2 Rick.....	30
2.8.2.3 Barjak, Li, Thelwall.....	30
2.9 The Nature of Personal Archives .....	31
2.10 Summary.....	33
 CHAPTER THREE: WEBSITE SURVEY .....	 35
3.1 Chapter Synopsis .....	35
3.2 Introduction .....	35
3.2.1 Defining the Personal Website.....	36
3.3 Methodology.....	37
3.3.1 The Need to Survey Researcher Websites .....	37
3.3.2 Subject Selection.....	38
3.3.3 Identifying the Website .....	39
3.3.4 Initial Survey.....	40
3.3.5 Final Site Survey Methodology .....	41
3.4 Data Analysis.....	42
3.4.1 General Organization and Design .....	42
3.4.2 Type of Content.....	45
3.4.3 Bibliographies .....	46
3.4.4 Navigational Structure .....	47
3.5 Level of Content .....	50
3.6 Discussion of Results .....	53
3.7 A Typology of Sites.....	56
3.7.1 The Basic Professional Site.....	57
3.7.2 Researcher / Lab Site .....	58
3.7.3 Extensive Site.....	61
3.7.4 Organization Template.....	62
3.8 Conclusion.....	63
 CHAPTER FOUR: INTERVIEWS .....	 65
4.1 Chapter Synopsis .....	65
4.2 Introductory Discussion.....	65
4.2.1 What the Survey Tells Us About the Sites.....	66
4.2.1.1 The Dominance of the Bibliography .....	66
4.2.1.2 Personal Disclosure Variety .....	66
4.2.1.3 Website Design.....	67
4.2.2 What the Survey Does Not Tell Us About the Sites .....	67
4.3 Methodology.....	68

4.4 Synopsis of Results.....	70
4.4.1 Strong Desire for Control.....	70
4.4.2 Little Institutional Pressure to Conform.....	71
4.4.3 Trust In Institutional Systems .....	72
4.4.4 Limited Long Term View .....	72
4.4.5 The Website is a Public Face, Not an Archival Point.....	73
4.5 Case Studies.....	75
4.5.1 The Individual/Lab Website: Saul Greenberg.....	78
4.5.2 The Basic White Site: Ravin Balakrishnan.....	82
4.5.3 The Extensive Site: Ben Bederson.....	85
4.5.4 The Organization Template: Jonathan Grudin.....	89
4.6 Conclusion.....	92
CHAPTER FIVE: ANALYSIS AND IMPLICATIONS .....	93
5.1 Chapter Synopsis.....	93
5.2 Impact of Findings.....	93
5.2.1 The Digital Preservation Problem at Creator Scale .....	93
5.2.2 How Do the Findings Relate? .....	96
5.2.2.1 The Digital Object and the Intellectual Entity.....	97
5.2.2.2 Lifecycle Preservation Impacts .....	98
5.2.2.3 Metadata .....	100
5.2.2.4 Limitations of the Website .....	100
5.3 Design Recommendations .....	101
5.3.1 Translating Findings To Design Recommendations.....	101
5.3.1.1 The researcher's identity and distinctiveness needs to be maintained at all times.....	102
5.3.1.2 Publications are the core content and most often updated. There should be a set of easy to use workflow tools that allow the researcher to update them. ....	103
5.3.1.3 Publications are the core archival unit. There should be ways to link related items that comprise the publications as a unit. ....	103
5.3.1.4 The publication list is needed for a number of purposes. There should be flexibility in presentation. ....	104
5.3.1.5 Researchers should be encouraged to create more archivable content through easy to use tools like wizards and graphical user interfaces. ...	105
5.3.1.6 Few researchers will start with little or no content. The ability to migrate from other systems or to other systems easily is a necessity as a result. ....	106
5.3.1.7 Researchers generally trust institutional systems but prefer substantial control over the system. The design of the system should reflect that. ....	107
5.4 Evaluation of Current Systems.....	107
5.4.1 Limitations of Current Approaches.....	107
5.4.2 Current Archiving Systems.....	108
5.4.3 Analysis of DSpace.....	109
5.4.3.1 The researcher's identity and distinctiveness needs to be maintained at all times.....	109

5.4.3.2 Publications are the core content and most often updated. There should be a set of easy to use workflow tools that allow the researcher to update them. ....	111
5.4.3.3 Publications are the core archival unit. There should be ways to link related items that comprise the publications as a unit. ....	114
5.4.3.4 The publication list is needed for a number of purposes. There should be flexibility in presentation. ....	115
5.4.3.5 Researchers should be encouraged to create more archivable content through easy to use tools like wizards and graphical user interfaces. ...	116
5.4.3.6 Few researchers will start with little or no content. The ability to migrate from other systems or to other systems easily is a necessity as a result. ....	117
5.4.3.7 Researchers generally trust institutional systems but prefer substantial control over the system. The design of the system should reflect that. ....	117
5.5 Conclusion.....	117
CHAPTER SIX: CONCLUSION .....	119
6.1 Research Questions .....	119
6.2 Thesis Contributions.....	120
6.3 Future Work.....	121
6.3.1 Broadening the Surveys and Interviews.....	122
6.3.2 Design Recommendation Refinement.....	123
6.3.3 Prototype System .....	124
6.4 Final Comments.....	128
REFERENCES.....	130
APPENDIX A: RESEARCHERS WEBSITES SURVEYED .....	138
APPENDIX B: SITE SURVEY FIELDS .....	139
APPENDIX C: INTERVIEW QUESTIONS .....	141
APPENDIX D: ETHICS APPROVALS .....	146
APPENDIX E: CONSENT FORM .....	148
APPENDIX F: LIST OF WEBSITES.....	152

# List of Tables

---

Table 2-1: Ways of Looking at Digital Preservation.....	15
Table 2-2: Threats to Digital Objects .....	18
Table 3-1: Types of Site Content.....	45
Table 3-2: Distance of Bibliography From Home Page .....	47
Table 3-3: Elements of the Navigational System .....	49
Table 5-1: Design Recommendations.....	118

# List of Figures and Illustrations

---

Figure 1-1: Searching on Google for Doug Engelbart .....	2
Figure 1-2: Searching on Google for Mark Weiser .....	3
Figure 1-3: Mark Weiser's Website .....	4
Figure 1-4: Doug Engelbart Website.....	5
Figure 1-5: Personal Archives .....	8
Figure 2-1: OAIS Diagram (reproduced from CCSDS, 2002).....	22
Figure 3-1: Researcher Website Type .....	42
Figure 3-2: Content Completeness .....	42
Figure 3-3: Type of Website Design .....	43
Figure 3-4: Number of Pages in Site .....	44
Figure 3-5: Personal Content Level.....	50
Figure 3-6: Project / Research Content Level .....	51
Figure 3-7: Teaching / Instructional Resource Content Level .....	52
Figure 3-8: External Content Level.....	52
Figure 3-9: Basic Professional Site (Peter Johnson) .....	57
Figure 3-10: Personal Page (Carl Gutwin) .....	58
Figure 3-11: Lab Page (Carl Gutwin).....	59
Figure 3-12: Extensive Site (Alan Dix).....	61
Figure 3-13: Organization Template Site (Michael Muller) .....	62
Figure 4-1: Ben Shneiderman's Website .....	76
Figure 4-2: Ben Shneiderman Archival Page .....	76
Figure 4-3: Saul Greenberg's Website .....	78
Figure 4-4: Saul Greenberg's Lab Page .....	78

Figure 4-5: Saul Greenberg's Publication Page .....	80
Figure 4-6: Saul Greenberg's Project Page .....	81
Figure 4-7: Ravin Balakrishnan's Website .....	83
Figure 4-8: Ben Bederson's Website .....	85
Figure 4-9: Ben Bederson Project Page.....	86
Figure 4-10: Ben Bederson's Blog.....	88
Figure 4-11: Jonathan Grudin's Corporate Page.....	89
Figure 4-12: Jonathan Grudin's Previous Website .....	91
Figure 5-1: DSpace at MIT .....	109
Figure 5-2: DSpace at the University of Toronto .....	110
Figure 5-3: DSpace at the University of Calgary .....	110
Figure 5-4: DSpace New Submission Screen.....	111
Figure 5-5: DSpace Choose Collection Screen .....	112
Figure 5-6: DSpace: Describe Item Screen .....	112
Figure 5-7: DSpace: Describe Item Screen 2 .....	113
Figure 5-9: DSpace: Browsing by Title Screen at U of C .....	115
Figure 5-10: DSpace: Browsing by Title Screen at U of T .....	115
Figure 6-1: Prototype System Data Flow .....	125
Figure 6-2: Prototype System Templates .....	126
Figure 6-3: Tumblr Content Type Bar.....	126



# Chapter One: Introduction

---

In this thesis, I will investigate the issue of preserving digital material created by researchers. Specifically, I will consider how human-computer interaction researchers construct their web presence and how this can be used to assist archivists and curators in preserving their work. To set the stage for the problem, I will discuss in this chapter the broader context of digital preservation, digital preservation at the level of the individual and the fate of the digital “estate” of two researchers. Finally I will outline the remainder of the chapters.

## 1.1 Background and Motivation

---

### ***1.1.1 A Tale of Two Digital Lives***

When someone sits down to write a novel, to compose a song, to photograph an event, to do scientific research, their tool is now, almost invariably, digital. We live in a digital world and the instruments we use to create information are digital. The challenge is how to save our digital creations. The speed of change and unstable technology environments create a situation where information is in danger of disappearing almost as quickly as it is created. This is the problem of digital preservation: how do we ensure that the digital material we create will be available for future generations?

Let us consider a concrete example: these days, if you want to find something out about someone, you’d likely start your search online. And you would likely find it there as more and more, people are posting information about themselves online. This is especially true for scientists given their *raison d’être* are to make their research public. Consider for a moment two seminal figures in Computer Science: Doug Engelbart and Mark Weiser. If you search on Google for the two scientists, you’d retrieve results like this (Figure 1-1 and Figure 1-2):

Google   [Advance](#) [Preferen](#)

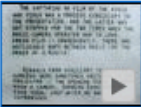

**Web** [Video](#)

**[Douglas Engelbart - Wikipedia, the free encyclopedia](#)**  
**Doug Engelbart's** career was inspired in 1951 when he got engaged and suddenly realized he had no career goals beyond getting a good education and a decent ...  
[en.wikipedia.org/wiki/Douglas\\_Engelbart](http://en.wikipedia.org/wiki/Douglas_Engelbart) - 84k - [Cached](#) - [Similar pages](#)

**[Douglas Engelbart](#)**  
**Douglas Engelbart** was born in 1925, in Oregon, where he grew up on a small farm. In 1942, he graduated high school and went to Oregon State University to ...  
[www.ibiblio.org/pioneers/engelbart.html](http://www.ibiblio.org/pioneers/engelbart.html) - 22k - [Cached](#) - [Similar pages](#)

**[Doug Engelbart 1968 Demo](#)**  
On December 9, 1968, **Douglas C. Engelbart** and the group of 17 researchers working with him in the Augmentation Research Center at Stanford Research ...  
[www.stanford.edu/dept/SUL/library/extra4/sloan/MouseSite/1968Demo.html](http://www.stanford.edu/dept/SUL/library/extra4/sloan/MouseSite/1968Demo.html) - 61k - [Cached](#) - [Similar pages](#)

**[Video results for Doug Engelbart](#)**

	<b><a href="#">Doug Engelbart: The Demo</a></b> 75 min <a href="http://video.google.com">video.google.com</a>		<b><a href="#">Douglas Engelbart : The Mother of All Demos (1/9)</a></b> 9 min <a href="http://www.youtube.com">www.youtube.com</a>
--	---	--	---

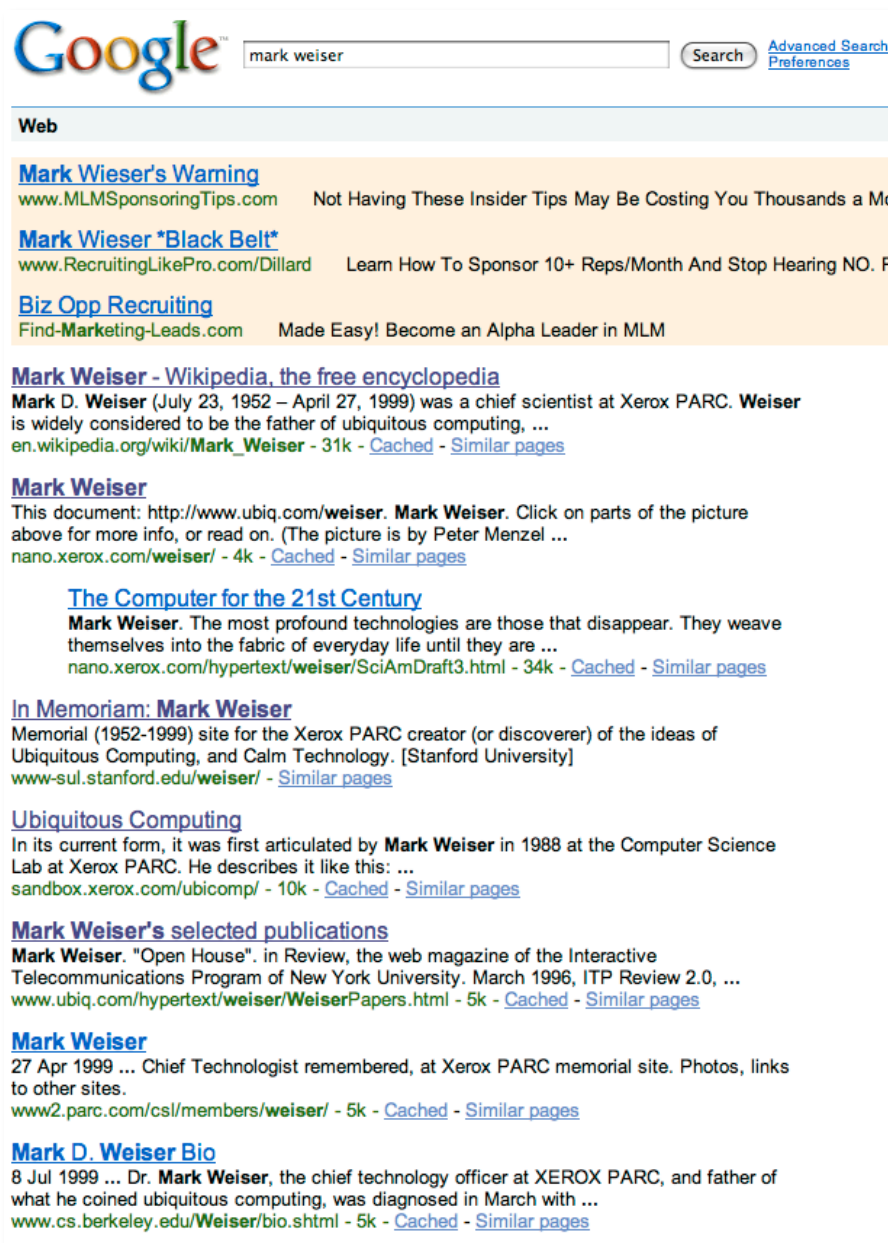
**[Bootstrap Institute: About BI](#)**  
The Bootstrap Institute was conceived by Dr. **Douglas C. Engelbart** to further his .... In accord with the above reasons for action, **Doug Engelbart** developed ...  
[www.bootstrap.org/](http://www.bootstrap.org/) - 24k - [Cached](#) - [Similar pages](#)

**[Biographical Sketch: Doug Engelbart](#)**  
At **Engelbart's** headquarters, his Bootstrap Institute.  
[www.bootstrap.org/chronicle/chronicle.html](http://www.bootstrap.org/chronicle/chronicle.html) - 40k - [Cached](#) - [Similar pages](#)  
[More results from www.bootstrap.org »](#)

**[Welcome - Doug Engelbart Institute](#)**  
Official website for the **Doug Engelbart** Institute -- a think tank for advancing collaborative tools, practices, and strategies for creating high-performing ...  
[www.doungelbart.org/](http://www.doungelbart.org/) - 18k - [Cached](#) - [Similar pages](#)

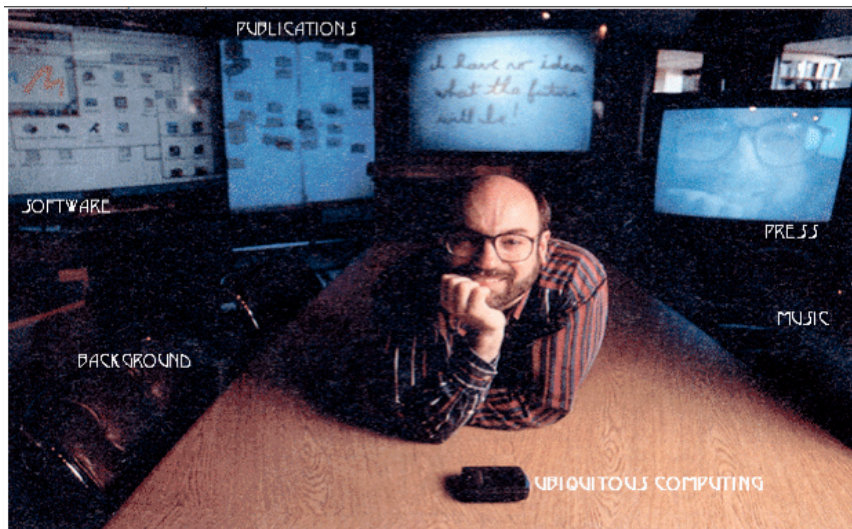
**[Who Invented the Computer Mouse - Douglas Engelbart](#)**  
**Douglas Engelbart** invented point and click computing with the computer mouse windows etc.  
[inventors.about.com/library/weekly/aa081898.htm](http://inventors.about.com/library/weekly/aa081898.htm) - 33k - [Cached](#) - [Similar pages](#)

Figure 1-1: Searching on Google for Doug Engelbart



**Figure 1-2: Searching on Google for Mark Weiser**

Both men figure prominently in the search results and majority of the top links are directly relevant to understanding the legacy of each scientist. However, there are significant differences in how the men are represented. The most relevant site for Mark Wesier that appears to provide a comprehensive overview of his work can be found at the Xerox PARC site. A screenshot of the site is captured here:



Click on parts of the picture above for more info, or read on. (The picture is by [Peter Menzel](#), and first appeared in a German magazine profile, without the words. I have adopted it for my web page with permission. Thanks Peter.)

[Pictures](#) from my vacation in Greece with my daughter Nicole.

[Pictures](#) from the vacation part of my recent trip to Brazil.

I ran the Computer Science Laboratory at Xerox PARC for seven years, stepped down in 1994 to found a startup, and I have now just started (August 1996) as [Chief Technologist of Xerox PARC](#). This will really be fun!

I am the drummer for [Severe Tire Damage](#), first live band on the internet.

I work in [ubiquitous computing](#).

[Slides for a keynote talk](#) to the International Conference on Software Engineering, ICSE97, on reaching agreement in software engineering, entitled "Software Engineering and People".

[Slides for a talk](#) on the Computer Science challenges of the next 10 years.

[Slides for a talk](#) on educational challenges for Computer Science for the next 10 years. Given to the biannual Snowbird conference for Chairs of Computer Science Departments and Industrial Research Labs.

My research interests are garbage collection, operating systems, user interfaces, and [ubiquitous computing](#). I used to work on software engineering and program slicing, but not much any more. (See [papers](#) for more info.) I am also active in the [Computer Research Association](#), where I organized the first workshop for heads of industrial computer science laboratories.

I was program chair for the [15th Symposium on Operating Systems Principles](#).

**Figure 1-3: Mark Weiser's Website**

However, while the image map (Figure 1-3) appears to contain links to the collected works, they are actually all dead links. Weiser's work must be tracked down from other sources. This is in large part due to Weiser's untimely death in 1999. What is left are just the vestiges of his life's work. Fortunately, his publications survive but the coherency he brought to the corpus through his website is lost.

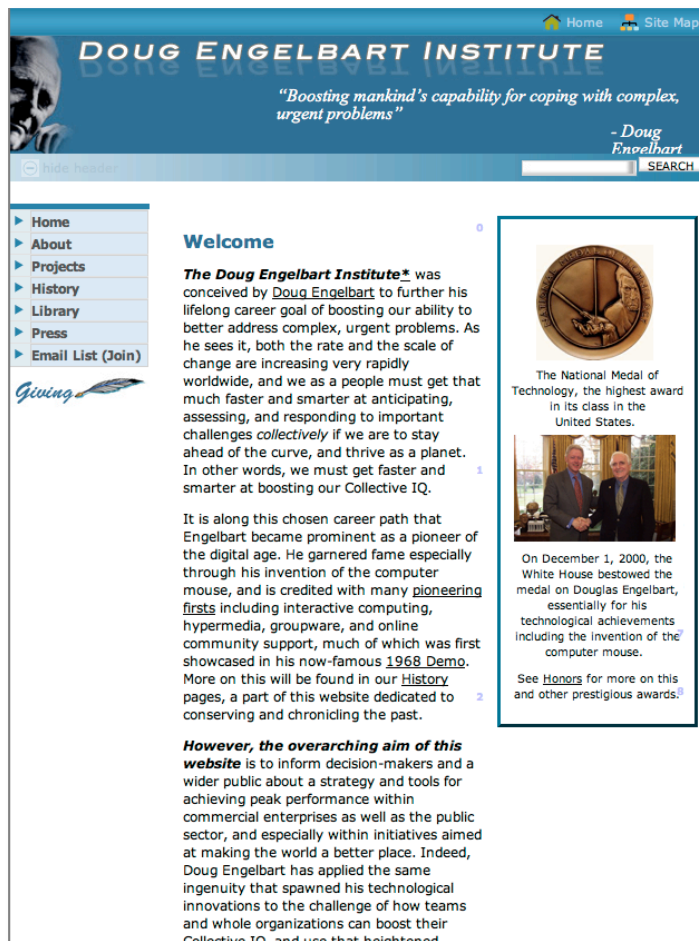


Figure 1-4: Doug Engelbart Website

In contrast, Engelbart's archive, found here, is a well-organized, curated retrospective with links to his bibliography, projects and a comprehensive autobiography. This is due in no small part to the fact that Engelbart is active in the curation of this archive. Engelbart has eliminated the guesswork as how his corpus fits together by providing the context and motivation behind his work. The differences between the two archives are striking. In the next section, I will explore how these issues play out in a wider context.

### 1.1.2 The Current State of Digital Preservation

In the past thirty years, the world has witnessed a rapid transformation from communicating and storing information on analog media to using digital media. This

shift in the medium has also brought a dramatic change in the way information is used. Lines of control over the dissemination of information have become blurred while the volume being created has grown exponentially. The environment has moved from a tightly controlled regime over dissemination to a plethora of creators broadcasting over multiple channels free of intermediaries. However, this democratization of information creation and dissemination has made it more difficult to ensure that the information is preserved for future generations. The growth in volume, the decentralized nature and the general ephemerality of the digital medium have all conspired to make this difficult (Waters & Garrett, 1996). The response has been a programme of research into digital preservation by many organizations around the world.

Currently, majority of digital preservation research has been done at the macro scale, with projects often spanning institutions and countries (Library of Congress, 2002; Planets, 2006). The primary emphasis of this research has focused on supporting the efforts of institutions to preserve large masses of digital content. Theoretical work includes describing how digital preservation systems should function (Consultative Committee for Space Data Systems [CCSDS], 2002) and assessing the kind of information needed for preservation purposes (The CEDARS Project [CEDARS], 2001; OCLC/RLG Working Group on Preservation Metadata [OCLC/RLG WGPM] 2002; PREMIS Editorial Committee [PREMIS], 2008). Practical work includes developing preservation techniques like emulation (Rothenberg, 1999; Granger, 2000; van der Hoeven, Lohman & Verdegem, 2007) and systems for storing digital content (Reich & Rosenthal, 2001; Staples, Wayland & Payette, 2003; RLG-OCLC Digital Archive Attributes Working Group [RLG-OCLC], 2002).

### ***1.1.3 Preserving the Individual***

In contrast to institutional preservation, little work has been done for individuals wanting to preserve their digital content. The majority of the work at the individual level has focused on personal information management. This includes organizing documents on the desktop (Henderson, 2004; Boardman & Sasse, 2004), e-mail



(Whittaker & Sidner; 1996, Mock, 2001), personal information (Bernstein, van Kleek, Karger & schraefel, 2007), web information (Bruce, Jones & Dumais, 2004) or the sum of all information in your life (Bell, 2001; Gemmell, Bell & Lueder, 2006; Marshall, 2007). However, this work tends to focus on the individual's use of digital information rather than on how to organize and preserve what he or she creates.

One emergent idea in individual digital preservation is the "digital estate" (Beagrie, 2005). In the realm of traditional archives, authors place their personal estate (correspondences, personal notes, rough drafts) in a library or archive for posterity. And for some, the presentation is as much part of the individual legacy as the content. While this is easy for paper, the digital equivalent is much more complex. Particularly in the case of posthumous deposit, the digital material can be indecipherable because of old and obscure formats (BBC, 2002). Moreover, multiple versions and unnoticed data loss can confuse the organization in the estate.

Moreover, the issue of the digital estate presents a particular challenge when the creator is a researcher. As Marshall notes (Marshall, 2008) research activity produces a cluster of archival artefacts. But because of the collaborative nature of research, ownership of the artefacts may not be clear, or the researcher may own only a part of the overall set. Other artefacts like data sets and software may require complex documentation and a specific technological environment to be viewed. This means that helping researchers organize and archive their digital material now, while the researcher is available, to decipher their estate is critical.

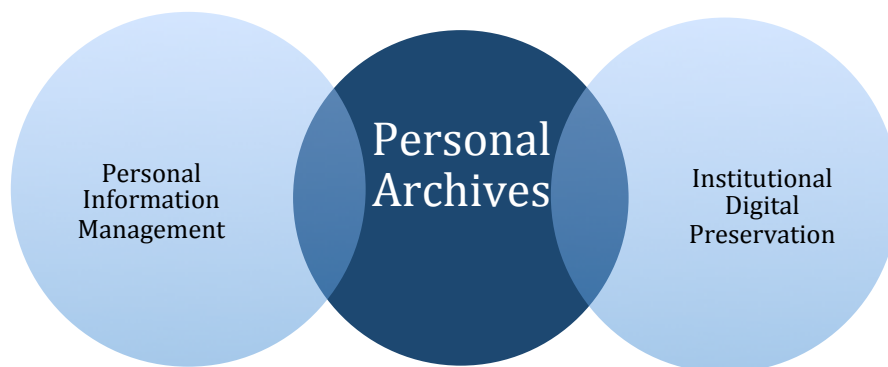
Unfortunately the current tools and strategies being developed at the institutional level do not support the way researchers do research. Where institutional tools demand an understanding of complex systems and standards, researchers need tools to be incidental and simple (Marshall, 2008). Where the institutional focus has been on the deposit of individual digital artefacts, researchers' work needs to be thought of as an aggregate for it to be intelligible.

## 1.2 Research Problems

---

As we have seen, there is a disconnection between digital preservation for the institution and the individual. The institutional focus has been on tools and standards with little consideration of the researcher. On the other hand, researchers have not done a good job of creating content in a way that makes it easy for institutions to archive their work. And they have demonstrated a reluctance to participate in institutional systems and tools that have the express goal of archiving. As we have seen (particularly in Engelbart's efforts), researchers are not only organizing the work but also making it available publicly. So if institutions are saving content and researchers are organizing content and making it available publicly, there ought to be an overlap.

More formally, we have a problem space that is at the intersection of how researchers manage and present their digital resources (particularly those of their own creation) and how institutions preserve them. It is possible to posit the idea of a personal archive where researchers organize their work for the purposes of creating a legacy. Such an archive could be created so that institutions could easily preserve it for future use. But are researchers creating personal archives?



**Figure 1-5: Personal Archives**

While researchers generally do not maintain formal archives (Marshall, 2008), we have seen that Engelbart does in fact create something that comes close to it in his personal website. This leads to the question: could a researcher's personal website



serve as a personal archive? Moreover, is it possible for institutions to support researchers in creating their websites while making the content more preservable? In order to answer this, we need to answer three questions:

1. **What is the current state of personal websites for researchers?** There are currently no concerted efforts to study researcher websites. For this reason, it is important to understand what researchers are making available on their websites and how that can impact the preservation of their corpus.
2. **Could researchers be motivated to create preservable content?** At the moment we know researchers are creating websites. But we do not know whether they are creating them strictly for current needs or whether there is a sense of legacy that might motivate them to create preservable content. Similarly we do not know what are the kinds of motivators that would encourage them to change or adopt new tools and systems that institutions might provide.
3. **How can institutions facilitate making the content preservable?** Even if the content is preservable, it does not mean it is usable by the institutions. Moreover, even if it is usable, it may require significant work on the part of institutions. What can institutions do to facilitate the researcher process while encouraging the creation of content that is easily usable for the institution?

### 1.3 Thesis Goals

---

Throughout this discussion, I have referred to researchers in general. However, trying to get a sense of these issues for all researchers may be an intractable problem in this context. Instead, I will focus on a narrower class of researcher to understand the problem at a more manageable level. Specifically, I will focus on senior researchers in the area of Human-Computer Interaction (HCI). In this thesis, I will address the questions above through the following methods:

1. **I will survey existing websites of senior researchers within the HCI domain.** This will allow me to develop a typology of the kinds of websites and general approaches that creators take to the organization of their sites.
2. **I will investigate the motivation of researchers in the creation of their websites.** I will identify what motivates researchers to create personal websites and the reasons behind the content and technology choices by interviewing them. I will look at what events or motivators would incite them to change their current system and adopt a new approach.
3. **I will identify design recommendations for systems based on the results of the first two goals and evaluate one exemplar against those recommendations.** By posing a set of design recommendations and evaluating current repository systems, I can begin to identify the challenge of integrating personal archives with institutional systems.

## 1.4 Definitions

---

While many of the terms used in this thesis are explained in the next chapter, several terms need to be defined early on. *Curators* and *archivists* in a traditional context have overlapping but distinct functions within cultural memory institutions. Curators and archivists are both responsible for managing collections within these institutions. Curators focus on developing expertise on the collection, often guiding the collecting and engaging in research on the collection. Archivists, on the other hand, focus on the management of received collections and try, as much as possible, to respect the organization of the collection received from the creator.

For this thesis, curator or archivist will be used for those individuals within cultural memory institutions, like libraries, archives and museums, who are responsible for managing digital resources for the long term. Within the digital preservation communities, the definition of the roles is often interchangeable; this is the by-product of how new the field is and the lack of clarity in the roles.

Similarly, a *researcher* in the context of this thesis is one who has formal responsibility for conducting research on behalf of or within an organization. While individuals outside of formal channels conduct research, it is more difficult to generalize the needs of these individuals and consequently, more difficult to suggest strategies for addressing this group. *Institutions* (including *institutional needs* and *expectations*) refer to cultural memory institutions that have formal obligations to preserve digital content for future access. As with researchers, while institutions other than cultural memory institutions often archive digital content, the obligation is an important aspect to understanding the behaviour and approach to digital preservation that these organizations take.

## 1.5 Thesis Organization

---

In the next 5 chapters, I will describe the current literature covering the topic areas of digital preservation, personal information preservation, and websites in academia. I will then discuss the two studies I conducted to address goals 1 and 2 followed by the design recommendations addressing goal 3 with the following organization of chapters:

### ***1.5.1 Chapter Two: Literature review***

Chapter two details the literature that underpins both digital preservation and the analysis of personal websites. I explore how these two things interact with the particular focus on researcher personal websites.

### ***1.5.2 Chapter Three: Survey of Researcher Websites***

Chapter three outlines the methodology used to survey researcher websites. I explore the results of the researcher websites both in terms of quantitative results and in terms of a typology of researcher websites.

### ***1.5.3 Chapter Four: Interviews with Researchers***

In chapter four, I describe the methodology used to interview researchers. I summarize the results into a series of general findings as well as four specific case studies with individual researchers and their views.

### ***1.5.4 Chapter Five: Design Recommendations and Analysis of Repository System***

Chapter five contains the synthesis of the survey and interview results placed in the context of digital preservation of researcher work. Out of that synthesis is a set of design recommendations that outline how a system could be built for serving researcher and archival institution needs. In the final section, I compare the design recommendations to DSpace, a popular institutional repository system.

### ***1.5.5 Chapter Six: Conclusion***

In the conclusion I revisit my initial research questions and describe the contributions I make to the literature. I also discuss possible future work.

## **1.6 Supplementary Content**

---

### ***1.6.1 Appendix A***

Appendix A lists researchers whose websites were surveyed in chapter three.

### ***1.6.2 Appendix B***

Appendix B lists categories used in the website survey in chapter three.

### ***1.6.3 Appendix C***

Appendix C lists the questions used in the survey in chapter four.

### ***1.6.4 Appendix D***

Appendix D contains the ethics approvals for the interviews in chapter four.

### ***1.6.5 Appendix E***

Appendix E contains the consent form used for the interviews in chapter four.

### ***1.6.6 Appendix F***

Appendix F lists the URLs for websites referenced in chapters two, three and four.

## Chapter Two: Literature Review

---

In the previous chapter I set the stage with a description of the research problems I will address and a set of goals to address those problems. In this chapter I will review the pertinent literature to put my research into context.

### 2.1 Overview

---

As I have outlined, the problem space occupies two distinct but related areas of research. The first area is *digital preservation*. Most of the literature in digital preservation is only tangentially related to my problem but without considering the broader body of literature, it is hard to contextualize the issues. Moreover while there is large body of literature, it is mostly unfamiliar to those who are not digital preservation practitioners.

The second area of research focuses on the *analysis of personal websites*. Unlike the digital preservation literature, there is little research on the nature of personal websites of researchers. In order to provide a sense of direction to my research, I will need to briefly consider the broader area of website analysis.

### 2.2 Digital Preservation

---

Digital preservation is a broad area of research and practice. It covers efforts to save both the vast realms of CDs and DVDs we have produced to store our commercial music and movies, and the magnetic tape used to store home movies and music. It covers the science of preserving the vast libraries of digital books, which contain the sum of human knowledge. It includes saving the documents stored on the personal computer. Digital preservation covers both tactical efforts to save local collections and global strategies to handle a nation's digital output.

The past twenty years has seen relatively intense activity involving national institutions, multinational consortia, large projects and a mountain of literature. Yet, for all of this research and activity, the number of evidence-based practises and

conclusive findings within this literature are few and far between. One reason is that the field is largely speculative: many of the reported research outcomes are built around tentative prototypes developed to address potential future problems and issues. However, despite the difficulty of predicting the future, researchers push on because the consequence of inaction is the potential loss of all digital content.

In this section, I will examine the wider context of digital preservation focusing on four areas:

1. Defining digital preservation.
2. Key concepts in digital preservation.
3. Issues in digital preservation.
4. Approaches to digital preservation.

I will then summarize how the current state of digital preservation impacts the individual creator and his or her digital content.

## 2.3 Defining Digital Preservation

---

One of the pioneering works in digital preservation, *Preserving Digital Information* (Waters & Garret, 1996), presents the challenge as “a grander problem of organizing ourselves over time and as a society to maneuver effectively in a digital landscape.” While this definition lends a sense of *gravitas* to the problem, it is not helpful in coming to terms with what digital preservation is and how it can be addressed. Yet the vagueness of the definition is reflective of the field as a whole. Practitioners often use three terms interchangeably: *digital archiving*, *digital curation* and *digital preservation*. To get a better sense of digital preservation, consider a number of more specific definitions for digital preservation. For instance, the Cornell tutorial on digital preservation (Kenney, McGovern, Entlich, Kehoe & Buckley, 2004) provides this definition:

...digital preservation is to maintain the ability to display, retrieve and use digital collections in face of rapidly

changing technology and organization infrastructures and elements

The Digital Preservation Coalition (Beagrie & Jones, 2001) uses this definition:

Digital Preservation [r]efers to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary.

And finally the American Library Association (Martyniak, Nadal, Ryder, Frangakis, Blood, Brown, Byrnes, 2007) adopted this as its medium-length definition of digital preservation:

Digital preservation combines policies, strategies and actions to ensure access to reformatted and born digital content regardless of the challenges of media failure and technological change. The goal of digital preservation is the accurate rendering of authenticated content over time

A number of key parameters can be gleaned from these definitions: that digital preservation is focused on ensuring accessibility; that it is a series of ongoing managed activities and that both technological and organizational solutions are required to address the challenges. Expanding on this, Lavoie and Dempsey (2004) suggest that there are in fact thirteen ways of looking at digital preservation:

As an ongoing activity	As a set of agreed outcomes
As an understood responsibility	As a selection process
As an economically sustainable activity	As a cooperative effort
As an innocuous activity	As an aggregated or disaggregated service
As a complement to other library services	As a well-understood process
As an arm's length transaction	As one of many options
As a public good	

**Table 2-1: Ways of Looking at Digital Preservation**

Some of these are simply motherhood statements and do not translate into things we can act on. But the ones that can be acted on are not technological solutions. In

fact, most of these imply human action, involving multiple actors over long periods of time. Using the definitions and their implications, I will address key concepts in digital preservation next.

## 2.4 Key Concepts in Digital Preservation

---

### **2.4.1 The Information Object**

The central concept in digital preservation is the idea of the *information object*. The *Preserving Digital Information* report (Waters & Garrett, 1996) argues that the *information* or *digital object* is the essential thing that is preserved and that to understand *how* to preserve a digital object, one needs to understand *what* essential attributes define a digital object. Examples of information objects include digital photographs, portable document files (PDF) of research publications, web pages and digitized books.

Digital objects are like physical ones in that they are created, modified, disseminated and discarded; in essence they have a lifecycle. Unlike physical objects, digital objects are typically made up of a set of digital files with different formats. This is an important distinction—a digital object is not just data (Quisbert, Korenkova & Hagerfors, 2007) but is a coherent unit bound by intellectual interpretation. While the loss of one constituent part may not dissolve the coherency of the object, one cannot shuffle the parts around separately without at some point losing the intellectual value of the object.

### **2.4.2 Born Digital versus Digitized Content**

While all digital objects require preservation, not all digital objects are created equally. An important distinction that needs to be made is between “born digital” objects created using digital tools and digitized objects that are transformed from an existing analog object into digital form. Born digital objects represent a much larger class of objects and can include things like e-mail, images from digital cameras and web pages. Digitized objects include books scanned into individual digital pages,



movies transferred from film to a medium like DVD or audio recorded from records to digital files.

There tends to be a bias in cultural heritage communities to digitized content due to their significant investment in digitization in the last decade. This bias extends to digital preservation standards that favour guidance on digitizing for archival quality over creating born digital objects. However, while digitized objects may be more heavily invested in, born digital objects are often more complex and are more valuable in that once lost, they cannot be replaced. In contrast, digitized objects can be re-digitized in the event of catastrophic loss.

#### ***2.4.3 Digital Object Lifecycle***

As already noted, digital objects possess a lifecycle from creation to disposal, much like a physical object. However, there are a number of key differences. The physical object lifecycle is a chain of custody (*provenance*) from one agent to another and is important for authenticity purposes. During that time conservation may be performed to preserve the object. The digital, on the other hand, can exist in multiple locations, and conservation in digital terms sometimes requires radical changes to the object.

To understand this we need to look at the phases of the lifecycle of a digital object. While there is not a single agreed upon lifecycle, a number have been proposed (Hodge, 2000; LIFE project, 2006; Knight, 2006; Higgins, 2007). There is general agreement for:

- a creation phase;
- an acquisition phase where the responsibility for the digital object passes to an institution;
- a preservation phase where activities are conducted to ensure continuous access to the digital object; and finally
- a reappraisal/disposal phase where a determination is made whether or not to continue preserving the object.

In this lifecycle, institutions may acquire and modify a version of the object, creating multiple versions that have authenticity and standing within the community.

With this basic understanding of key concepts in digital preservation, I now turn to look at what issues exist in digital preservation.

## 2.5 Issues in Digital Preservation

---

It is easy to generate laundry lists of threats to digital objects. One such list (Rosenthal et al., 2005) includes:

Media Failure	Software Obsolescence
Hardware Failure	Operator Error
Software Failure	Natural Disaster
Communications Errors	External Attack
Failure of Network Services	Economic Failure
Media and Hardware Obsolescence	Organizational Failure

**Table 2-2: Threats to Digital Objects**

Handling a list like this may result in piece-meal solutions to a problem that requires an integrated, comprehensive approach. One way of dealing with the multitude of threats presented to digital objects is to categorize them. Doing so allows for a coordinated effort to address the entire category rather than individual threats. One such typology (Besser, 2000) groups the threats into five broad problems in digital preservation.

The first problem is the **viewing** problem. The naked human eye can view a physical information object (like a book), but digital objects are reliant on technology in order to be viewed. If the technological context no longer exists to view the object, then even if the object exists in pristine digital form, it is still no longer accessible and is, for all intents and purposes, lost.

A second problem is that of **scrambling**, where the digital object may have a layer of complexity added to it that makes it harder to preserve. Often the reasons

for this include compressing the object to save space or encrypting the object due to security or intellectual property rights concerns. While these decisions are sometimes necessary, the impact of short-term decisions made for expediency can impair long-term preservability of the object.

The third problem in Besser's ontology is the problem of **inter-relation**. With analog media, the object tends to be a singular, discrete item but digital objects can be composed of many individual digital files bound together only by descriptive and contextual metadata. The overall solvency of the digital object requires that all of the constituent parts be accessible, which may require separate actions on parts of the object, especially as obsolescence and other problems may only affect certain files within the whole object. There is also another aspect to the inter-relation problem, that of a relationship to the context of the object.

Consider the example of a webpage: one can treat the webpage as a discrete object from a conceptual standpoint but a single webpage is actually a composite object with the marked-up text and associated image and multimedia files as its constituent parts. Moreover, the web page exists in both the context of the website that it resides in, as well as the target to links from other pages external to the website. These links and the positioning of the webpage in the context of the site provide information about how to understand the page: if this contextual information is lost, the individual webpage—even if perfectly preserved otherwise—will have lost at least some of its meaning.

The fourth problem is the **custodial** problem—this being directly related to provenance. Where institutions have divided the landscape of preserving analog material on a well-organized basis, no such orderly divisions exist in the digital realm. This can be problematic as multiple actors may claim to possess the most current and authentic object, leaving the viewer confused as to which is the best source to go to. In addition, the actors may undertake different preservation actions, leading to divergence as to which version of the digital object is the authentic object.

Finally, there is the problem of **translation**. To avoid obsolescence of file formats and digital standards, digital objects are often moved from one format to another. As noted in the lifecycle discussion, preservation activities may include a significant transformation of the digital object, a transformation that can result in information loss. As a consequence, the current viewer may experience something very different than the original intention.

In summary, the above typology of digital preservation problems provides a useful framework. In turn, this framework will help us understand the current approaches to digital preservation and, more importantly, what kinds of problems remain to be addressed.

## 2.6 Approaches in Digital Preservation

---

As of yet, there are no definitive conclusions about the correct way to handle digital preservation despite a multitude of projects aimed at providing digital preservation solutions. Because most of these research projects are large, multi-institutional (and in many cases multi-national) exercises, the field is not necessarily evolving as quickly as one might expect given the urgency of the problem. However, a number of directions are developing in digital preservation with some consensus in terms of approaches and methodology.

### ***2.6.1 Open Archival Information System Model***

Any discussion of digital preservation approaches is not complete without a discussion of the Open Archival Information System (OAIS) reference model (CCSDS, 2002). The OAIS model is the de facto starting point for the design of virtually every digital preservation system. The model was initially developed at NASA to address the specific needs of space data but the broader community found it generalizable enough for a wide range of digital preservation problems. Its adoption reflects a need in the digital archiving community for a common language to describe the features of the systems being built. Moreover, the broad acceptance of the OAIS model imputes a level of implicit quality to systems conformant to the model.

At its core, the OAIS deals with the movement of information. This information is represented as an entity by the *information object*. The information object is composed of a *data object* (which is the raw encoding of some knowledge) and its *representation information* (which is the information necessary to decode the data into a form that can be understood by a viewer). *Preservation Description Information* (PDI) is added to the information object (content information or CI) to assist in preservation activities. The PDI represents information about where the CI came from, who created and modified it, how the CI relates to other information, how the CI is identified, and finally, information about its current integrity. The two (CI and PDI) together represent an *Information Package*, the core unit on which an archival system works. The model identifies three kinds of information packages:

- the submission information package (SIP) provided by the creators / producers of the content containing the content itself and pertinent information describing the content;
- the archival information package (AIP) maintained by a preservation system which combines the SIP with information necessary for managing preservation processes; and
- the dissemination information package (DIP) that is used to provide access to the content to end-users by the preservation system and is a transformation of the AIP.

The model also defines a number of roles that interact with the packages. A *producer* creates the SIP and then submits it to the archival management. The *management* is responsible for transforming the SIP into an AIP with information necessary to manage the package. A *consumer* then requests the object from the archival information system and is presented with a DIP.

Finally, the model defines a number of key components in the archival information system required to handle the information package. An *ingest* system is required to receive the SIP from the producer and do the initial work to transform the SIP into an AIP. The AIP is placed into an *archival storage* system for

safekeeping. *Preservation planning, administration and data management* systems are required to organize and handle the AIPs as necessary. An *access* system is then required to provide the DIPs to the consumer. Figure 2-1 lays out the relationship between the different systems, information packages and actors.

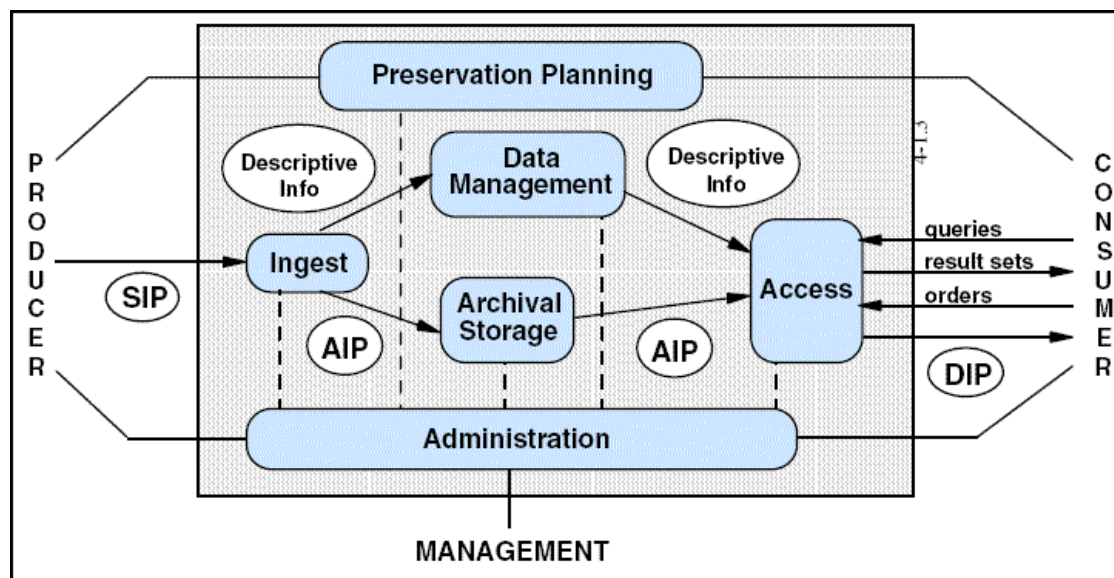


Figure 2-1: OAIS Diagram (reproduced from CCSDS, 2002)

While OAIS ostensibly covers all aspects of the roles and functions, there tends to be an emphasis on the management side of things. This is understandable given the goal of system-building but results in producer- and consumer-side issues being glossed over. The net result is that the model addresses some digital preservation issues better than others. The issues of *scrambling* (transformation into archival formats) and of *viewing* (access systems for current viewing) are reasonably handled. *Custodianship* is addressed by defining roles, even as it oversimplifies. The issue of *inter-relation* is only marginally dealt with and the broader issue of intellectual context (such as the case of the webpages and the external links) is not addressed at all. The issue of *translation* is glossed over by assuming that the existence of archival information attached to the object will facilitate translations.

### **2.6.2 Preservation Metadata**

The second area where the preservation community has identified a clear direction is in preservation metadata. As most digital preservation is done within the context of institutions, managing large volumes of material is often a necessity. In order to coordinate activity, information is needed to manage the objects. This information is referred to in OAIS as preservation description information or, in more common practitioner terminology, metadata. Haynes (2004) identifies five key functions for metadata: resource description, information retrieval, management of information resources, documenting ownership and authenticity of digital resources, and interoperability. All of these functions are essential to digital preservation tasks. As has been noted, “[e]ffective management of all but the crudest forms of digital preservation is likely to be facilitated by the creation, maintenance, and evolution of detailed metadata in support of the preservation process” (RLG-OCLC, 2002).

A number of early projects developed rudimentary metadata standards (CEDARS, 2002, Lupovici & Masanes, 2000). However, these early efforts focused on the requirements of their respective projects. The need for a more general standard led to the creation of the OCLC/RLG working group on preservation metadata (OCLC/RLG WGPM, 2001). The result of the working group is the PREMIS data dictionary for preservation metadata (PREMIS, 2008). While PREMIS enjoys broad recognition from the preservation community, its actual implementation remains limited, as evidenced by the paucity of entries in the PREMIS implementation registry (<http://www.loc.gov/standards/premis/premis-registry.php>).

Preservation metadata address some of the aspects of the custodial problem by providing a log of the actions performed on any given object, so that even if the custodianship is unclear, it is possible to remove transformations to an object that reflect improper custodianship. It also addresses some of the issues of translation, viewing and scrambling by recording what was done to an object, how it was done and by whom, allowing for later custodians to either reverse actions or gather more information from the original agents for preservation actions. In a worst-case

scenario where the object has been irreversibly damaged or lost, preservation metadata can help in reconstructing the object from previous versions. Inter-relations can be recorded using preservation metadata as well. One key issue is that not all of the actors in the process are fluent in the complex PREMIS standard and thus the quality of current metadata may be suspect.

### ***2.6.3 Trusted Digital Repositories***

As noted in the discussion on the OAIS model, the centrality of the archival information system has lead to a focus on how to build systems and the kinds of institutions that could house and adequately support such a system. The result of this is the idea of a trusted digital repository (RLG-OCLC, 2002). Trusted digital repositories (TDR) are much like libraries for books or museums for artefacts. As such, trusted digital repositories encompass ideas like administrative responsibility, organizational viability, financial sustainability, technological and procedural suitability, system security and procedural accountability. The formal definition of a trusted digital repository is “one whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future” (RLG-OCLC, 2002). This definition implies a number of things about the institution, namely that it:

1. actively manages digital objects in its care;
2. is part of a community, with its attendant standards and expectations;
3. is responsible for making digital objects available to its community. This may involve the creation of tools necessary to view the digital object in the future.

The last implication suggests that planning aspects of digital preservation would be part of the activities of trusted digital repositories. This planning, which is intended to identify the tools need for the goal of persistent, long-term access, would result in these institutions being engaged in active research and development efforts. With the attributes of a trusted digital repository defined, it is possible to establish an auditing mechanism for evaluating both existing and proposed repositories (Center for Research Libraries & OCLC [CRL/OCLC], 2007).



Trusted digital repositories primarily address the problem of custodianship by articulating the most important attributes of the custodian. However the remaining problems are subsumed under the general blanket statement that TDRs would act to ensure that the other problems are addressed. However, it would be up to the individual TDR to act accordingly.

#### **2.6.4 Preservation Activities**

At the core of both OAIS and TDRs is the idea that archival management is executed through a series of preservation activities. As noted above, preservation metadata provides the information required to guide these activities. There are three kinds of methods for solving preservation problems (NINCH Working Group on Best Practices [NINCH], 2002; Hodge & Frangakis, 2004).

The first method is *technology preservation*. In technology preservation, the goal is to save the actual environment required to view a digital object. This may involve saving the actual hardware and software and placing it in an environment where it can be maintained and allowed to run past the normal failure point of the hardware. While this is useful for extreme situations, it can be a costly proposition when the hardware is no longer produced en masse and replacement hardware must be manufactured solely for the goal of preserving the object.

A second method is *technology emulation*. Emulation involves creating an alternate method of viewing the digital object that simulates the original viewing environment. There has been some discussion as to the practicality of creating emulators for every file and format that exists (Besser, 2000) but there have also been some experiments that have successfully demonstrated emulation (*Seeing Double*, 2004).

In *data migration*, the actual digital object is changed to allow it to run with modern software and hardware. This can be a complex process, especially if the original technical specifications are no longer available. As well, it can be risky as the migration process can be lossy. To ensure that no critical information has or will be

lost, the migration process must undergo a risk assessment process (Lawrence, Kehoe, Rieger, Walters & Kenney, 2000) to validate migration results. This can be an issue as the time necessary to analyze the migration process for risks of data loss is also substantial and requires considerable resources. However, central registries for file formats like the Harvard Global Digital Format Registry provide a means to consolidate the effort. This means that individual organizations do not have to replicate the work.

What is common to all three methods is that they primarily address the viewing and scrambling problems. None of these activities address either the inter-relation issue (except where the object is largely standalone) or the custodial problem.

## 2.7 How Digital Preservation Relates to the Researcher

---

Based on the preceding section, an ideal world scenario for institutions would be that researchers deposit their research into trusted digital repositories. If they did so on a regular basis using archival formats with sufficient metadata, there might be a workable preservation situation. However, researchers do not work this way (Marshall, 2008) and trying to convince them to do so has generally been futile (Foster, Gibbons, Bell & Lindahl, 2007).

However, even if researchers could be convinced to use the systems and processes institutions want them to, there are still problems. There is a mismatch as current repository systems focus on the deposit of individual digital objects while researchers often create clusters of related documents (Marshall, 2008). Similarly, the researcher websites appear to have contextual information not related to an individual object, which might also get lost.

As we saw with the websites in chapter one, a framing narrative is created by the choices in presentation and supplementary information on the researcher's website. The presentation also provides an organizational structure that binds individual documents together into a coherent whole. One supposition would be to

simply save the site as a whole. This is the practise for web harvesting projects. However, this would limit the usefulness of the individual items like the publications. As well, individual items may have external relational context that would be difficult to capture by treating the entire website as a single unit. In order to understand this, we need to look in greater depth at the website itself. To provide a theoretical basis for how to approach the websites, the next section will review the current literature on website analysis.

## 2.8 Analysis of Websites

---

### ***2.8.1 Research on Websites***

Given the explosion of personal websites, there has been surprisingly little research on the nature of them. Of the work that does exist, most falls into one of two areas: exploring the construction of identity through web activity, and content analysis of the website.

#### *2.8.1.1 Identity Construction*

The genre of work covering the construction of identity on the web spans the disciplines from sociology and psychology to more literary understandings of the nature of the personal website. At the core is an attempt to understand the people behind the websites—what motivates them to expose themselves on the web, what they hold in reserve, and what they hope to achieve. The voluminous nature of personal content creates a rich field for research. It is not just the content that gives away the identity of the creator, but also the form of the content (Chandler, 1998) from the way the code is formed to the choice of design and usage of media. These things all give clues as to the person behind the page.

Similarly, Dominick (1999) notes that personal websites are not just random content but the form of the site is an intentional self-presentation. In his view, to understand the website is to understand the incarnation of self that the creator wishes to project. One interesting finding from Dominick's study is that the most common feature on a website is a feedback mechanism (visitor counter, e-mail

address, guest book or rant page); this suggests that personal websites are not merely one-way communication channels, but are invitations to interact with the creator.

The creator's cultural and demographic allegiances often have an impact on the nature, content and form of the websites (Buten, 1996; Stern, 1999; Ishii, 2000; Alexander, 2002; Kim & Papacharissi, 2003; Stern, 2004; Bortree, 2005; Jung, Youn & McClung, 2007; van Doorn & van Zoonen, 2007; Hodgkinson & Lincoln, 2008). Within their cultural and demographic groups though, website creators are homogenous within the group (Marcus, Machilek & Schutz, 2006). This means that the behavior of early adopters in a group is predictive of the larger group.

#### *2.8.1.2 Content Analysis*

The previous section demonstrated that personal websites are indeed a viable topic of study because their creators reflect their social and cultural subgroups. The next question is how to analyse the content of the web pages. Asirvatham and Ravi (2001) noted that there have been a number of attempts to classify web pages. The approaches can be broadly organized as follows:

1. manual classification by domain specific experts,
2. clustering approaches,
3. META tags (which serve the purpose of document indexing),
4. a combination of document content and META tags,
5. document content, and
6. link and content analysis.

While their findings suggest that automatic classification is possible, it is still unclear as to how effective automatic classification would be where no initial categories have been identified. Such techniques are used more to validate initial findings once categories have been established.

The classification of personal websites has not yielded any clear categories for analysis; this is perhaps reflective of the size of the problem space as noted previously. However, other kinds of websites like corporate websites are easier to classify (Marco, 2002) likely due to a greater set of commonalities in purpose and function. Similarly, political websites are easier to classify (Jensen & Helles, 2005) than personal websites. This suggests that it is viable to focus on a narrow subset of personal websites where there is a greater commonality, and be able to classify and develop categories successfully.

#### *2.8.1.3 Academia and Websites*

The majority of the discussions in the context of academia and web publishing have focused largely on the needs of teaching and the transformation of scholarly communication. There has been significant discussion on things like open access to scholarly information (Alperin, Fischman & Willinsky, 2008; Swan & Brown, 2005), citation analysis and persistence (Lawrence, Coetzee, Glover, Pennock, Flake, Nielsen, Krovetz, Kruger & Giles, 2000; Wilkinson, Harries, Thelwall & Price, 2003; Piwowar, Day & Fridsma, 2007), the design of academic websites (Peterson, 2006) and the relationship between the web and the classroom (Greenhow, Robella & Hughes, 2009).

#### **2.8.2 Research on Academic Websites**

While the broader research on academic websites has a more institutional focus and is not especially germane to the current discussion on individual researcher websites, there are a small number of studies of direct relevance.

##### *2.8.2.1 Dumont and Frindte*

Dumont and Frindte's (2005) study of academic psychologists evaluated the type of content found on their pages. They focused on content in six areas: results-oriented research, process-oriented research, publications, teaching, links and private content. First, 350 homepages across four European countries were surveyed by students to identify whether content existed in the 6 categories. Second, psychologists reported the type of content available on their websites through a

questionnaire. In both, they found that most website content focused on the publications of the psychologists, and on disseminating result-oriented research information (research that yielded results). They also found little disclosure of private or personal information. Dumonte and Frindte (2005) suggest this is due to the social convention within science to separate the personal from the research. Thus websites are not used to represent the researcher. Rather, they represent the research itself.

#### *2.8.2.2 Rick*

Rick's (2007) study focused on fellow academics' adoption of a website content management technology (AniAniWeb), which he developed at the College of Computing at Georgia Tech. System adopters were interviewed about its use and about their general approach to their home websites. While Dummont and Frindte's (2005) study suggested that academics do not reveal much in the way of personal information through their website, the AniAniWeb system encouraged more disclosure of a personal nature by including elements like blogging, third party editing of the pages, and social networking features. Despite this, the pages produced generally reflected more professional concerns, indicating that the social convention found in Dummont and Frindte's study is very much corroborated in Rick's study. Rick also compared those who use static web page methods to those who use the wiki-based approach, where the later tended to create more pages on their site.

#### *2.8.2.3 Barjak, Li, Thelwall*

Barjak, Li and Thelwall (2007) examined the impact of websites as measured by the links to the website. As with the Dummont and Frindte (2005) study, they examined the website in conjunction with a questionnaire. They assumed that the greater the number of links to the researcher website, the greater the impact of the website. What they found was that in general, pages that contained links to research output tended to receive the most links.

## 2.9 The Nature of Personal Archives

---

Finding out whether or not a researcher's personal website is useful for digital preservation purposes is part of the goal here. Based on the previous studies, it is not yet clear whether the website is useful or not. The second question raised is how researchers think of their websites. Perhaps if they viewed them as personal archives, they might be motivated to cooperate with institutions to create preservable content. While there is little in the way of research on personal digital archives, there is some research on personal analog archives of researchers. In the next section, I will briefly discuss the current research on personal archives and researchers.

One finding for paper archives is the value of context. 36% of people's archives consist of reference material that is not unique to the person or the archive (Whittaker & Hirschberg, 2001). In the case of material that is publicly available, one reason people keep documents is that the material influences their work. Clearly this identifies one area where a given object (physical or digital) has a lineage; that to understand the current work one must have available the work that influenced it.

This leads to a broader discussion of both the reasons for archiving material at a personal level and how this archiving materializes. A recent study of the archiving practises of academics (Kaye, Vertesi, Avery, Dafoe, David, Onaga, Rosero & Pinch, 2006) found that there were five primary goals for creating archives.

1. *Finding it again* was perhaps the most obvious reason for the archive, and the one that overlaps most closely with institutional needs. Yet Kaye et al. found that most researchers were unhappy with their archival systems' ability to find things, suggesting this is an unrealized value. They also found that in general the systems all performed about the same in real world tasks.
2. *Building a legacy* was another commonly stated goal for constructing a personal archive. While the person was unlikely to use the material again, the

archival system provided a way to communicate the accomplishments of the person to the viewer. Often there is a rigid organization to the archive to convey a sense that everything has a rightful place. Context is clearly important here: It is less about an individual object than how everything fits together. This will be important when I evaluate how institutional systems mesh with the creator, especially if those systems lose that sense of context.

3. *Sharing resources* lets the person share their materials for others to use as a resource. While this reason overlaps heavily with the aims of institutional systems, personal archives add a sense of personal ownership maintained over the collection as a whole by the creator.
4. *Fear of loss*. By concentrating critical material in an archive, a person can selectively save what is most critical in the event of disaster. Here is where the digital vs. physical contexts diverge. With physical objects, the act of gathering into one spot for quick escape is also an act of selection. Those not part of the critical pile by default are of lesser value. These constraints are less pressing in digital media. Digital storage has great capacity and necessitates a different selection dynamic. Surprisingly, the perception of the dynamic can be the reverse of reality: in one study, people tended to treat electronic space as more limited than the physical equivalent (Whittaker and Hirschberg, 2001).
5. *Identity construction*. As almost an extension of legacy building, the personal archive is largely an expression of the researcher. The researcher often engages in careful curation of what is presented to ensure that his or her identity comes through. While most things included in an archive are significant from a research perspective, some items may be included because of personal meaning.

One takeaway from the above points is that the current digital tools only address a subset of the values of a personal archive. The majority of the tools primarily



support the ability to find a document later (reason 1). However, today's tools generally do not address the other four reasons, suggesting an opportunity for tools that do support the other values.

The values are also useful as a point of comparison and contrast between personal archives and the personal website. For instance, in a study of the popular photo-sharing site Flickr (van House, 2007), users viewed their photo streams as self-representation. Here users carefully groomed their stream to reflect self-identity in a manner similar to identity construction identified above (reason 5). Similarly, project websites by oceanographers (Lamb & Davidson, 2002) are often meticulously maintained and up to date, demonstrating a legacy function. Even when web pages give little information about the individual, the creator is aware that the page represents something of who he or she is (Walker, 2000), suggesting identity construction. Finally Marshall (2008) points to where researchers have used online sources to reconstruct lost digital files. In addition, the bibliography is often the only complete index of the researcher's work. These last two points line up with the goals of finding again and fears of loss. All of this suggests that there is utility to thinking of the researcher's web presence as a form of personal archive.

## 2.10 Summary

---

I have noted that digital preservation is a broad area covering all aspects of the life of a digital object and kinds of digital objects. The definition of digital preservation that emerges is as a series of managed activities to ensure accessibility to the objects via technological and organizational means. Within this, there are three key concepts. First, the basic unit for preservation is the information (or digital) object: a composite consisting of a set of files held together by contextual metadata. Second, there are both born digital objects that are created by digital tools and only exist in digital form, and digitized objects that are transformed from analog objects. Born digital objects represent a particularly critical problem, as once they are lost, they cannot be recovered. Third, digital objects exist in a lifecycle with similarities to the

physical object but for two key differences: that the chain is one of responsibility rather than custody, and that multiple copies of the object can exist along that chain.

While there are many issues in digital preservation, Besser's (2000) typology of viewing, scrambling, inter-relation, custodial and translation problems represent a good set of categories under which most individual issues can be grouped. In response to these problems, a number of approaches have been developed which include the OAIS model as an architectural blueprint for how archival systems might work, preservation metadata to manage and facilitate the process of preserving digital objects, trusted digital repositories as agents of preservation and activities like technology preservation, emulation and migration to actually preserve access to the objects. One issue that stands out as being largely unaddressed is the problem of inter-relation, which may require the creator.

However, all the above represent institutional concerns about preserving digital objects, and these concerns are not necessarily those of the individual researcher. Instead, the focus of the researchers is on building websites. But those websites might be considered archival and that is important to changing researcher behaviour. As such it is important to understand why the creator would archive. Five reasons for a personal archive are identified: to find things, to build a legacy, to share resources, to avoid loss, and to construct an identity. More importantly, if a researcher's view of his or her personal website is close enough to these values, then he or she might consider altering approach and adopt tools that are better suited for preservation.

The challenge is finding the bridge between the institution and the creator, which in turn requires much greater knowledge of how and why creators create and manage personal websites. These issues form the bulk of the next chapters. I will seek to uncover researchers' reasons behind their personal websites, investigate the missing elements that would allow personal websites to be transformed into archives, and suggest what kinds of tools might bridge the gap between personal and institutional archives.

# Chapter Three: Website Survey

---

## 3.1 Chapter Synopsis

---

In the previous chapter, I outlined the research on digital preservation and on personal homepages and how it was applicable to my current research. In this chapter, I will provide an overview of the methodology used to survey the websites and discuss the results from the survey, including the identification of four types of websites in common usage.

## 3.2 Introduction

---

As discussed in chapter two, current efforts in preserving digital information are mostly by large organizations either harvesting swaths of websites from the World Wide Web (Web) or acting as repositories for deposited research material. Most of the work has focused on saving individual documents, with little regard for how these documents fit into the process that created them or the context of other similar documents by the same creator. This is the inter-relational problem (Besser, 2000), described previously.

However most documents worthy of being saved by a cultural heritage institution do not exist in a vacuum. They are generally produced in conjunction with other material or spawn more documents later in the process. This is especially true of research activity. Consider documents associated with a research study. Typically, a study builds and expands on those that came before it, and is further expanded by research that follows. A document or digital object created in the research process is usually part of a larger cluster of related documents and objects. If the research is to be saved (as opposed to the document), all of the constituent elements need to be saved. Yet thus far the primary response by the institutions has been to focus on the individual files. Associating other documents with those files is

done crudely, usually by putting out the call to researchers to deposit their files in the archive.

This is insufficient because the ongoing challenge to convince researchers to deposit their works in these repository systems remains difficult. The “if you build it, they will come” axiom has yet to work for cultural heritage institutions. Simply put, researchers either often do not participate at all (Foster et al., 2007) or contribute only one or two papers (Thomas & McDonald, 2007). Clearly, if there is to be success in acquiring digital material from researchers in an organized and coherent fashion, there needs to be a better motivation: personal benefit.

When there is personal benefit, researchers appear to be more than willing to invest in the effort to craft good descriptions of their work and organize it in a coherent fashion. A cursory scan of researcher personal websites seems to indicate this. So if true, then one solution to the institutional archiving problem is to use the researcher’s existing personal website, a suggestion I made in chapter one. One approach could be to retrieve the necessary context, organization and description for preserving a researcher’s work from the actual website via some type of data mining. Another solution might be to provide tools that facilitate the construction of a personal website, while engineering those tools to also provide institutions with preservable content. To determine whether there are feasible approaches or not, it is important to understand both what researchers are currently doing with their personal website and what tools currently exist for researchers to build those sites.

### ***3.2.1 Defining the Personal Website***

One important definition must be made at this point: what is a researcher’s personal website? To study what exists, we need to define what a personal site is and what it is not. For the purpose of this thesis, we define a *researcher’s personal website* as a website that is the online, public face that represents the researcher and his or her research to both the specific research community and to the broader public with interest in the research domain. This definition provides a number of important demarcations.

First, there is an audience for the website and, more importantly, the audience is the research community. This allows us to eliminate websites aimed at his/her hobbyist community of (say) model builders or the photo collection of border collies shared with fellow dog enthusiasts. It also eliminates purely personal websites for family and friends.

Second, the website is aimed at representing not only the researcher as an individual, but also his or her research. This is important because we are looking to collect and preserve a collection of works as opposed to everything about an individual. In this case, the collection of research is expected to be a coherent corpus of value to a broader audience.

Third, the site is the public face of the researcher and therefore defines the intent: to make the researcher's work public. From this, we begin with the assumption that materials available on the site are meant to be in the public domain: to be read, to be viewed, and to be used by others.

### 3.3 Methodology

---

#### ***3.3.1 The Need to Survey Researcher Websites***

While the Dumont and Frindte (2005) study of website content is particularly useful in understanding the kind of content available, it tells us little about the form and structure of the content. Similarly, Rick's (2007) study tells us much about the underlying motivations that drive researchers to create and add content to their website but less about the system choices, as the platform is pre-determined. Barjak, et al.'s (2007) study tells us almost nothing about the form and content of websites.

In order to determine what is preservable, we need to know what kind of content is available, how it is structured and managed and what motivates researchers to invest effort into both the content and its structure. These questions can only be answered by understanding in much greater detail what content is on researcher websites and how that content is structured.

One study that focuses on the composition of researcher websites is by Anthony Gray (2009). Gray, a student of Ben Shneiderman, looked at the common elements on academic websites and catalogued them, looking particularly at their frequency. He notes that every website contains contact information for the researcher and content relating to publications. Most have between 11 and 100 pages. Most (93%) have an image of the researcher. In contrast few (13%) have an institutional logo and none have site maps or search systems.

However, this is still insufficient to determine how to approach using researcher personal websites to facilitate the preservation of their works, as we do not have a clear sense of the structure of the websites nor do we have a good sense of the weighting and completeness of the categories of content on the websites.

### ***3.3.2 Subject Selection***

Given the number of disciplines and the variety of practises, the breadth of researcher websites is still be too broad for analysis. There are simply too many disciplines and too many practitioners. To reduce the number of sites studied to a reasonable number while maintaining descriptive power, it makes sense to limit this study to the personal websites of one discipline. While this will limit generalizability of the results, the sample size required for a multi-discipline analysis would be resource prohibitive.

I chose computer science for a number of reasons. First, computer scientists have long been pioneers in sharing things online and it was likely that more content would be available. Second, computer scientists are more familiar with the tools. This familiarity means that the choice of tools, design and approach are more likely a reflection of the personal preferences of the academic instead of a standard institutional offering.

However, even computer science is a broad discipline so the choice of researchers was further narrowed to human computer interaction (HCI) researchers. This choice reflects two advantages: that there exists a resource that

identifies most of the top researchers within the area (the HCI Bibliography – <http://www.hcibib.org>) and that researchers within HCI tend to be a little more diverse in terms of varying backgrounds and interests.

The final selection criterion in the choice of subjects was to select senior researchers within HCI as defined by the number of publications. The first reason is that the top researchers in the area would be better established and have a greater ability to do as they choose, as opposed to either following a convention or being forced to act according to institutional policies. Secondly, senior researchers have a greater breadth and depth to their content and thus, it should be easier to identify themes and patterns in the organization and presentation of the content. As well, they may have more cause to address archiving their work given the point they are at in their careers.. Third, they will likely have had the opportunity to experiment with different ways of organizing their sites and they will have arrived at some form of best practise. Finally, others in the field are more likely to emulate them due to their seniority and fame.

Based on the above set of criteria, approximately 70 names were initially selected to review. Further analysis of the list eliminated a number of names where the entry in the HCI bibliography actually represented multiple authors. The final count of researcher websites evaluated was 65. The final list of researchers' websites selected can be found in Appendix A.

### ***3.3.3 Identifying the Website***

Once the names of the researchers had been identified, it was necessary to find their website. Starting with the list of names, I used search engines to locate the researcher websites. While many of the researchers had names that are distinctive enough and their work prolific enough to be found on the first page of a search set, a number had relatively common names and had to be filtered through additional keywords.

What was more challenging was identifying the personal website. In many cases, the researcher had websites in multiple locations, either through multiple affiliations or over time as they moved from organization to organization. Often websites were left orphaned but still findable through search engines. This required establishing a set of criteria for the identification of the official current personal website as follows:

1. Recency of the website. The more current the site was, the more likely it was that the site was the current personal website. In some cases, the bibliography on the site was the best indicator of recency.
2. The degree of personalization. The greater the personalization, the more likely it was to be chosen. Conversely the more like an institutional / organizational site the website looked, the more likely that it was a required site for the organization and maintained by the organization; these were passed over.
3. The amount of content. The more content on the site, the more it was favoured as the personal website.

It should be noted that the sites were reviewed a number of times over a period of almost a year. In a number of cases, the individual either moved institutions or changed roles resulting in a change of website. In these cases, the new site was reviewed again, replacing the existing entry in the database.

### ***3.3.4 Initial Survey***

Once the list of websites was identified, an initial survey of a small set of websites on the list (20 sites) was reviewed to gather impressions about the layout, the content and organization of the websites. Screen captures of the websites were taken and free form notes were recorded, describing impressions of the content and structure of each site. The notes were then informally coded to identify the site's key themes. This coding produced a list of elements and features of the sites reviewed.



A second pass over another limited set of sites (~10) was used to refine the elements. In this second pass, the goal was to determine if the elements and features could be categorized, and controlled vocabularies were developed to describe patterns in the site elements. This second pass produced the final survey instrument (appendix B) used to document the sites.

### ***3.3.5 Final Site Survey Methodology***

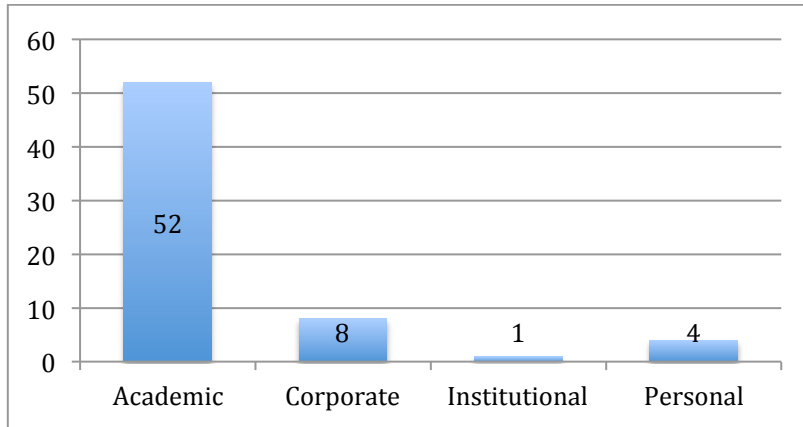
The final survey instrument was designed as a series of questions about the features of the sites. While the goal was to develop controlled vocabularies to assist in quantifying answers, the final survey allowed for open-ended answers to ensure all key data was recorded.

Each of the sites was then evaluated using the instrument. If an answer to a particular question could not be accommodated by the controlled vocabulary, it was either noted in the notes section or in the field itself; the goal was not to preclude possible anomalies and outliers. The source HTML of the websites was not systematically inspected to identify trends within the coding. Once all sites were surveyed, the resultant data was reviewed and, where possible, normalized. As the initial data entry was open-coded, similar concepts and ideas often were encoded with variations in terminology. The normalization process was used to select a single term to unify these similar concepts. This allowed for rudimentary quantitative analysis of the results. Fields with controlled vocabularies and normalized data were then tabulated through pivot tables in Microsoft Excel. The pivot tables allowed for automated counting of the categories established either through the controlled vocabularies or the normalized data.

## 3.4 Data Analysis

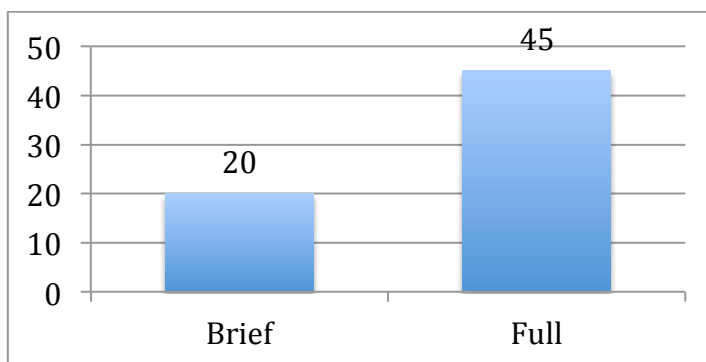
---

### 3.4.1 General Organization and Design



**Figure 3-1: Researcher Website Type**

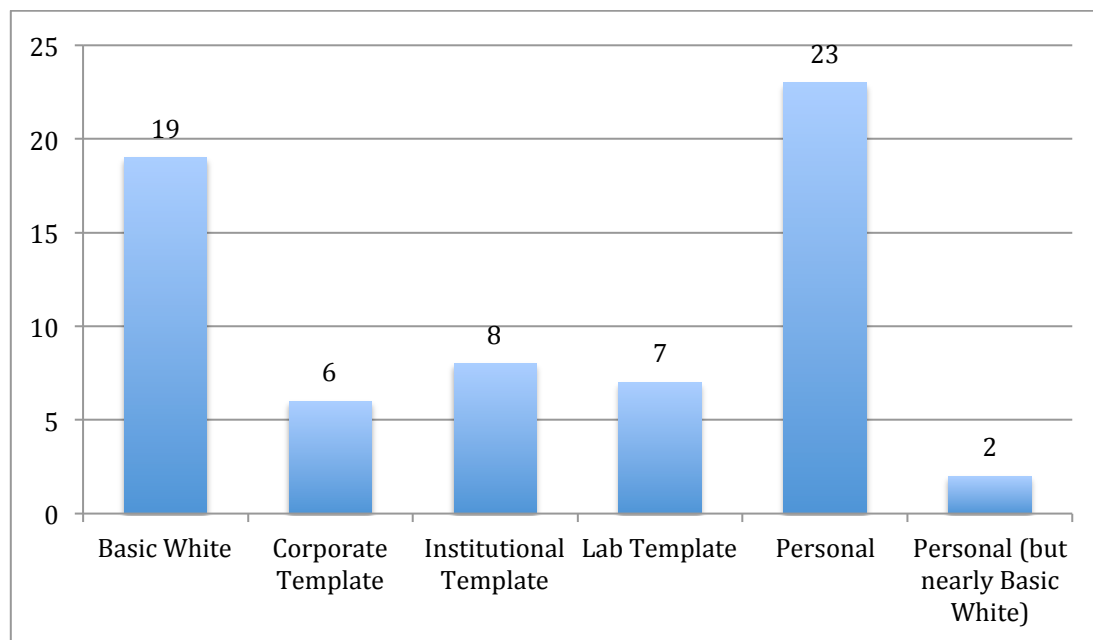
As noted above, a total of 65 sites were surveyed through the web content analysis process. Of these 65 sites, 52 were websites of active researchers at an academic research institution, 8 were researchers in a corporate environment, 1 was affiliated with a governmental research institution and 4 were personal websites with no clear affiliation to an organization (Figure 3-1).



**Figure 3-2: Content Completeness**

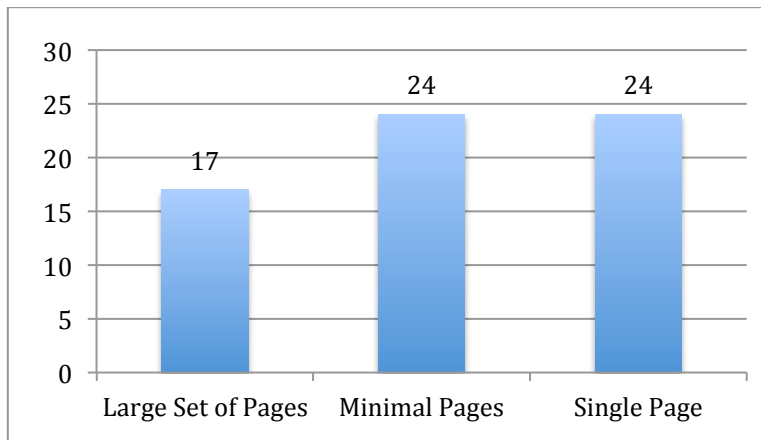
I judged 45 sites as having *full content*, i.e. a relatively complete bibliography and sufficient information to understand the researcher's areas of interest and the nature of past and current projects. I judged 20 sites as *brief content*, where the information merely provided a sketch of the researcher. In the brief content cases,

preserving the site would not suffice to preserve the full research and publication history of the researcher (Figure 3-2).



**Figure 3-3: Type of Website Design**

I found six types of designs for the websites (Figure 3-3). The most common (23) was a *personal design*: the layout, colour choices and use of imagery reflected the personal preferences of the researcher. The second most common was a *basic white design* that used minimal typefaces, single column design, and white backgrounds. If there was an image, it was a photo of the researcher. Another 21 comprised *template design*, where the page used a template. These were further distinguished by the origins of the template: *institutional* (8), *corporate* (6) or *lab* (7). Finally, 2 sites were *personal but nearly basic white designs*, i.e., a mix of the personal and basic white.



**Figure 3-4: Number of Pages in Site**

Sites with a *large set of pages* were relatively uncommon (17 of 65 sites), with sites having either only a *single page* or *minimal pages* (48 of 65 sites) being the norm. While the landing page on some sites gave the appearance of being rich, multipage sites, their links were often to external sites or to other sections within the landing page. I judged a *large set of pages* to have at least 5-10 pages of content although some leeway was given if a site had a page that linked significant locally hosted content like publications or videos. *Minimal page* sites were defined as sites with a primary landing page and a small (less than 5) number of supporting pages (Figure 3-4).

### ***3.4.2 Type of Content***

Element Type	# of Sites with Element Present	% of Sites with Element Present
Bibliography	62	95%
Extensive Contact Information	54	83%
Identification Photograph	54	83%
Identified Research Projects	40	62%
Professional Biography	38	58%
Project Pages	34	52%
Primary Navigation	31	48%
Teaching / Class Resources	18	28%
Links to Students	13	20%
Personal Photos	13	20%
Directions to Office	7	11%
Personal Biography	6	9%
Personal Blog	4	6%

**Table 3-1: Types of Site Content**

The types of content available on the websites are listed in Table 3-1. The majority of the websites contained a bibliography of some kind (62 out of 65 sites). While the organization and style of the bibliography varied significantly (as will be discussed later), for most sites this list of publications represented the most significant part of the site.

The majority of websites also contained a photograph of the researcher (54 out of 65 sites) that was usually featured above the fold on the first page of the website. An identical number of sites also contained extensive contact information for the researcher, where I define extensive contact information as containing at least two points of contact (e-mail, telephone, fax number, mailing address or office

location) as well as an institutional affiliation identifier. Although the contact information and identity photo was found on an equal number of websites, it should be noted that some sites contained one or the other but not both.

Beyond the above three items, the consistency of elements dropped significantly. 38 sites had a professional biography (a biography that could be used for grant applications or inclusion in conference literature). 40 sites identified their research interests and/or current projects in detail. 34 had distinct project pages. For project pages, these were often not located within the personal website but in lab or institutional websites or as completely separate sites (24 out of 40 sites).

31 sites had some form of primary navigation system, where a distinct table of contents was presented. I initially considered only navigation systems that ran through the entire site on multipage sites; however, given the large number of either single page sites or minimal page sites that had a primary navigation system, I counted these as well as it was important to capture the structure they presented.

18 sites had teaching resources, usually in the form of class pages with syllabus, calendar and readings. Only a small number of sites included historical class offerings: the majority provided resources only for current classes. 13 offered links to the home sites of the researchers' graduate students. 13 had photos of personal nature or links to photo hosting sites like Flickr. 7 provided explicit directions to offices or labs. 6 had biographies of a more personal nature mentioning spouses, children or interest areas. Finally 4 had a personal blog as part of the website.

### ***3.4.3 Bibliographies***

As noted, bibliographies were prevalent, and thus deserve further analysis (Table 3-2). A bibliography was present on most sites (62). Of the sites with bibliographies, the majority had bibliographies located 1 link away from the main page (50 sites). In 7 cases, only a selected bibliography was located on this main page, with the full bibliography located at another link. Only a small number of sites (4) had

bibliographies located 2 or more links away from the main page. 9 sites had either full or partial bibliographies located on the main page.

Bibliography Distance	Total	Bibliography Distance	Total
0	4	1 recent, 2 full	1
0 recent, 2 full	1	1 selected	2
0 selected, 1 full	2	1 selected, 2 full	1
0 selected, 1 to lab listing	1	1 selected, 3 full	1
0 very recent, 1 selected	1	2	3
1	43	3	1
1 lab, 2 personal	1	Not applicable	2
1 online, 2 full	1		

**Table 3-2: Distance of Bibliography From Home Page**

The organization of the bibliographies were predominantly chronological (25) or by the type or venue of publication (20). In a few cases (6) the viewer had the option of changing the organization of the bibliography.

Sites also varied in how the researcher made the full text of the paper available. 24 sites included links to the full text of the majority of the papers, while 17 didn't include full text versions at all. The remainder were somewhere in the middle, i.e., a scattering of citations included full text.

Researchers also varied in their practice of including additional materials, such as: links to prior versions of a paper, supporting material including videos or presentation files related to the paper, and multiple versions of the paper. As a general rule, the bibliographic entries displayed a wealth of diversity in what was included in a given entry and how the entry was formatted. This variance could be potentially reflective of either local or discipline practise.

#### ***3.4.4 Navigational Structure***

As noted above, 31 sites had some form of navigational system. In some cases, the navigation structure helped to divide sections in a single page. For others, the

navigational structure spanned the entirety of the site and provided an overall unifying structure. In a small number of cases, the primary navigation system had both primary entries and sub-entries under each main entry.

To get a better idea of the structure, I compiled a list of links in the navigational system. The link names were normalized as much as possible to allow for quantitative analysis. For instance, if the terms “papers” and “publications” lead to the same kind of page (publications), they were both tallied under the same category. The final list of navigational elements and the frequency is in Table 3-3. While I tried to make everything as consistent as possible to allow for this analysis, a large number of elements were distinctive and had to be left as single items.

The number of common elements argues for a common base structure for the majority of the sites. 24 of the 31 sites had an entry for research and projects, 22 for publications, and 19 for teaching. A number of interesting points can be made about this data. 4 sites separated out books as an entry distinct from the overall publication list, indicating that some researchers give a special preference or weighting to books over other publications.

Another interesting result is that the research/project link occurs as the most common navigational element. In comparison, the bibliography and contact information occurs more frequently as content on the sites (see Table 3-1). This can be accounted for in a number of ways. If the research/project content consisted of only a brief mention on the landing page, it was not counted as site content. However, the brief mention often was linked in the navigational system. In other cases, the bibliography was only presented as a set of selected items on the landing page and therefore not included in the navigational system. Finally, only sites with significant content areas typically had navigational systems; this could explain the strong correlation between project pages and navigational systems.

A similar finding occurs with the teaching pages where 18 of 65 sites have teaching content (Table 3-1) but 19 of 31 sites (Table 3-3) list it as a navigational



item. The discrepancy can be accounted for in the same way as the research and project links vs. content. Often, the teaching link in the navigational system points to a small section on the landing page while Table 3-1 only lists sites with substantive content associated with courses or teaching. Indeed the 18 sites where teaching content is identified is not a subset of the 19 sites with teaching links in the navigational system. In some cases substantive teaching content existed on the site but either no navigational system exists or teaching is not included in the navigation. On the other hand, some sites that have links to a teaching section only have a couple of lines on the primary landing page devoted to teaching content.

There is a strong argument that the discrepancies show there is only weak correlation between the amount of content available for an area and identification of that area on the navigational system. This could mean a few things. It could be that the navigational system is patterned on a template that the researcher adds obligatory content to satisfy having a link. It could also mean that the navigational elements are placeholders for future content. Either way, analysis of site content through the navigational system is problematic and would be so for automated harvesting that relied on the navigational structure to classify content.

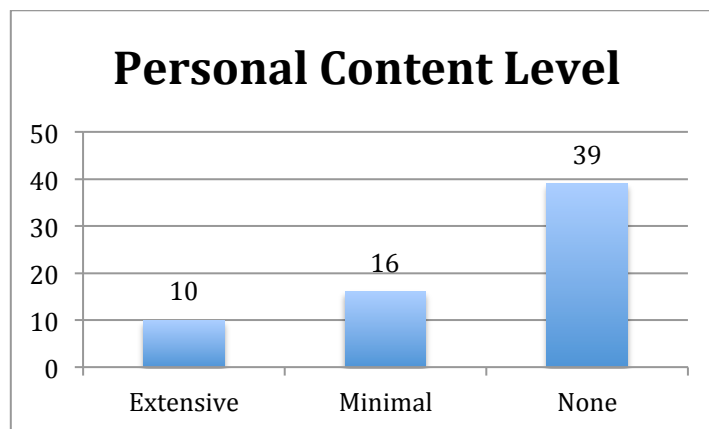
Navigation Element	# of Sites	Navigation Element	# of Sites
Research / Projects	24	Fun	5
Publications	22	Books	4
Teaching	19	Press	3
Home	13	CV	3
Biography	11	Organizations	3
Contact	11	Other	2
Personal	10	Photos	2
Duties & Activities	8	Positions	2
Research Team / Students	8	Additional Information	2
Notes/Presentations/Lectures	7	Research Lab	2
Resources	6		

**Table 3-3: Elements of the Navigational System**

### 3.5 Level of Content

---

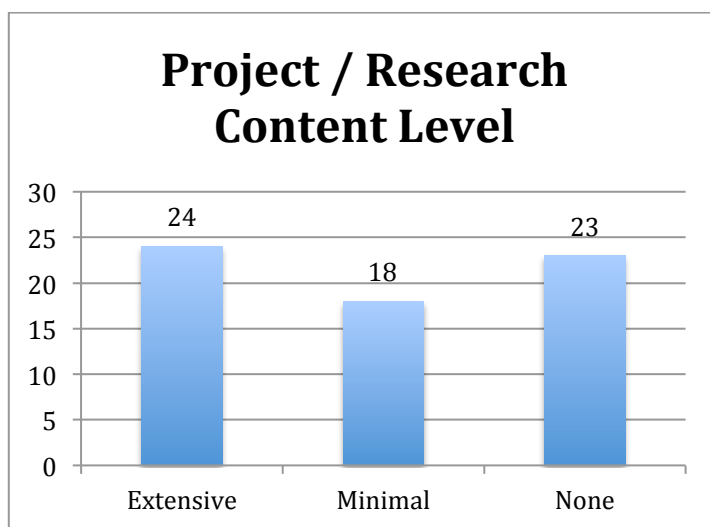
Finally, I quantified the level of content in four areas: *personal content*, *project involvement / research interests*, *teaching and instructional resources* and *external content* related to the researcher. Three levels of classification were chosen based on prior evaluation of the sites and a desire to keep the classification relatively simple. A content level of *none* was given if no content of that type existed or if the content was no more than a sentence or two. A *minimal* level was given to sites where there was content in that category but either as a simple listing or as a short synopsis. An *extensive* level was given to sites that provided extensive content in that category, where either a significant amount of content was provided for at least a subset of a larger list, or for which a large variety of content was provided. The results are described below.



**Figure 3-5: Personal Content Level**

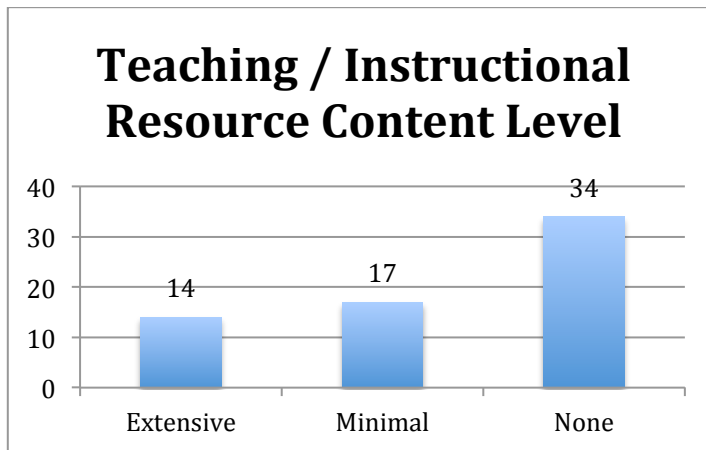
In the case of personal content (Figure 3-5), few researchers (10 of 65 sites) provided *extensive* personal content. For most sites, content was strictly professional in nature. 16 sites provided a *minimal* amount of personal content such as a few family photos or brief mention of external interests or family composition. *None* was the largest category with 39 out of 65 sites. Clearly, the majority researchers are not interested in disclosing personal content.

In a number of cases, the content straddled the line between professional and personal, where the content was of a personal nature but the context was professional. Examples of this include candid photographs taken at professional functions. However, since the content itself communicated something personal, this was considered personal content for the sake of classification.



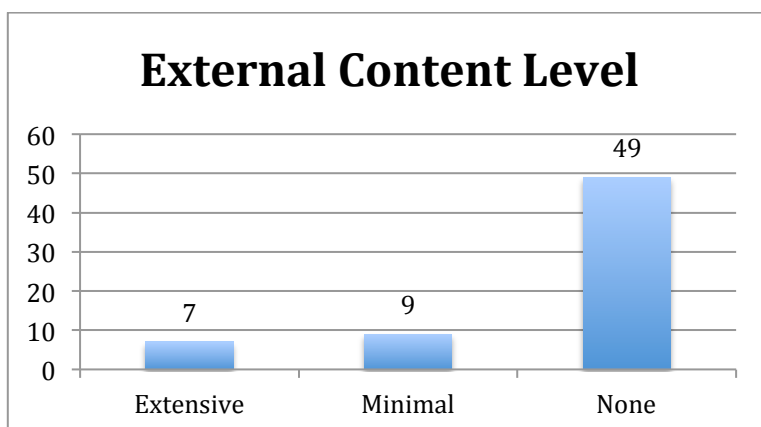
**Figure 3-6: Project / Research Content Level**

The project and research content level described the amount of description and content related to the researcher's area of research interests and projects. In the case of an *extensive* amount of project/research content (24 of 65 sites), researchers would often provide a list of the projects they had worked on. Each project would link to a separate project page with listings of related publications, resources and project team members. It should be noted that in many cases these project pages were not necessarily located within the researcher's pages, but on entirely separate pages or on pages located within the research group pages. However, in most cases it was relatively clear that researchers had participated in the creation of the project pages either directly or through their graduate students. A smaller number of sites (18 of 65 sites) provided *minimal* information about their projects or research interests but the content was generally limited to a page with one to two sentence synopses of a set of research interests or projects.



**Figure 3-7: Teaching / Instructional Resource Content Level**

Despite most researchers being teaching faculty within higher education institutions, the amount of teaching and instructional resource content available on was relatively minimal. Only 14 out of 65 sites provided *extensive* content that included lists of courses, including past courses, with their attendant syllabi and course readings. Some researchers did provide additional information for students above and beyond course notes including suggested reading lists, helpful topical pages and information for graduate students. More common (17 out of 65 sites) were sites that provided *minimal* content; a list of courses taught with a short descriptive paragraph or a syllabus for only the most current courses.



**Figure 3-8: External Content Level**

One final area of interest was how much external content related to the researcher he or she attempted to gather on the site (Figure 3-8). This area was particularly challenging to assess, as even the most extensive collection of external content typically was a list of links to external sources. This is understandable from the point of view that most external content is covered by copyright and therefore, the researcher would be unlikely to have permission to reproduce the content of a site. Therefore the primary distinguisher was the level of care applied to the curation of these lists. As can be seen by the *none* category (49 out of 65 sites), few researchers included related external content. For those providing *minimal* content, this usually consisted of a list of awards won or committees served on. Only 7 out of 65 researchers provided *extensive* content through listing news articles and other online sources where they are either mentioned or discussed extensively. A few included links to videos and audio streams from news organizations where they had been interviewed or their work discussed.

### 3.6 Discussion of Results

---

For those familiar with academia, the results are not particularly surprising. However, moving from generalities to specifics is important. So codifying the website structure and frequencies helps to establish the general nature of personal websites of researchers, including commonalities and differences.

As expected, the bibliography is dominant on the researchers' websites. This would flow from the presumed goal of an academic online presence to make one's publication list available to granting agencies, to potential students or collaborators, and to others in the field (to ease access to publications and to encourage citations). Given the value of the curriculum vitae in the life of an academic and the regularity for which the CV is called for, one would assume a similar importance placed on the online equivalent.

What is surprising is the variety in the presentation of the bibliography. In some websites, the complete bibliography could only be found on a PDF version of

the CV. The navigable, HTML version of the bibliography only contained selected entries—typically the most recent. This suggests that for some researchers, the CV still has primacy. As well, there was considerable inconsistency and variation in how the publications were organized, whether there existed links to online versions of the paper or to locally downloadable copies, and the way the entries were presented. This goes doubly so for the underlying HTML used to display the publication entries. All of this suggests that neither a standardized format nor machine readability are currently a consideration for the way a researcher maintains publication lists.

Some of the variations in what is offered for the full text of the publications can be attributed to a number of factors. It is likely that older publications may have been typeset using analog methods and no digital version exists. For the researcher to put up a large historical publication list might require significant effort to go back, locate copies and scan them. All this requires considerable effort for which a busy researcher is unlikely to have time. There is evidence of this in a number of the bibliographies, where more current publications have a digital full text version, while older entries do not.

Another factor in the presence of full text might be related to the complexity of copyright. Publisher policies on allowing researchers to provide a digital version on their personal websites are varied, with some allowing postings and others disallowing them. Researchers might not wish to incur publisher accusations of copyright infringement by reposting their own work. However, a number of websites appeared to provide either all or a majority of the publications in full text form. This may be due to the fact that ACM, a major publisher of HCI material, allows researchers to self-archive their publications on their own site. For researchers who do the majority of their publishing “within the fold” of computer science, the blanket permission by such a significant organization would certainly embolden one to make all publications available.

Two of the other major content areas are teaching and research. The web has become the primary source of information for students in higher education, and thus students now expect to be able go to the professor's website to get readings, course schedules and so on. This expectation would influence the behaviour of the researcher as teacher. Similarly, the paucity of personal or frivolous content indicates the researchers' intent to convey a professional persona. Thus having research interests and current projects available both reinforces the projection of the researcher persona and provides useful information to prospective students and collaborators. It also provides an easy pointer to a carefully considered explanation of work that five minutes in a chance encounter at a conference could not provide. In contrast to the publication list (which while having much variability still has some comparative level of consistency), neither the teaching content nor the research / project content has a level of consistency for meaningful comparison. This is not surprising since publications reflect a shared social convention that tends to enforce consistency while neither teaching nor research content have an overarching set of social conventions to dictate their presentation. Moreover, the content itself varies significantly between researchers. Given that HCI is a field that spans many kinds of approaches (e.g., pure technology development, ethnographic studies), it is reasonable to suggest the kinds of projects, their overall composition of team members and the kind of outputs created all influence the presentation on the website.

Beyond these core areas, most websites contained an identifying photograph and extensive contact information. Given this, one could easily imagine a scenario where an automated system could gather the photos, the contact information, the research interests and the publications and present visualizations on nearest neighbours in terms of either geography or interest areas. While the researchers themselves are typically senior in their field and would be familiar with people doing related work, it could be useful to other researchers who might not have the familiarity, either because they are new or outside of the HCI community.

Unfortunately while these elements are all in place, the lack of consistency and the way HTML is coded for display mean that easy machine readability is unlikely.

One other possibility of identifying underlying structure for machine readability might be to look at the primary navigation structure. However, the analysis of the navigational structure demonstrated the need to normalize navigation entries. Consider the publication list: some researchers list the publication section as “papers” while others list it as “publications”. Others separate it into two or more sections like “articles” and “books”. Nor is there consistency in the sections themselves. Out of 189 navigational entries identified covering 31 sites, there are 22 entries distinctive enough to be singletons. In order to have automated crawling of the sites to gather content, the navigation terms would need to be mapped to a common vocabulary. Based on this limited set of terms, that mapping would currently require human intervention. On the other hand, a sufficiently large database of navigational elements might begin to yield enough regularities to create mapping rules, e.g., lists of synonyms.

### 3.7 A Typology of Sites

---

While the results of the survey identify a significant amount of variation from site to site, there was enough commonality to suggest a typology of sites. The typology comes from the confluence of the type of institutional affiliation, the design of the website, and the amount of content on the website. While the typology does not emerge directly from the quantification of these attributes, the categorization was increasingly obvious after repeated inspections. I suggest 4 types of sites as detailed below.



### 3.7.1 The Basic Professional Site

## Peter Johnson

Professor of Computing Science

Head of [Department of Computer Science](#) Head of [HCI Group](#)

### Contact

[Department of Computer Science,](#)

*University of Bath*  
BA2 7AY  
UK,  
EU.

Email: [P.Johnson@bath.ac.uk](mailto:P.Johnson@bath.ac.uk)

See also

[Departmental Home Page](#)

[Systems Engineering for Autonomous Systems](#)

## History

**Further Qualification:** [Bachelor of Fluencing \(Unseen\)](#).

Joined University of Bath in 1999 from Department of [Computer Science](#), Queen Mary and Westfield College, University of London (1984 - 1999). Research Fellow at University College London, [Ergonomics](#) Unit (1981-4). PhD in Cognitive Psychology at [Warwick](#) University (1978-81)

## Research

### Current Funded Projects

#### *Current EPSRC research grants*

GR/R40739/01 Human Computer Interaction and Flight Deck Safety. £223,000 Completed June 2005. EPSRC Research Cluster - Creativity in Design for the 21<sup>st</sup> Century - with Dr. H. Johnson £61,000 - STARTED 1/1/05.

EPSRC Engineering Doctorate Centre in Systems Engineering- in collaboration with Loughborough, BAE systems, Leicester, Strathclyde, OUB. Four year funding of Centre --- £3.5m APPROVED for Funding by PANEL 6/5/05

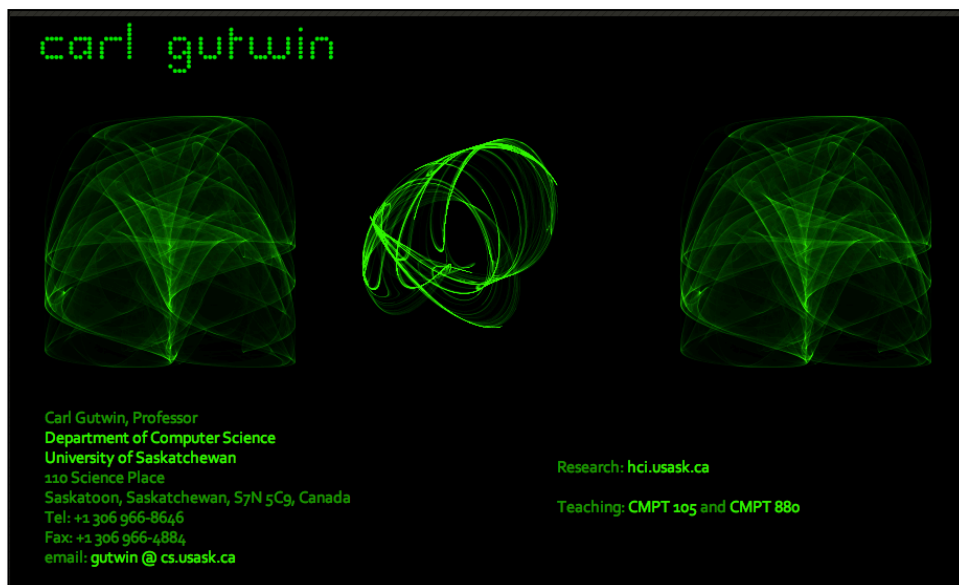
**Figure 3-9: Basic Professional Site (Peter Johnson)**

One example of a basic professional site is the site of Peter Johnson as depicted by the above screenshot (Figure 3-9). The basic professional site shares a great deal of similarity with the traditional CV. Most likely, the layout is either deliberately modeled on the CV, or the choices and motivation for creating a website is similar

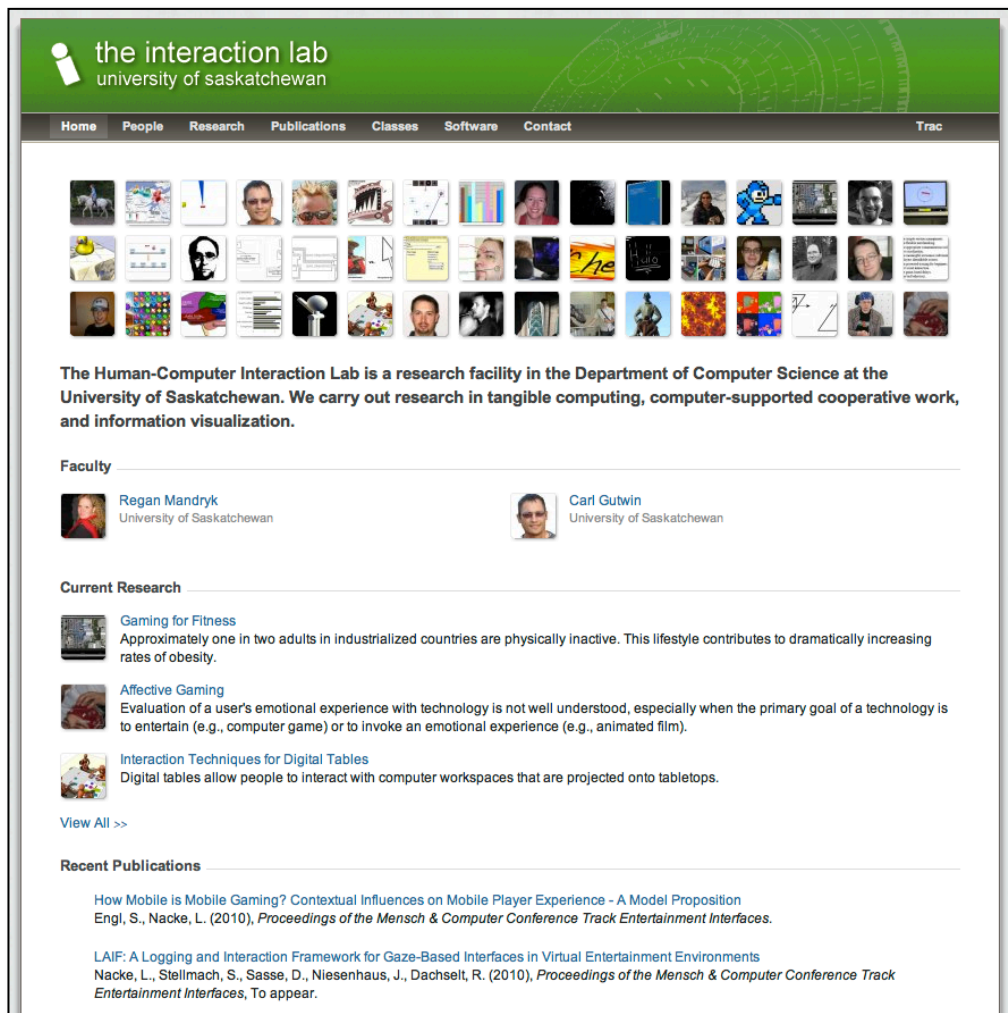
enough to a CV to yield a similar result. Typically, the basic professional site uses a layout involving a single column of text punctuated by appropriate section headings against a white background, i.e., the basic white design described previously. Most basic professional sites have only one page or, at most, supporting pages that supplement sections on the landing page. For instance, the landing page may have a set of selected recent publications with a link to a supporting publications page with more entries. Generally, there are only a couple of supplementary pages and they need to be contextualized by the linking section from the landing page to make sense.

Navigation on the basic professional page tends to be minimalistic. More often than not, navigation is the direct function of the section headings being visually different than the surrounding text. A few sites had a simple table of contents with active links to the section heading. Typically, the primary focus is on the bibliography; indeed, the bibliography is often the only other page in the basic professional site. Project and research interest information tends to be brief synopses. In most cases the biography, if it exists, is similarly brief.

### ***3.7.2 Researcher / Lab Site***



**Figure 3-10: Personal Page (Carl Gutwin)**



**Figure 3-11: Lab Page (Carl Gutwin)**

Carl Gutwin's site represents an archetypal example of the researcher / lab site. In the researcher / lab site, the researcher's personal page (Figure 3-10) is often relatively simple and contains minimal information, e.g., a biography and contact information. It is the associated lab site (Figure 3-11) that contains the bulk of the content. Gutwin's page is unusual in that the personal page is particularly distinctive as compared to the lab site, whereas in most other researcher / lab sites, the researcher page often utilizes the same design template as the broader lab site. Since the lab pages have more active content like ongoing project information and the publications, they tend to be updated more often than the researcher's personal page.

One particular feature of the lab site is that publications are usually contained in an aggregate bibliography. Typically the bibliography is driven using a database or content management system. When presented, the list of publication often intermingles the work of all of the researchers in a chronological ordering. In many cases the publications also appear in project pages in a way that suggests everything is being drawn from a single database. If the researcher has a separate bibliography, it is more often than not a static document in CV form. Otherwise, filtering for the researcher's work requires using the search function if it is available. This raises a number of interesting questions that are unclear from surveying the sites alone.

The first question is how do researchers view themselves in the context of the broader picture of the lab? It is clear from the literature and the survey that researchers are largely defined by their publication list online. For the senior researcher in a lab, it is likely a reasonable assumption that he/she views the entirety of the lab's output as reflective on him/herself. But what of other researchers in the lab?

The second question is one of workflow. Who is responsible for creating the entries in the database? Is each researcher responsible for his or her publications or is there a coordinated workflow that drives the updating of the content? Marshall (2008) raises the question of ownership of the discrete parts of a project and this particular template makes that issue very clear.

While those questions cannot be answered from surveying the site alone, the usage of a database / content management system is hopeful. It opens the possibility of being able to harvest the content of the researchers' sites with automated tools. As well, it indicates researchers' willingness to create metadata, even if it is for their purposes.

### 3.7.3 Extensive Site

**Alan Dix**  
Professor  
Computing Department  
Lancaster University  
Lancaster, LA1 4WA, UK  
[a1d@lancaster.ac.uk](mailto:a1d@lancaster.ac.uk)

send a virtual cracker!  
now on [facebook](#)

too

scramble reset  
puzzle help and info

Professor Alan's puzzle square:  
can you solve it? (click arrows)  
put a puzzle like this [on your own page](#)  
or [make one](#) with your own picture

1571681340  
about life/counter

Journal: [flonphort](#) and [nails on Good Friday](#)  
words: [the carpenter](#), [more](#)

Alan elsewhere: my [calendar](#) || my [blog](#) || see my [SnipIt Channels](#) || on [Twitter](#) || on [slu.ma](#) (RDF search) || on [FaceBook](#)  
Publications: ag. eng.: 1982-1986, computing: 1985-1989, 1990-1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010.  
Other places and pages: [things I've done](#) - a mini-CV || [blog](#) || [my research topics](#) || HCI education || [eBusiness Bulletin](#) || [Magisoft Wand](#) || [meandeviation.com](#)  
Notes for lectures etc.: [CSC252 - Human-Computer Interaction](#) || [CSC355 - Artificial Intelligence](#) || [MSc/MRes project suggestions \(will update soon\)](#) || [MRes](#) || [CSC221 - Software Engineering](#) || [CSC224 - Interactive Systems Engineering](#) || [MSc HCI / MRes AISD](#) || [Research and Innovation Techniques](#) || [tutorial and short course notes](#) || [Rome 2003](#)  
Serious and fun things to do: [Query-by-Browsing on the Web](#) || [Professor Alan's puzzle square](#)  
Personal: [words](#) - just me writing, [essays](#) - tangential research: cognition, imagination, etc.

Firefly in Lancaster City  
Centre  
Light fantastic! live and  
flashing at CityLab in  
Lancaster - go see them  
soon  
... [more on Firefly](#)  
... and [more on the CityLab display](#) ...

VieMaster launches video  
([www.vismaster.eu](http://www.vismaster.eu))

Flash

study HCI @ lancaster  
look at our [MSc in Human Computer  
Interaction](#) (formerly MRes design and  
evaluation of advanced interactive  
systems) or [MSc in Advanced Computer  
Science](#)  
or mail me if you are interested in PhD  
studies

Events and Calls ...

desire International Summer  
School: "Models of Creative  
Design" For innovation in science and  
technology  
Aveiro University, Portugal, 19-25  
September 2010  
application deadline 3rd May  
... [more](#) ...

**HUMAN-COMPUTER INTERACTION**  
third edition  
ALAN DIX - JANET FINLAY - GREGORY ABOUD - RUSSELL BEALE

alan@Large ...  
Invited Talk: [Touching Technology: taking the physical world seriously in  
digital design](#), (full text notes also available) [Greek SIGCHI Workshop "Human-  
Computer Interaction: Theory and Practice of design of usable and accessible  
technologies"](#), Athens, 5th March 2009.  
India Visit: part of the EPSRC funded UK-India Network on [Interactive Technologies  
for the End-User](#), Bangalore, 2-5 Feb 2008.  
Invited Tutorials: [Winter school on Usability Engineering](#), Anzere, Switzerland, 28-  
30 Jan 2009.  
SIGCHI Ireland Inaugural Lecture: [Human-Computer Interaction in the early 21st  
century: a stable discipline, a nascent science, and the growth of the long tail](#),  
(full text now available) Dublin, 2nd Dec. 2008.  
distinguished lecture series: [Human-Computer Interaction: as it was, as it is,  
and as it may be](#), St Andrews, Scotland, 6th Nov. 2008.  
keynote: [Using the Web of Data at WOD-PD 2008](#), Web of Data Practitioners Days,  
Vienna, Austria, 22-23 Oct. 2008.  
Invited tutorial: [Interaction with and through the mobile](#) at [MobiKUI 2008](#) - First  
International Workshop on Mobile and Kinetic User Interfaces, Fribourg, Switzerland,  
13-14 Oct. 2008.  
keynote: [Tasks = data + action + context: automated task assistance through data-  
oriented analysis](#), at [Engineering Interactive Systems 2008](#), Pisa, Italy, 25-26 Sept.  
2008.  
keynote: [As We May Code - The art \(and craft\) of computer programming in the 21st  
century](#), keynote at PPIG08 at The 20th Annual Psychology of Programming Interest  
Group Conference, Lancaster University, UK, 10th/12 September 2008..  
masterclass: [From Formalism to Physicality](#) part of [UPC North](#): Understanding People  
and Computers, UCLanc, Preston, 30th April 2008.  
Invited talk: [Designing for adoption and designing for appropriation](#), University of  
Technology of Berlin, 12th Feb 2008  
panel keynote: at [Social Technologies Summit](#) part of FutureSonic2007, Manchester,  
May 2007.  
keynote: [the brain and the web - intelligent interactions from the desktop to the world](#), at  
Simpósio de Fatores Humanos em Sistemas Computacionais (IHC 2006), Natal Brazil,  
19-22 Nov. 2006.  
I'll be at [Misuse and Abuse of Interactive Technologies](#) and [Designing for Collective  
Remembering](#) at CHI 2006, Montreal, 22nd and 23 April 2006

Alan on Twitter  
wondering why this never made the news? <http://bit.ly/cS5F4d>  
#takeitback about 2 hours ago  
added skype support to SnipIt! 1 day ago  
References and HR stuff all day - time for late (4pm!) lunch -  
hungry 2 days ago  
[follow me on Twitter](#)

Alan's Blog  
and they said they would protect front line services  
Friday, May 14, 2010  
Just been at a public meeting about imminent cuts in the school  
here on Tires. In a small school like this (120 pupils) losing  
several posts isn't just a matter of shrinking slightly, but means  
that whole subjects, such as French, drop off the curriculum.  
There are two issues here. One is for the... [read more](#)...

update: (im)migration Holyrood vs Westminster  
Sunday, May 9, 2010  
Since post last week on migration Holyrood vs Westminster ,  
found link on the BBC website to the the BBC News "Reality  
Check" on immigration that showed net outflow of non-EU. That  
is migration is out of the country not in! Also Mark Easton's blog  
@ the BBC, which gives more info. Bottom ... [read more](#)...

language, dreams and the Jabberwocky circuit  
Thursday, May 6, 2010  
If life is always a learning opportunity, then so are dreams. Last  
night I both learnt something new about language and cognition,  
and also developed a new trick for creativity! In the dream in  
question I was in a meeting. I know, a sad topic for a dream, and  
perhaps even sadder it had started with... [read more](#)...

migration Holyrood vs Westminster

Figure 3-12: Extensive Site (Alan Dix)

The extensive site represents the stereotypical image of a website, particularly from researchers with a computer science background. Typically, the design of these extensive sites is idiosyncratic. They vary in layout (as in the case of Alan Dix's site, Figure 3-12), overall visual look, and type of content. They tend to be large, multipage sites. They have a navigational system repeated on each page in the site. They usually have rich bibliographies, and are most likely to include the full text stored locally on the site.

In addition to comprehensive bibliographies, extensive sites often contain the most personal content in the form of blogs, fun or interest pages. As such, they

provide the most contextual information about the researcher and represent a valuable resource to archive. Given the variety of site layout and organization, the challenging aspect of this would be that machine reading of sites to extract information (like publication information or projects and research interest) in all likelihood would need to be customized for the site.

What is most surprising is that the extensive site is not more prevalent. We would expect computer science researchers, give their technical knowledge, would want an extensive site because it would demonstrate their mastery over the medium as part of the projection of the expert persona. However, it is clear that the extensive site requires a considerable amount of time and effort to create it, which may dissuade researchers who otherwise might want such a site.

### 3.7.4 Organization Template

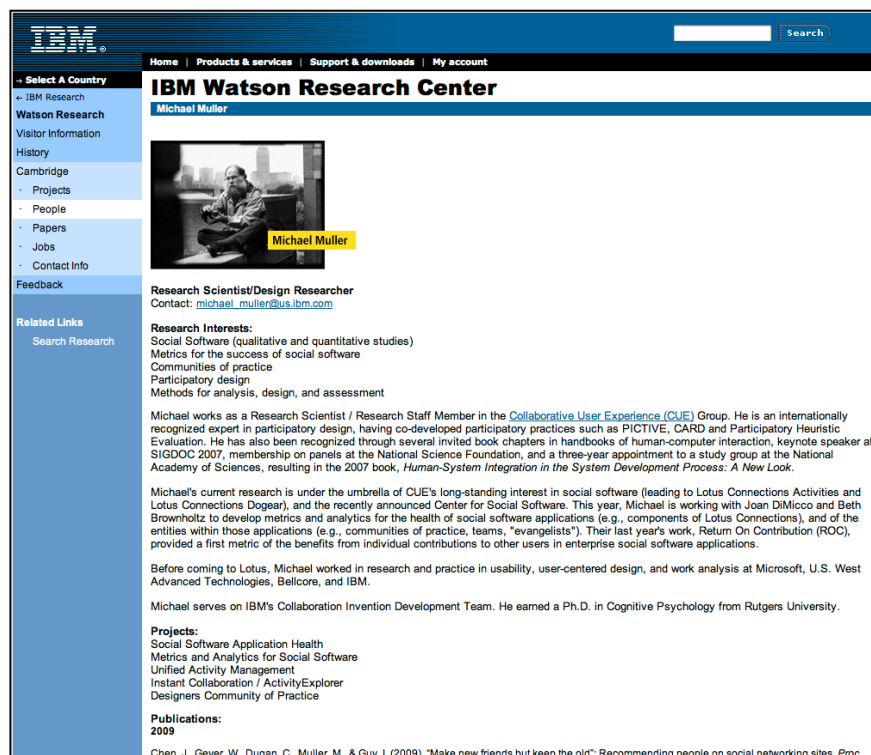


Figure 3-13: Organization Template Site (Michael Muller)

The final type of site is the organization template. As with the example above of Michael Muller's page at IBM (Figure 3-13), the organization template is driven by the branding of the organization. It often represents the website of a corporate researcher. However, the rigidity of adherence to the template varies from organization to organization. Thus some corporate researchers do have the freedom to stray from the norm. Most researchers will likely use the template for the sake of consistency, to display organizational affiliation, and because the tools are already in place to use the template. The researcher is also likely constrained by an organization-wide content management system that limits the variability and flexibility of content. As well, corporate researchers may be driven less by the need for external exposure.

Content on these sites are typically limited to a short biography and a list of recent publications. With the focus on recency, bibliographic organization tends to be chronological in nature. Personal content is minimal, if any exists at all. Project information usually consists of a list of current projects that may or may not link to richer project pages. As an interesting side note, this type of page also tends have the least contact information, usually little more than an e-mail address.

### 3.8 Conclusion

---

This chapter explored the results of a survey on personal websites of senior researchers in the field of human-computer interaction. The general composition and content on the websites hold few surprises. They are professional sites predominantly providing information on the researcher's publications, teaching and research / project interests, as well as contact information. Many of the sites are fairly simple, which is surprising given the technical nature of the group being surveyed.

4 types of sites were identified:

1. The Basic Professional Site
2. The Researcher / Lab Site
3. The Extensive Site
4. The Organization Template Site

In the next chapter, these types will guide the interview process. I talk directly with selected researchers to garner a better sense of the choices they made in content selection, presentation, organization, and management.



# Chapter Four: Interviews

---

## 4.1 Chapter Synopsis

---

In the previous chapter, I discussed the results of the survey I conducted on the personal websites of HCI researchers. Among the findings, I identified a number of elements that the websites shared in common as well as a set of standard website types. In this chapter, beginning in Section 4.3, I will discuss efforts to understand the rationale and approach to the creation of the websites by interviewing individual researchers. The interview was largely (but not completely) guided by high-level outcomes suggested by the survey from the prior chapter as summarized and discussed in Section 4.2.

## 4.2 Introductory Discussion

---

In the previous chapter, I evaluated a number of researcher personal websites looking for commonalities and differences between the websites to identify trends. To reiterate, the key findings are as follows:

1. Most websites were academic websites.
2. Most sites generally were content rich.
3. The design of most sites were either personal in nature or of a basic white design.
4. The most common elements contained in the websites were the publications, contact information, an identifier photograph and a list of research projects or interest.
5. The bibliography or publications page was typically quite close to the front of the website.
6. The amount of personal content on the websites was minimal.

As the results of the survey provide the basis for the next step in understanding the websites, it is important to place the findings of the survey in the context of what the prominent elements of the website reveal about the researchers.

#### ***4.2.1 What the Survey Tells Us About the Sites***

##### *4.2.1.1 The Dominance of the Bibliography*

The primacy of bibliography is significant for understanding the intent of the sites. In the majority of sites, it is either at or near the front page of the site, indicating its general importance. A large percentage of the sites have the bibliography as the best-developed and fullest source of content. As the best maintained section, it is also one of the best indicators to the currentness of the site.

The investment into the bibliography compared to other aspects of the websites is indicative of a number of possibilities. It suggests the websites could be tools for professional networking. It may also be a social convention within the academic community. A third possibility is that it is simply required by organizations the researcher deals with. Since researchers in corporations or governments do not have as complete a bibliography as those in academic institutions, this tends to give more credence to the latter two possibilities.

##### *4.2.1.2 Personal Disclosure Variety*

In contrast to the consistency of the bibliography, there is great variation in the kind and amount of personal disclosure on the websites. Many of the websites have minimal or no personal disclosure. A small number have personal information of a non-professional nature (largely being photographs of family members and a few with personal blogs and other personal information). However there is very little pattern to the context and the construction of content of personal nature. This suggests that unlike the bibliographies, there is less social convention or institutional requirement or control over personal content. The general infrequency of personal content suggests that the primary intent of the website is professional in nature.

#### *4.2.1.3 Website Design*

A large percentage of the websites use a basic white or a template design approach. Following with the theme in the previous two discussion points, the prevalence of this design template supports the ideas that either there are social conventions or institutional requirements in play. Even in the case of the more personalized websites, the basic design elements seemed to resemble those of the basic white design.

However, one argument against institutional requirements is the relative lack of institutional markings and colours that would demonstrate a required template approach. This suggests that the general impetus and control of the design, structure and content of personal researcher websites is dictated by social convention within the researcher community as opposed to direct institutional control. However, another possibility for the prevalence of the basic white template could be the ease of creating and maintaining a simple to code site. This raises a question of how much of the content is dictated by the effort (or lack thereof) to create and add content to the website.

One final possibility in terms of the choice of design and layout might be that the design is intended to support both human and machine readability. However, the lack of consistency in the formatting of specific pieces of content like the publications or projects seems to argue against this as a motivation. On the other hand, many of these sites were developed before many web design best practices were formulated and it is possible the choices represent a best guess as to how to make a site machine-readable.

#### ***4.2.2 What the Survey Does Not Tell Us About the Sites***

The possibility of machine readability as design choice raises questions as to the intent of the content's audience and this clearly is one of the limitations of the survey methodology. While it is possible to make inferences about the intent and the choices underlying the design and content, it is impossible to determine whether the

choices made reflect the limitations of technology, time or knowledge or the result of conscious decisions.

This lack of understanding of the goals and aims of the creators limit our ability to provide useful tools that would encourage the preservation of their content. For instance, it would be good to know the reasoning behind the lack of consistency and completeness in the project pages. Unlike the bibliography, project pages are far less likely to be included in the website and when they are included there is little in the way of a common pattern. As well, the project pages often are out of date or only reflect the most current projects.

Yet there is a strong argument that project pages may be more valuable from a preservation perspective than the bibliography pages. The content emanating from the bibliography pages are often preserved elsewhere whereas project pages may contain content that does not reside anywhere else. If there were some way to assist researchers in maintaining these project pages and give them some consistency, it could make the job of preserving the contextual aspect of the researcher's work easier. But do researchers want the project pages maintained? Understanding the context and motivation in the choice of content, how it is structured and the reason for the choices made in presentation and management are all-important for creating tools to support preservation. And so the logical extension of this is that we need to ask the researcher directly.

### 4.3 Methodology

---

The goal of this part of the study is to interview researchers to find out their motivations for the creation of the website and for the choices made in terms of inclusion of content as well as for the presentation and management of the content. There are two ways to do this: either by a questionnaire or by interview. Other methods like focus groups would be impractical given the geographic dispersion of the subjects.

In Kaye et al.'s study (2006), they identified 5 values of a personal archive. If the personal website does resemble a personal archive, then the values guiding the creation of the website should be similar. With that in mind, a set of questions was constructed that explored the five values with an initial mix of open-ended and closed questions. The majority of the questions were made open-ended after a pilot study to facilitate gathering more information and allow for deeper exploration. A copy of the questions used can be found in appendix C.

While open-ended questions can be administered via interview or questionnaire, the selection came down to the kind of information sought. As noted, there has been little work done in this area and it was important to solicit as much information as possible to understand exactly why the researchers are doing what they are doing. Given this, the best choice was the interview, as it would allow the interviewer to probe for further explanation or clarification should an answer be unclear or incomplete. Moreover, the initial pilot identified that significant information could be obtained without biasing the interview if the subject were allowed to freely discuss their website without structure or prompting.

The choice of subjects came directly from the survey's list of researchers. Although the initial goal was 10-15 interviews, logistics (such as scheduling constraints) ultimately allowed for 9 interviews to be completed. Researchers chosen were prioritized based on the site type they used with the goal of having at least one for each type.

All of the interviews were conducted via Skype for at least a portion if not the entire interview. Interviews were generally an hour in length and interactive. Initially the subject was asked to reflect on the choices and motivations for maintaining a website and the methods used in the maintenance. Each subject was encouraged to be as broad and free ranging as they liked as long as the focus had some relevance to the website. Once the initial comments were done, the subject was then asked the questions in Appendix C. Questions were omitted if the subject had already answered it as part of another question or as part of the initial

conversation. At times, questions would be modified to probe interesting or unusual answers. Answers were occasionally repeated to the subject in a synoptic or clarified form if the interviewer deemed the answer unclear in the context of the question.

Most interviews were recorded with two devices (computer and digital camera) to ensure that the answers were accurately captured. The interviewee was asked to share their Internet browser window with the interviewer and bring up appropriate pages on their site when discussing them. The recordings captured both the audio of the conversation and the video of the shared screen. In addition, an interview document containing the questions was created for each interview and notes and responses were recorded on the spot where answers were relevant to specific questions. Once the interviews were completed, the question response sheets and the interviews were analyzed to identify major themes and draw out pertinent quotes. Unlike the site surveys, the interviews were not formally coded into categories. Instead, an informal process was used where themes were transcribed from the interviews as they were reviewed. The list of themes was then reviewed to identify the dominant themes that recurred most consistently. These dominant themes form the basis for the results in the next section.

## 4.4 Synopsis of Results

---

As with the results of site survey, there are few surprises. Indeed, in most cases the answers to the common questions could be anticipated given that the domain should be familiar to most academics. However, there are some interesting insights as to the choices and in particular to where the values of the website differed from the findings identified by Kaye et al (2006).

### ***4.4.1 Strong Desire for Control***

There is an almost universal desire to maintain strong control over the website. Most respondents maintained the website themselves. A number of reasons were cited for choosing to update on their own. Most often, respondents pointed out that

assistants did not always do the updates to specifications or the amount of instruction meant it was just easier to do it themselves. Some felt that the site was personal and not an imposition they felt comfortable with.

This desire for control extended primarily to the content of the site itself and in particular to the publication list. The same care was not applied to the visual design of the website. Half of the respondents indicated that someone else (most often a student) had designed the website. In one case, the design of the site was borrowed from a free template available on the web.

While most of the respondents are happy to host their sites on institutional servers, most still view the sites as "theirs" in the sense that they have direct control over the site. In a majority of the cases, they rely on institutional backup systems to provide insurance against loss for the site.

#### ***4.4.2 Little Institutional Pressure to Conform***

While a number of respondents indicated that one of the primary motivators to maintaining their website is to support the annual reporting process and thus is indirectly mandated, institutional templates are rarely utilized in the construction of the site itself. When asked, none of the respondents felt they had been under any pressure to conform to institutional guidelines and even if there had been pressure to use the templates, none would agree to do so. This corresponds with the repeated assertion that "the site is mine". As one respondent notes: "[professors] have to create their personal brand to be successful".

Only in the formatting of the publication page is there any indication of pressure to conform to a standard. A number of respondents noted that grant and reporting requirements drove the specific formatting of the publication page. While this may be the case, respondents have a variety of approaches as to what to include and how to present their publications.

#### ***4.4.3 Trust In Institutional Systems***

Most respondents are happy to have their sites hosted on institutional systems, particularly those in university environments. As one respondent notes, commercial entities may come and go but generally universities remain. One reason that there is not a strong concern for the potential failure of the servers or loss of data is that none of respondents could recall a catastrophic loss to the contents of the websites. And even those who reported the loss of a file stated that it was relatively easy to retrieve the file from the backups.

Only one respondent had his website hosted on a commercial service. However this case is not because of lack of confidence in the university services but rather that due to a leave, he was unable to access the servers. The solution was to host his site on a commercial service so he could access it. Similarly, only one respondent expressed a lack of confidence in the backup services provided by the institution. However, the general confidence in the institutional systems may be borne more of optimistic expectations of the institutions' handling of the content than concrete knowledge of operational policies. When asked, none knew the institutional retention policies for digital content.

#### ***4.4.4 Limited Long Term View***

The general consensus of the respondents is that the website provides an immediate rather than a long-term benefit. Respondents generally perceived the activity in the present. One common use of the site was point someone to more detailed information instead of just giving a quick answer. Other reasons given for the utility of the website include recruitment of graduate students, attracting funding, generating interest in their research and even attracting outside consulting work.

Most, when asked about longer-term prospects for the site, expressed some doubt that the site would be around that long. The sense of most respondents is looking forward to the next thing rather than looking back. This is particularly so for project pages. Most project pages are constructed toward the middle to end of a project rather than near the beginning. As such, virtually every respondent that had



a set of project pages expressed some disappointment at the state of the pages as often being incomplete and out of date. Usually the reason cited for the state of the project page was that the pages are student maintained and once the student is done, work stops on that page.

Another area where the content and the motivations reflect short-term needs rather than long term goals is in the inclusion of research data. There has been a strong effort at the institutional and agency level to get research data available for meta-analysis and review (for instance, the DataOne project in the US). For the most part, respondents had little awareness of this. Most had not considered adding data to their site. Except in cases where the research ethics prohibit it, most were quite happy to supply the data but typically made it available by e-mail rather than including it on their site.

Finally some express chagrin at how out of date some sections of their sites are. In this context, respondents are more likely to want to delete the section or content rather than to spend the time updating the content. This is generally the case in terms of teaching sections. The practise in the teaching sections is to replace content with the most current version. Only a few maintain historical teaching pages. Similarly, in the few instances of pre-publication, the publications are replaced afterwards.

#### ***4.4.5 The Website is a Public Face, Not an Archival Point***

In going back to Kaye et al.'s (2006) values of a personal archive, there are some similarities between the personal website and the office space. But there are also striking differences that can be attributed to the source of the content in each venue. As Kaye et al. (2006) noted, the contents of an academic's office are as much about the stuff he or she collects as the stuff he or she produces. On the other hand, the contents of a personal website are almost entirely about the stuff a researcher produces. This results in the website being more curated, with material meeting a certain level of completeness before it appears on the website. Publications typically will have appeared on the publication venue or at least have been accepted for

publication before the researcher places them on the website. This contrasts sharply with the office and personal digital resources where the contents and associated material with a publication will have lived for a long time prior to publication. This change means that while the personal website shares some of the values of a personal archive, it is generally not viewed as an archive by the researcher. Rather, as mentioned by a number of respondents, the website is part of the researcher's public face.

If we compare against Kaye et al.'s (2006) archive values, the personal website is rarely used for finding stuff again nor is it used as a hedge against fears of loss. With the exception of the contextual information, publications can be retrieved elsewhere. Of the other content, the majority are either time limited (teaching material) or incomplete (project content). In terms of finding it later, most respondents do not use the website to store links to other content and if they have links on their site, the purpose is usually to share sites of relevance in the research domain.

In terms of sharing with others, the majority of respondents do not share content produced by others. When asked, respondents usually indicate that they use other methods to share non-publication content whether via e-mail or file sharing services. So while the website is used for sharing, it is largely of their publications.

In terms of creating a legacy, personal websites exhibit much more in common with a personal archive. Often publication pages conform to a format rigidly adhered to. As one respondent notes, while creating a visual representation for each entry was a pain, it helped to distinguish each entry and so served a valuable purpose. Moreover some respondents indicate that as soon as something is ready, it immediately goes on the publication page.

Finally, personal websites share much in common with physical archives in the value of constructing identity. As one respondent notes, the website is one of the core ways that a professor can build his or her brand and gain attention. Similarly,


one respondent insists his students build their websites. In his view, it is the professional responsibility of being an academic. There is less evidence of the tokenism found by Kaye et al. (2006) in the context of the office space however. This is understandable as there is less sentimental value associated with digital objects. One respondent did point out a game running on his site that he had written years ago. While it was not relevant to his research, it clearly held sentimental value in a way that suggested tokenism.

In summary, based on the interviews I found that personal websites are most closely associated with the values of identity construction and building a legacy. There is some sharing with others but only in terms of the publications and tools that the respondent creates. It should be noted that some respondents used research group websites that served the function of sharing content instead of their own sites. The personal websites did not demonstrate evidence of fear of loss as a motivator nor did respondents utilize websites to aid in finding things later.

## 4.5 Case Studies

---

The previous section identifies broad themes in the responses of the subjects to the interview. In the following section, I will highlight a number of case studies and the researcher's views and approaches to the construction of the personal website. Before I discuss the case studies, it is useful to highlight one exception in the websites surveyed. The general pattern of the websites is that the researcher maintains strong control over the publications and the website itself does not have indicators of a concern for archiving the work in an orderly fashion. However, Ben Shneiderman's website (Figure 4-1) is an exception to this rule.



## Ben Shneiderman




FIG (70 K)  
[Other pictures of Ben Shneiderman](#)

**Email:** [ben@cs.umd.edu](mailto:ben@cs.umd.edu)

**Current Position:** Professor, [CS, UMIACS](#); Founding Director [HCIL](#) (1983-2000)  
 Affiliate Professor: [Institute for Systems Research](#)  
 Affiliate Professor: College of Information Studies – [Maryland's iSchool](#)

**Academic Degree:** Ph.D., SUNY at Stony Brook, 1973.

**Research Interests:** Human-computer interaction, user interface design, information visualization.

Ben Shneiderman receives honorary doctorate on [February 9, 2010](#)

Special Issue of [International Journal of Human-Computer Interaction](#) in honor of Ben Shneiderman's 60th birthday: [Press Release](#)  
[University of Maryland Press Release](#)

Ben Shneiderman elected to the [National Academy of Engineering](#) on [February 17, 2010](#)


[University of Maryland Libraries: Papers of Dr. Ben Shneiderman](#)  
 Includes database of links to academic publications in published and technical report versions.  
[Slide Presentations](#)  
[Video Presentations](#)

Follow me on Twitter: [benbende](#)


A. V. Williams Building, Department of Computer Science  
 University of Maryland, College Park, MD 20742  
 Phone: (301) 405-2680 Fax: (301) 405-6707

Ben Shneiderman is a Professor in the [Department of Computer Science](#), Founding Director (1983-2000) of the [Human-Computer Interaction Laboratory](#), and Member of the [Institute for Advanced Computer Studies](#) at the University of Maryland at College Park ([full resume](#)). He has taught previously at the State University of New York and at Indiana University. He was made a [Fellow of the ACM](#) in 1997, elected a Fellow of the American Association for the Advancement of Science in 2001, and received the ACM CHI (Computer Human Interaction) [Lifetime Achievement Award](#) in 2001. He was elected to the [National Academy of Engineering](#) in 2010: "For research, software development, and scholarly texts concerning human-computer interaction and information visualization." He was the Co-Chair of the [ACM Policy 98 Conference](#), May 1998 and is the Founding Chair of the [ACM Conference on Universal Usability](#), November 16-17, 2000. Ben Shneiderman's interest in creativity support tools led to organizing the June 2005 [NSF workshop](#) and to chairing the June 2007 [Conference on Creativity & Cognition](#).


Figure 4-1: Ben Shneiderman's Website

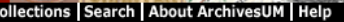


[How do I...?](#) [Site Index A-Z](#) [Search](#)




[Home](#) [Catalog](#) [Research Port](#) [Ask us!](#)





[Ben Shneiderman](#)



**Contact Information**

Dr. Ben Shneiderman  
 A. V. Williams Building  
 Department of Computer Science  
 University of Maryland  
 College Park, MD 20742  
 Tel: (301) 405-2680  
 Fax: (301) 405-6707  
 Email: [ben@cs.umd.edu](mailto:ben@cs.umd.edu)

**Papers of Ben Shneiderman**

- [Archives/Papers](#) (Finding aid to the personal papers of Ben Shneiderman, includes biographical sketch)
- [Career Review](#)
- [Curriculum Vitae](#)
- [Lectures](#)
- [Service](#)
- [Teaching and Advising](#)

**Publications**

- [Books and Chapters \(authored & edited\)](#)
- [Publications \(conferences, refereed\)](#)
- [Publications \(journals, refereed\)](#)
- [Publications \(journals, unrefereed\)](#)
- [Publications \(by author\)](#)
- [Publications \(by first author\)](#)
- [Publications \(by second author\)](#)
- [Publications \(by year\)](#)
- [Other media](#)
- [Tech Reports, unpublished](#)

University Libraries, University of Maryland, College Park, MD 20742-7011 (301) 405-0800  
 Please send comments and suggestions to [ArchivesUM](#)  
 Last edited

© 2005 University of Maryland Libraries/MITH (Maryland Institute for Technology in the Humanities)

Figure 4-2: Ben Shneiderman Archival Page

Highlighted on the Shneiderman's website is a link to the University of Maryland Libraries site where Shneiderman has collaborated with the U of M Archives (Figure 4-2) to organize and catalogue his works. This is the only case in the entire survey where I have identified a relationship where the researcher has taken an especial effort to preserve his work for posterity. While it is beyond the scope of this study to discuss this unique relationship, a few points can be made about Shneiderman's views on the archiving of his work.

Shneiderman notes that the archives were more than happy to accept the papers but were not as well equipped to handle the born digital material. This is important: Shneiderman's experience seems to be echoed by other experiences (Foster et al., 2007) and could be indicative of the broader interaction between archives and researchers.

For the publications in the archival page, there usually is a link to the final publication location as well as to the technical report version. The goal is to have people cite the final publication version but one source of frustration is that often the technical version gets cited. As well, there is relatively little traffic that comes from the archive page to download reports and publications. More traffic comes directly from either Shneiderman's personal site or the lab site; it could well be that Shneiderman's personal site is better known than the archive site. This argues strongly for the idea of researcher brand.

One final note should be made about Shneiderman's experience. In order to keep everything up to date, the publications need to be updated in 4 locations: the CV (as a word processing document), the website, the technical report database and the archives database. As he notes, there needs to be a "better structure of things" in order to keep everything synchronised.

### 4.5.1 The Individual/Lab Website: Saul Greenberg

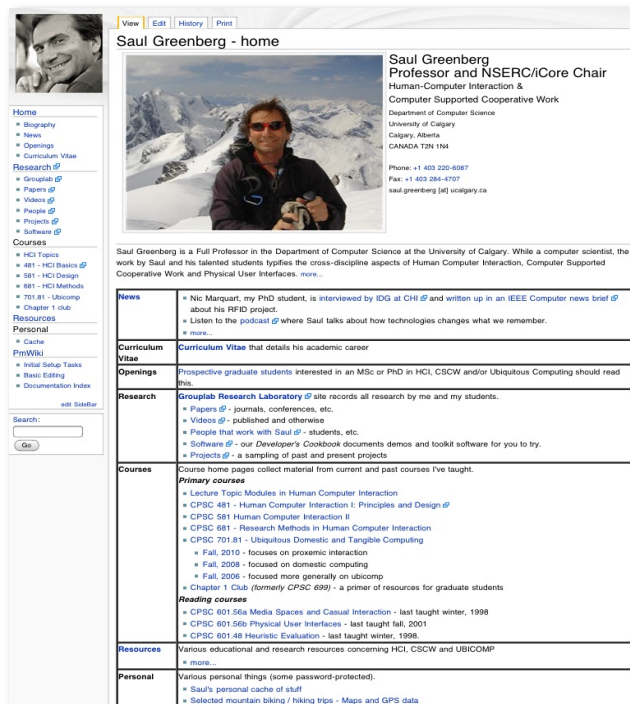


Figure 4-3: Saul Greenberg's Website

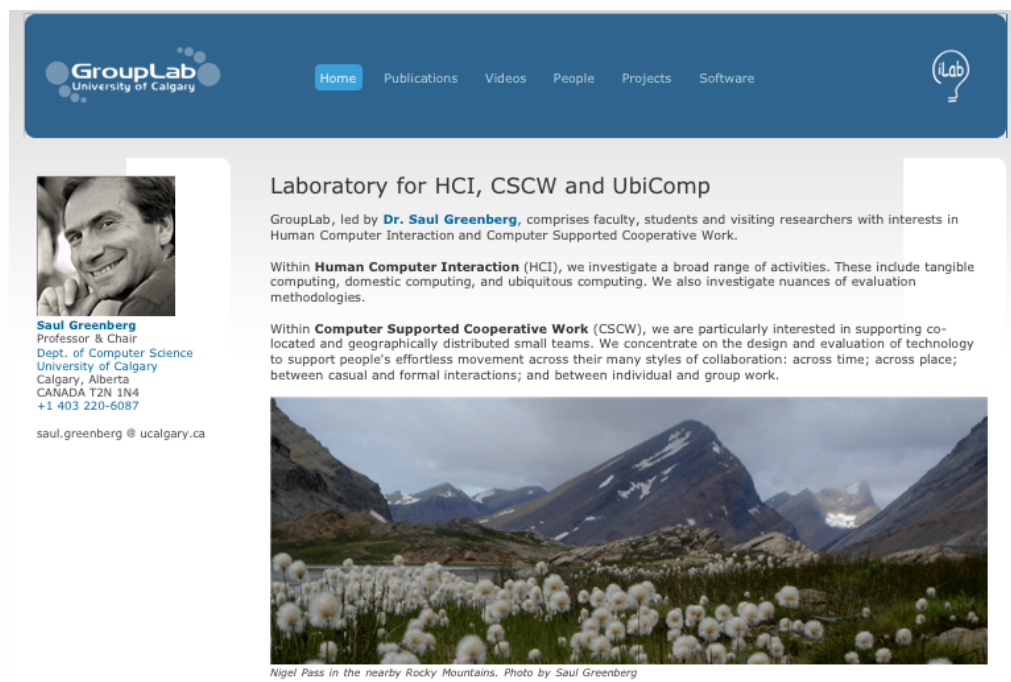
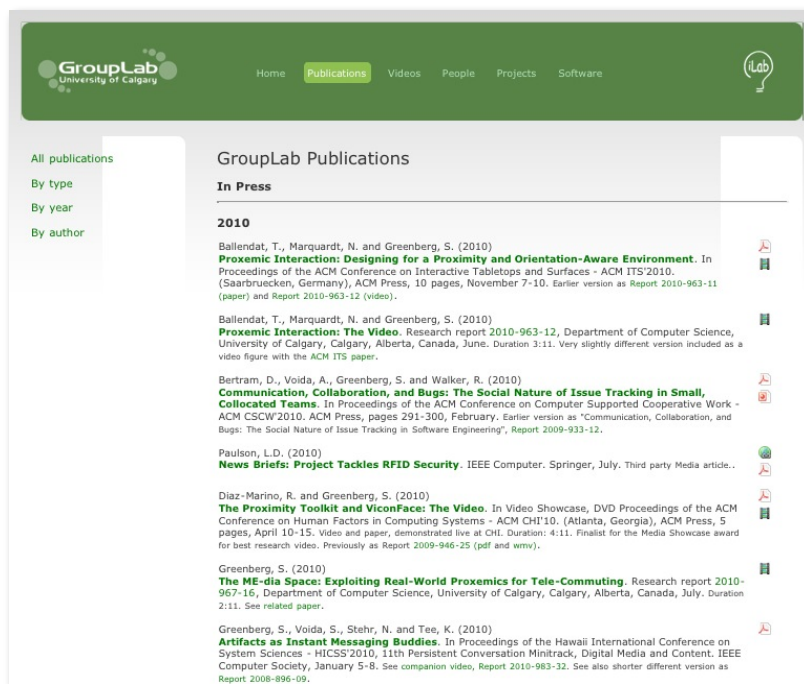


Figure 4-4: Saul Greenberg's Lab Page

Saul Greenberg's website represents a rich example of the individual and lab combination website type where the subject has two websites, one to represent his personal work (Figure 4-3) and one to represent the collective work of the research group (Figure 4-4) he is part of. It is important to note here that the subject sees both sites as being his own as evidenced by Greenberg's statement: "But still, [the research group website] is my website. I have full ownership of it and no one else is allowed to touch it." So it is not the case that the lab's website is separate from the personal website but rather as acknowledgement of the collaborative nature of the lab's research with Greenberg at the center.

One interesting thing is that Greenberg is the only respondent who sees the website as an archive of material much in the way that Kaye et al. (2006) reported of the researcher who had a box of material to grab in the event of a fire. In this way, what is on the website represents what is irreplaceable to Saul although not all of the components of what is irreplaceable are on the website. So while the site itself represents "what I can't lose", the original source documents are located on the file system of a personal system as distinct from what is available on the website. Another point is that not all of the content is directly hosted on the servers that hosts the site itself as the videos are provided on YouTube because of better usage statistics than what is currently available on the local servers.



**Figure 4-5: Saul Greenberg's Publication Page**

The publications page (Figure 4-5) is interesting in that Greenberg views this as the core legacy but the publications page is actually located on the group website. Greenberg does have his own publications in the CV on his site though. Another notable aspect of the publications page is that all publications have a local version of the document so that colleagues and students can get them directly from Saul's site. When asked about concerns with copyright, Greenberg indicates that the majority of the documents are covered by the ACM digital library policy that allows for self-archiving. However, in any case, Greenberg would provide them as he feels he has a right to do so. He suggests this is a common understanding within the researcher community.

Videos, which Greenberg argues are equally important in his specific domain, are also grouped together with the publications. This is one example of where objects placed in an institutional repository might omit the researcher's choice in presentation and lose the contextual information as a result. Here the visual presentation encodes contextual meaning; the videos, the presentation files and the



publication all belong together. Therefore if an automated harvesting system is used to gather Greenberg's website for posterity, capturing this aspect of the context would be important.

**GroupLab**  
University of Calgary

Home Publications Videos People **Projects** Software

**GroupKit**  
a groupware toolkit

GroupKit is a free, easy-to-learn Tel/Tk groupware toolkit from the University of Calgary. With GroupKit, programmers build applications for real-time, distributed computer-based conferencing. Examples include drawing tools, text editors, and special meeting tools that are shared simultaneously among several users. GroupKit has been used for prototyping groupware, investigating multi-user architectures and interfaces, and as a CSCW teaching tool.

GroupKit was constructed from our belief that programming groupware should be only slightly harder than building functionally similar singular-user systems. We have been able to significantly reduce the implementation complexity of groupware through the key features that comprise GroupKit. A runtime infrastructure automatically manages the creation, interconnection, and communications of the distributed processes that comprise conference sessions. A set of groupware programming abstractions allows developers to control the behaviour of distributed processes, to take action on state changes, and to share relevant data. Groupware widgets let interface features of value to conference participants to be easily added to groupware applications and are built by developers to accommodate the group's working style. Example GroupKit applications in a variety of domains have been implemented with only modest effort.

The diagram and the code fully defines a Hello World program written in GroupKit. What makes this a *real* Hello World program is that when one person presses the hello button, all people in the conference see that person say hello! Of course, there are many complex applications written in GroupKit, and it has been used to develop our many research prototypes.

**Primary Investigators**  
Mark Roseman (Chief Architect)  
Saul Greenberg (Supervisor)  
And many contributors over the years.

**Milestones**

- Full system implemented and freely available for downloading.
- We are now on GroupKit version 5.1, which was rewritten from the ground up to include support for Unix, Windows and Macintosh platforms, and also a new flexible meta-architecture.
- Many other researchers outside of GroupLab have used GroupKit.
- A variety of papers and a videotape were produced.

**Current Status**

- GroupKit has moved from research into groupware toolkit construction into a stable tool for developing groupware prototypes.
- Expect only minor enhancements and bug fixes.

**Figure 4-6: Saul Greenberg's Project Page**

Another key aspect of the publications is that Greenberg has put them in a content management system (in this case a wiki). While Greenberg acknowledges that this resulted in extra work both in transferring the content from the previous system to the current platform as well as in adding each publication to the CMS, the benefit is that he can quickly reconfigure the list. This quick reconfiguration is important, as different organizations require reporting in their own formats.

In terms of context, another area of interest in Greenberg's sites is the project page (Figure 4-6) that gathers together a visual representation of the projects with brief synopses, related publications and collaborators. While the project pages do have importance to Greenberg, he expresses concern because they are incomplete and difficult to maintain due to:

1. The project pages not being under personal control.
2. Projects evolving.
3. Bounds of the project being unclear.
4. Involves other people, hard to get students to update once done.

So while the project pages may be important particularly for identifying the context of sets of files, they provide an incomplete snapshot. When asked about the possibility of adding research data to the appropriate project, Greenberg expressed a desire to do so but felt that between the ethics concerns and the extra effort required to make the data meaningful, it would not be worth the effort. In addition, the software related to specific projects is located on a third site.

#### ***4.5.2 The Basic White Site: Ravin Balakrishnan***

Ravin Balakrishnan's website (Figure 4-7) represents a basic white layout with a few small exceptions. The standard basic white layout does not have a table of contents like the one in the upper left section nor does it include images with each publication entry. These are indicators that Balakrishnan's choice to do a basic white layout is deliberate and reflects a desire to keep an easy to maintain, easy to navigate site. One of the key features of the site is the publication section. Unlike other organization schemes, the publication list is in reverse chronological order numbered counting down from the most recent to the first entry. This was done to facilitate easy referencing of the site. Balakrishnan points out that in other approaches, the referrer often needs to copy down a complex URL or the entire citation to give to another person whereas the simplicity of his numbering approach

allows him to quickly tell a student to look up reference number 43 on his site and the student can navigate to that publication with minimal effort.

## Ravin Balakrishnan

Associate Professor & Canada Research Chair  
[Department of Computer Science](#)  
[University of Toronto](#)

### [Publications](#)

### [Courses](#)

### [Students & Postdocs](#)

### [Brief Biography](#)

### [Professional Activities](#)

### [Directions to my office](#)

email: [ravin at dgp.toronto.edu](mailto:ravin at dgp.toronto.edu)  
 voice: (416) 978-5359  
 fax: (416) 978-5184  
 office:  
 Room BA5270  
 40 St. George Street, Toronto  
 lab: [Dynamic Graphics Project](#)

*postal mail & courier address:*  
 Ravin Balakrishnan  
 Department of Computer Science  
 University of Toronto  
 10 King's College Road, Room 3302  
 Toronto, Ontario  
 Canada M5S 3G4  
 Tel (for courier): 416 978-6025

## Publications

### Refereed papers



110. Daniel Vogel, Ravin Balakrishnan. (in press). Direct pen interaction with a conventional graphical user interface. *To appear in Human-Computer Interaction.*



109. Jeremy Bimholtz, Abhishek Ranjan, Ravin Balakrishnan. (in press). Providing dynamic visual information for collaborative tasks: Experiments with automatic camera control. *To appear in Human-Computer Interaction.*



108. James Scott, Shahram Izadi, Leila Rezai, Dominika Ruszkowski, Xiaojun Bi, Ravin Balakrishnan. (2010). RearType: Text entry using keys on the back of a device. *To appear in Proceedings of MobileHCI 2010.*



107. Abhishek Ranjan, Jeremy Bimholtz, Rorik Henrikson, Ravin Balakrishnan. (2010). Automatic camera control using unobtrusive vision and audio tracking. *To appear in Proceedings of GI 2010 – the Graphics Interface Conference.*



106. Daniel Vogel, Ravin Balakrishnan. (2010). [Occlusion-aware interfaces](#). *Proceedings of CHI 2010 – the ACM Conference on Human Factors in Computing Systems*. p. 263-272.  
**CHI 2010 Best Paper Award**  
[video](#)

**Figure 4-7: Ravin Balakrishnan's Website**

Balakrishnan acknowledges that the primary effort needed to maintain the site itself is the creation of the visual representation of the publication beside each entry. In his opinion though, the image aids navigation and recall. However, he also notes that it increases the size of the page and increases loading times when away from a fast Internet connection. This is one case where the design choice in the site reflects going beyond the standard white design to accommodate a useful feature.

The effort to maintain a simple but memorable site is a reflection of the value Balakrishnan attaches to the site. He suggests that a personal academic website is "advertising in many ways," particularly in terms of recruiting graduate students.

The site also serves as a useful tool for Balakrishnan to answer questions particularly when networking at a conference. Finally, it serves as a repository of his work and in particular, videos associated with publications as libraries (in his view) are currently not archiving them. This is similar to Greenberg's efforts in associating videos (often a demonstration of the software or method) with publications. Again, this points to the existence of contextual metadata encoded in visual representation aimed at human access rather than machine access.

In another nod to the problem of institutions relating to the individual researcher, Balakrishnan states that the website is useful as a public point for things of his that do not fit elsewhere. This reinforces the idea that the site serves as a personal communication tool. Unlike Greenberg, Balakrishnan says that "[he] doesn't view it as an archive but more as communication with the outside world". Further, he states that he hasn't really thought about a building a legacy. In contrast, Greenberg has given careful consideration to having the website be the long-term reflection of the contributions of both his lab and his own work.

In terms of control, Balakrishnan shares many of the same desires. As with Greenberg, he is the sole updater for the site. The production process starts with a mirrored version of the site that resides locally on his laptop. He updates the local version and then copies it over to the server. He admits that he doesn't maintain the web server itself but that staff in his lab maintains it. This local control is important so that "if I want something changed, they'll do it as I pay their salary" and firmly states that he has "almost absolute control."

One final note is that Balakrishnan often utilizes a just-in-time approach to managing the site. In particular, Balakrishnan notes that he does not migrate files to a newer version unless he gets a request to. Similarly, while he is happy to make the data files available, the effort required to make the data files understandable cannot be justified for all projects. Only when he receives a request for a data file does he invest the work in documenting it.

### 4.5.3 The Extensive Site: Ben Bederson

Ben Bederson's site is an example of an extensive site blended with a personal / lab site. The site (Figure 4-8) serves as a good exemplar that covers multiple areas including personal content, prominent featuring of project pages and social media features like a blog. As well, the design of the page is atypical with an informational sidebar containing news, links to research and teaching as well as a contact box. There is another inset on the right side featuring books from a children's digital library project Bederson participated in.

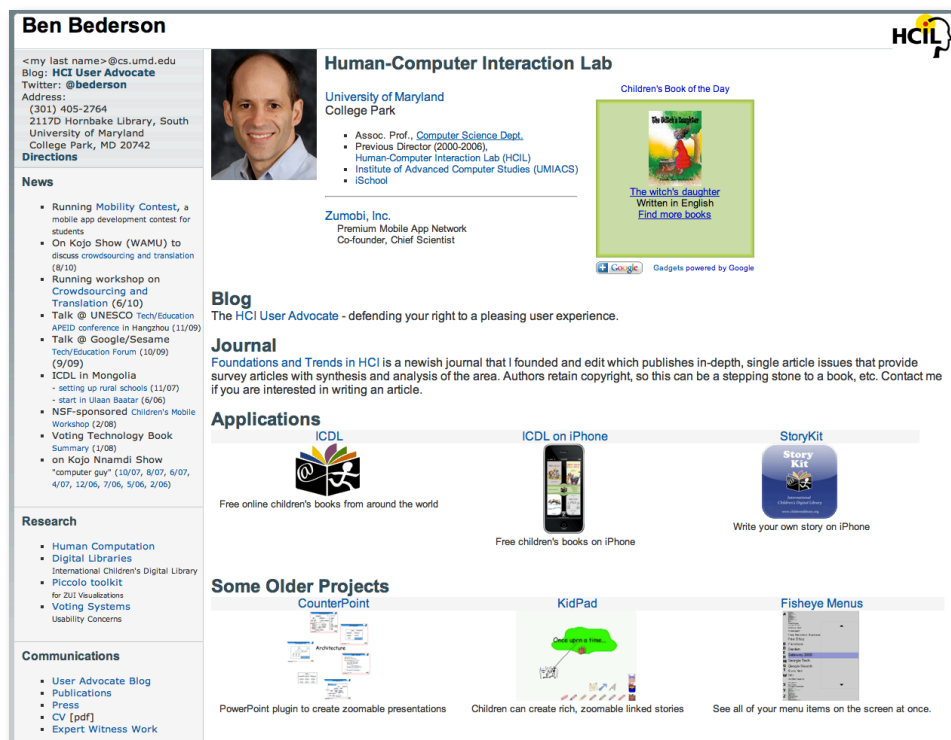
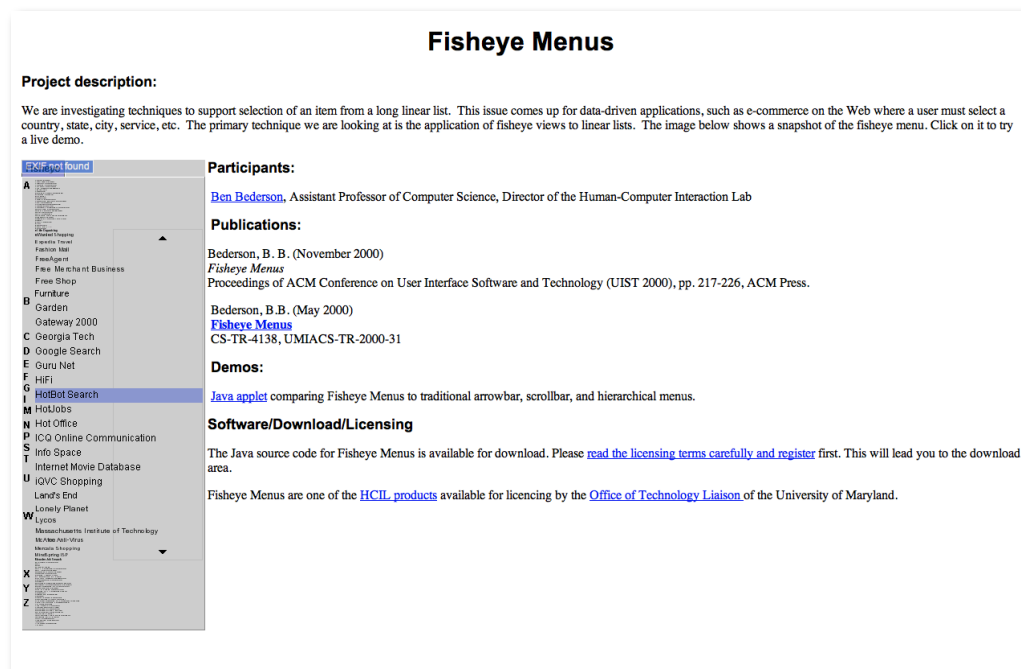


Figure 4-8: Ben Bederson's Website

In many ways, Bederson's site has a strong resemblance to Greenburg's in terms of layout and organization. However, unlike Greenburg's site where there is a distinctive separation on the main page between things like the project pages (which are grouped together in the lab section on Greenburg's site), Bederson's site appears to have a set of integrated pages that include the projects. It is only on

navigation to the project page itself that it is clear the project pages reside on lab pages or independent sites.

This leads to an interesting observation that Bederson makes in the interview. At the beginning of his career the primary focus of Bederson's web presence was on the projects as the only content available. However, the emphasis shifted from projects to information about him as a researcher as he advanced. As Bederson notes, at this point in his career, the balance of content is likely at a midway point between emphasis on projects and emphasis on him. At some point in the future, like his colleague Ben Shneiderman, Bederson speculates that his site might be entirely about himself with no mention of projects.



**Figure 4-9: Ben Bederson Project Page**

Figure 4-9 represents a typical project page on Bederson's site. The pattern of the visual representation, list of participants, publications and software is repeated through the project pages and represent a relatively full content set that could be useful for preservation efforts. The project page gathers a great deal of contextual information that would aid to preserving the individual files into a coherent whole

and well as providing information to future users. For instance, the visual representation can be used to get a sense of the software without having to actually install the software. Archivists could use that representation to determine the need to conduct preservation activities or consider if the representation is sufficient for access purposes.

Bederson's project pages serve as a good exemplar of the lifecycle challenge in archiving digital material. Pages in collaborative projects tend to be built with wikis or similar technology and reside in an independent location. However, these are built only when there is something to publicize; otherwise communication is done through other channels. When the project is complete, Bederson may extract elements of the project page to make a project page for his site. Given the use of wiki technology, a history of things like edits is available. But because it is located on an external site, archiving just content in Bederson's personal site will miss this kind of information.

In terms of visual appearance Bederson notes that the current website design is motivated by wanting something simple and easy to maintain. It is interesting that he grabbed the basic template from "somewhere on the web" demonstrating that while heavily concerned with the content, the layout and visual feel is not as much of an issue. Bederson points to an older version of the website which is more idiosyncratic and based on the motifs of one of his projects. While the older version was more visually unique and representative of his research, it was ultimately abandoned, as it was too difficult to maintain.

Bederson's website is also relatively unusual for the amount of personal content. On the site he links to his blog (Figure 4-10) as well as to a number of personal pages, including a series of photos of his daughter and a memoir of hiking in Alaska. Bederson also tracks mentions in the press about his research.



**Figure 4-10: Ben Bederson's Blog**

One observation of interest to digital preservation that Bederson raises is the issue of link permanence and persistent access. Rather than hosting the blog locally, Bederson chose to use a commercial service that allowed him to pull the content back to have a URL in one location. However, when the service discontinued the option, Bederson was faced with the choice of either having persistent URLs spanning two domains or changing the old URLs and likely causing links to them to break. As I described in the case of Mark Weiser, the issue of link breakage becomes even more significant when the creator is no longer able to update the links.



#### 4.5.4 The Organization Template: Jonathan Grudin

##### Jonathan Grudin

I work in the [Adaptive Systems and Interaction Group](#) at [Microsoft Research](#), part of the [Microsoft Corporation](#). My research is in human-computer interaction and computer supported cooperative work, with a particular focus on the design, adoption and use of group support technologies. Some of the work below was done in the [Collaborative and Multimedia Systems Group](#).

Prior to joining Microsoft Research, I was Professor of Information and Computer Science at University of California, Irvine. I have taught at Aarhus University, Keio University, and the University of Oslo, and worked at the MRC Applied Psychology Unit, Wang Laboratories, and MCC since earning my Ph.D. at UC San Diego. ([Life before Microsoft](#))

My wife Gayna Williams is a Usability Manager at Microsoft. Our daughters [Eleanor](#) and [Isobel](#) are not yet considering career options.



##### Projects

**Supporting Interaction In Organizations:** I am interested in challenges in designing and using technology to support people in group and organizational settings. These are quite different: we have interacted in groups for millions of years and the challenge is to support natural capabilities that are unlikely to be changed. Organizations have only existed for thousands of years and are very much subject to modification. In both cases, technology design, adoption and use require attention to context, including features that promote awareness of the relevant people, objects, and events, and through social conventions that elicit such information. As awareness technologies produce more efficient and effective interaction, they bring to the fore issues of privacy, trust, reciprocity, and accountability. Another challenge is the need to support work at a more detailed, individual level while at the same time enabling people to interact intelligibly. Among the papers below, some current interests appear in those on [information displays](#), [emerging conventions in technology use](#), [effects of digital representation](#), and [challenges in building widely useful platforms](#).

**Collaborative Information Retrieval:** Most examination of information retrieval is from the perspective of individuals. In this [NSF-sponsored project](#) investigators at Microsoft, Boeing, RISO, and the University of Washington are considering the way that workgroups or teams determine needs and retrieve and disseminate information.

##### CHI Academy and ACM TOCHI

In 2004 I was inducted into the [ACM SIGCHI CHI Academy](#). From 1997 through 2003 I was the Editor-in-Chief of [ACM Transactions on Computer-Human Interaction](#).

##### Recent Publications and Papers

###### Field Studies and Analysis

- [Managerial Use and Emerging Norms: Effects of Activity Patterns on Software Design and Deployment](#). J. Grudin, 2004. *Proc. HICSS-37*, CD-ROM, 10 pages. ([PDF](#))
- [Return on Investment and Organizational Adoption](#). J. Grudin, 2004. *Proc. CSCW 2004*, 274-277. ([PDF](#))
- [Personas: Practice and Theory](#). J. Pruitt and J. Grudin, 2003. *Proc. DUX 2003*, CD ROM, 15 pages. ([PDF](#))
- [Messaging and Formality: Will IM Follow in the Footsteps of Email?](#) T. Lovejoy and J. Grudin, 2003. *Proc. INTERACT 2003*, 817-820. ([PDF](#))
- [Information Seeking and Sharing in Design Teams](#). S. Poltrock, J. Grudin, S. Dumais, R. Fidel, H. Bruce and A.M. Pejtersen, 2003. *Proc. GROUP 2003*, 239-247. ([PDF](#))
- [Leaders Leading? A Shift in Technology Adoption](#). J. Grudin, 2003. *CHI 2003 Extended Abstracts*, 930-931. ([PDF](#))

Figure 4-11: Jonathan Grudin's Corporate Page

Jonathan Grudin's current website (Figure 4-11) is typical of the webpage of a researcher in a corporate environment. The layout is simple with a single column, white background and section headers. As described in the general overview, the bulk of the content is a selective publication list. The projects and research interests section is much smaller and describes them without providing much in the way of supplementary material. There are a number of exceptions to the typical corporate page though:

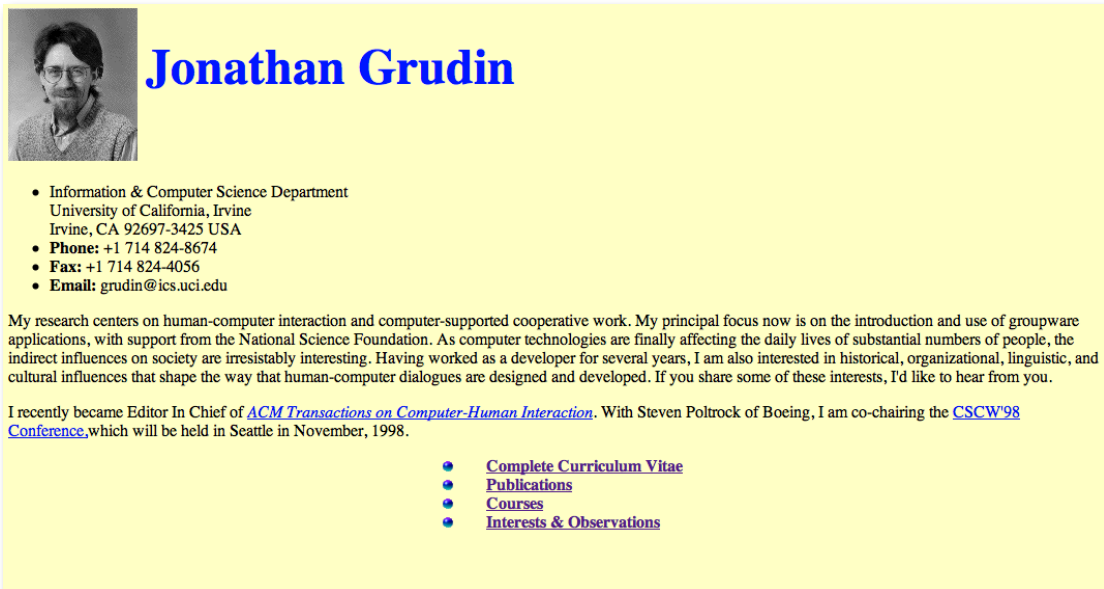
1. There is more content on the pages than is typical of a corporate webpage. As Grudin notes, Microsoft encouraged the researchers to maintain a webpage and as maintainer of the group website, Grudin felt obligated to add content

to the site. This suggests that a change in cultural expectation may be necessary to encourage researchers to archive their work.

2. There is only minimal corporate branding. Again, Grudin notes that while Microsoft would issue guidelines, they did not have strict policies on branding.
3. Grudin has some personal content on the site in the form of family photos.

Grudin's experience with his website is an interesting discussion on the issues of technology platform migration and site usage feedback. The site has not been updated in six years. Even though Grudin continues to be an active researcher, the abrupt cut-off is the result of a switch from one content management system to another. Rather than having content in two places, Grudin elected to wait for migrating the content from the old system to the new system to continue updating the site. Although the migration would only have required a couple days, Grudin never did find the time to migrate and so the site has not been updated since.

One additional lesson that might be learned from the experience is that the content management system did not provide a lot of feedback in terms of visitors to the site. One of the reasons Grudin gives for not committing the time to update the site is the lack of feedback. Given other priorities, the lack of feedback meant that Grudin never moved the maintenance of the site higher on the list of priorities. This may be an important lesson for archives wanting to encourage researchers to be more active in the preservation of their work. To motivate researchers, they need to have feedback about what is going on, either with how their work is being archived or the benchmarks of effort being invested.



**Figure 4-12: Jonathan Grudin's Previous Website**

Another note about Grudin's web presence is that he has a website at the former institution he taught at (Figure 4-12) and so his publications are currently spread over two different sites. This raises an interesting question as to which is the authoritative site or if in fact, both are required to understand Grudin's work. This is compounded in that neither site has a complete publication list; the current site has selected publications from work before Microsoft. In order to understand Grudin's full publication history, one would need to look at both sites.

This phenomenon of moving on may be one possible reason for the more general lack of concern for future legacy prevalent among computer scientists. Grudin states that over the course of his career, the technology has transformed dramatically. Because of this transformation, Grudin feels that that his early work with systems driven by punch card would not be useful to current problems in HCI. It is likely that this experience amongst computer scientists is common and might be why as a group there is not a strong concern for preserving the content.

## 4.6 Conclusion

---

By comparing these four case studies, we can see that there is a general consistency to how the personal website is viewed by researchers. While there are differences in the specific conceptualization, those differences may be related to the individual approach and understanding and also what role the website plays in regards to the activities of the individual. For the most part, the emphasis for the website is as a communication tool and not necessarily an archive for the future. The function of the websites tends to be for immediate needs like reporting publications to granting agencies and academic annual reviews.

To reiterate, though, the interviewees articulated a set of broad themes in terms of the value of the website and motivators. These themes were:

1. Strong desire for control.
2. Little institutional pressure to conform.
3. Trust in institutional systems.
4. Limited long term view.
5. The website is a public face, not an archival point.

For the most part, the case studies serve to reinforce these points. For instance, all of the researchers in the case studies exercise strong control over their web presence. As with any group, there are differences but these do not detract from the majority consensus and simply serve to highlight the challenges in providing a solution that will meet the needs of all concerned.

In the next chapter, I will take the results and analysis of the interviews coupled with the results from the survey of websites to derive some implications of the findings. I will discuss how it impacts the preservation of their digital content, issues of building tools and systems to support their work. Finally I will conduct a brief assessment of a tool currently in use that supports digital preservation.

# Chapter Five: Analysis and Implications

---

## 5.1 Chapter Synopsis

---

In the previous chapter I discussed the results of the interviews including general findings and specific case studies related to each of the types of sites identified in chapter three. In this chapter I will take the results from the interviews and the surveys to synthesize a set of recommendations for building systems for assisting with researcher personal websites. I will apply those recommendations to an existing system (DSpace) to identify matches and mismatches.

## 5.2 Impact of Findings

---

### ***5.2.1 The Digital Preservation Problem at Creator Scale***

For the past two chapters, I have discussed the nature of researcher personal websites without referencing the impact on digital preservation. We need to return to how researcher behavior, in regards to the design and maintenance of their websites, impacts the preservation of their work. As noted in the initial review of digital preservation in chapter two, the goal is to take digital objects of an intellectually atomic nature, gather sufficient metadata and manage the object over its lifecycle so as to ensure continued access over the long term. In order to relate how researchers are creating their websites to the digital preservation problem, how *atomicity*, *metadata* and *lifecycle* relate at a pragmatic level needs to be understood. In essence, this section reconsiders some of my statements in chapter two, where I now account for activities done at the creator level.

The first concept that I need to unpack is *atomicity* or the idea of the distinct (atomic) intellectual entity. The PREMIS data dictionary defines an intellectual entity as "a set of content that is considered a single intellectual unit for purposes of management and description" (PREMIS, 2008). Most people are familiar with dealing with digital files; it is relatively easy to grasp the idea of how to save and

work with a file. However with an intellectual entity, the challenge is in its definition. We may know that an intellectual entity is composed of multiple digital files. But which files belong? And which files do not? Where do the boundaries mark the end of one intellectual entity and the beginning of the next, particularly where the entities are part of the same project? Can projects be intellectual entities? And over time, can the bounds of what constitutes a single intellectual entity change?

Despite the above questions, the intellectual entity is understandable at the level of the work of the researcher. He or she begins work on a project and identifies one research question that can be explored in a study. Once the study is completed, he or she gathers the results and writes a paper. Later the paper is submitted to a conference and a presentation is prepared for it. The raw data, administrative documentation like ethics approvals, the analysis, the final paper and presentation file are distinct artifacts of the process but still belong together as a coherent whole.

The second concept to unpack is the idea of the *lifecycle* of the objects. As long as objects are deemed to be viable, work must be done to make them accessible. However, at some point, the object requires too much effort to make it accessible and is no longer viable. At this point, the responsible organization needs to either cede responsibility or retire the object. The life and death of a digital object represent two phases in the lifecycle of a digital object; phases for which much research has been done. The literature also talks about the object's birth or creation phase but aside from those created in institutions, there is not a lot of work done in the area.

Yet what is done at creation can have an impact on the rest of the lifecycle. For instance, if the creator mandates a clear retirement point for the object, it simplifies the determination of its "death". Similarly, if the creator articulates what is valuable in an object, it eases tasks like migration. The creator can also determine which institutions should have the responsibility of managing the digital object, allowing for smooth transition as necessary for those needing to cede responsibility.

The final concept to unpack is the idea of *metadata*. In discussing the lifecycle, it is clear that many of the difficulties encountered in the lifecycle issues of digital objects can be addressed by sufficient documentation. However unless that documentation is articulated in the language of digital preservation, its use may be limited. At an institutional level, this requires the use of standards that formalize the documentation (metadata). However creators are rarely trained to use these standards and more importantly the effort of creating such metadata is tangential to the research activity of the creator. This leads to a disconnection between the intents of the creator and what is required to do adequate digital preservation. Even if creators supply sufficient metadata, there is no guarantee that what is supplied is encoded in a way that allows for easy use. Similarly, terminology the creator uses often differs from what is required by archivists and curators. Even if the terms are the same, the assumptions and meanings behind them may not be.

As a result, it is doubtful that researchers can generate metadata for digital preservation purposes even if willing to do so (and this is unlikely). Moreover it may not be realistic to ask creators to do so given that their priorities are elsewhere. This leaves the option of having archivists create metadata for digital content. As this has been the traditional division of labour, it seems a reasonable compromise at first blush. However, digital preservation has a significantly compressed timescale compared to preserving analog materials. This means acting earlier to ensure that nothing is lost. Secondly, the volume of material is substantively greater. This is compounded by the ease of creating digital objects that are simply versions of an existing object. Without documentation to explain the relationship between the current and the prior versions from the creator, the archivist might spend significant resources either preserving the wrong version or trying to identify the correct version. This will lead to a delay in processing the content, extending the gap between an unmanaged and managed digital collection and even leading to errors in the handling. All of this suggests the need to involve the researcher in the preservation metadata creation process.

### ***5.2.2 How Do the Findings Relate?***

Having unpacked the issues with digital preservation at the creator level, they need to be related to the findings of researcher websites patterns. When I started the question was (and still is): *can we utilize what researchers are putting in their personal websites to ease the challenge of preserving their work?* From my findings, I can extend the summary in chapter five to this list:

1. Most websites are academic websites. There is a sense from the interviews that the primacy of academic websites within the result set is the product of the expectations for academic researchers to have a website gathering their work. As well, having a website generally benefits the individual researcher in other ways.
2. The websites are content rich in some areas, with content poor in others. The richness of the content focuses almost entirely on the bibliography. Researchers generally maintain the bibliographies fairly well. Other areas like project summary pages are not as well maintained if at all.
3. The designs of most sites are either personal in nature or of a basic white design. The almost universal goal of the website design is to keep the maintenance simple. While the creator almost always wants to be the one in control of their site, they have little time to maintain it. Thus simplicity is key.
4. The most common elements contained in the websites are the publications, contact information, an identifier photograph and a list of research projects or interest.
5. The bibliography or publications page is typically quite close to the front of the website. It is clear from the interviews that the publications are the most important element of the site.
6. The amount of personal content on the websites is minimal. Researchers generally identify the audience of the site as professional. When researchers



combine existing personal and professional sites into a single site, personal content tends to disappear.

7. Researchers tend to exercise substantive control over the website both for the purpose of saving time and to ensure that the information gets on the site correctly. There is also a sense that the work to maintain the site should be a personal effort.
8. There is little conformity to institutional pressures or templates in the design of the website. For the formatting of the publications though, there are some concessions to the needs of granting agencies and annual reporting formats.
9. Researchers generally trust the institutional systems to provide backup for their work although most maintain a copy on their own systems.
10. There is only a limited sense of a long-term time scale for the sites. If anything, there is a sense that the sites are likely to disappear in the future.
11. The websites are communication tools as opposed to archival tools. This biases the design and presentation of the sites to be primarily human readable with little consideration for machine readability. It is not (in general) that researchers are specifically choosing against machine readability but rather is not thought of at the time of design and construction.

#### *5.2.2.1 The Digital Object and the Intellectual Entity*

The findings suggest that researchers see the publication as the primary target for their intellectual output—this is not surprising to anyone with familiarity with the academic field. There are only a few cases in the site survey where other elements have equal or greater weight. This dominance tends to result in sites with publications as the anchor point. In an interesting comparison, few researchers invest in project pages as an online representation of their work even though most talk about the projects they are working on. In some ways, project pages would be tremendously useful as an archival unit as they sit at a level of granularity above that of the publication and allow for an understanding of the context of the

individual publication. However treating projects as units for preservation would be problematic given that project pages are incomplete and some of the interviewees lament the problems with maintaining the project pages.

Publications, on the other hand, tend to be fairly atomic in nature and can be preserved more easily in comparison. Based on the site survey, the publication as an intellectual entity is comprised of the document with the publication metadata and in some cases, associated presentations and videos. There is a definite visual clustering to the presentation of a publication that could allow the binding of these elements together for machine reading. With the right tools, it would be relatively easy to encourage researchers to create these publication units in a way that could be harvested as a whole.

However, a focus on preserving publications as archival units means that some related content does not have a fit. For instance, research data at a project level may not fit into any one publication as the data may be used for several publications. Whether the data should be archived separately, with one of the publications, or with all of the publications raises problems no matter the approach. If the data is archived separately, it is disconnected from its explanation, as it is likely the methods used to collect and analyze the data are contained in one of the publications. Archiving it separate from the methodology means losing the association that may make the data unusable. Archiving the data with its primary publication may cause the loss of context with other publications. Finally archiving the research data in every publication archival unit will significantly inflate the costs of preservation; data sets are often the largest part of a publication archival unit.

#### *5.2.2.2 Lifecycle Preservation Impacts*

From the standpoint of looking at the lifecycle of digital objects, the findings indicate that the website in its current form does not provide a particularly useful tool for harvesting digital objects created by researchers. The publication represents an end point for a group of related files created in preparing the publication, but the website rarely identifies these files. Similarly, the publications' lifecycle is not

captured by the website as most interviewees noted that they do not post pre-publication versions. Thus access to previous versions of a publication through the website is unlikely.

Similarly, interviewees noted that project pages are generally initiated at a point where there is something to communicate to the outside world. However, this often coincides with an initial publication and may be relatively late in the life of a project. A large number of digital files may have been created prior to the availability of the project pages. Thus project pages are not particularly useful for tracking objects over their lifecycle.

If anything, the primary use of the researcher's website is to provide a reasonable inventory of what might be available. One area it can assist is in locating the publications. Some researchers are particularly difficult to locate because of the commonality of their names. This is reflected in the ranking of researchers in the HCI bibliography as it uses the first initial and last name to locate and identify publications. Locating publications attributed to a researcher using this scheme is a matter of guesswork as a result. However, assuming that the researcher does not provide the publications as part of the site, the publication list should provide sufficient information to locate the publications in publishers' databases while identifying only those that belong.

Unfortunately, because researchers do not consider the website to be archival, much of the material that will be a challenge to archive is unavailable to automated harvesting of the website. As the interviews have brought out, the data files and other supporting source files tend to reside on local machines and servers. The association between the source files and the final publications tends to be only through file system conventions like named folders. This is one area where having a system that allowed a researcher to keep all the material together and expose to the public a limited subset might encourage the kind of bundling necessary to unify the archival units.

#### *5.2.2.3 Metadata*

Ultimately, the potential of exploiting the researcher's personal website rests in its ability to provide metadata. Contextual information about what digital objects belong with each other can be identified through the visual groups. A sense of the intent of the researcher in the handling of individual files in their archive might be identified in how projects and files are described on the site. One of the primary efforts of an archive is to create an index to a researcher's work so that it is accessible. In reducing the effort to make the work accessible, it increases the odds that the material will be viewed in the future. This in turn increases the odds that preservation activities like migration will be performed on the material. If it is possible to extract the kind of information from the researcher's website, it may mean easing the effort of archiving their work and increasing the odds it will be preserved.

The general preference of archivists is that metadata be clearly labeled to its respective field and formatted in specific ways with terms controlled by vocabulary lists. This is a high expectation to have of a researcher who is typically neither trained in the specifics of metadata creation nor in the broader issues of information science. As well, having metadata so clearly delineated is atypical of how publication metadata is presented on researcher websites, and does not meet the conventions of the researcher community. However, it is possible to embed this information into the HTML such that it is invisible to the viewer yet usable to the machine. Another possibility is to have the metadata entered into a database that then generates the display appropriate to the researcher community. Either way, these are options to enrich a researcher website to provide the necessary metadata that currently do not exist.

#### *5.2.2.4 Limitations of the Website*

The above discussion reflects the potential of using the creator's personal website as a curated archive. If all researcher websites reflected the ideal, it would be possible to systematically crawl their webpages on a regular basis. This would allow a

system to extract new and updated material and build an organized and coherent legacy archive of the researcher (like what the University of Maryland Archives has done with Ben Shneiderman's work as described in chapter four). This would provide a ready index to later materials deposited by the researcher.

However, researchers do not consider the website a personal archive. Consider teaching material. In many cases, the researcher will develop a fairly comprehensive set of pages for a given course. The pages will often include a set of presentation files that cover a majority of the lectures, supplementary readings and in some cases, even the students' work in the course. The course website is often the only record of what is taught. Since the site's value is in the present, these pages are often removed when the course is over. The course material might actually be the primary resource on a topic area, but because of the limited view, the resource is lost.

Another area that deviates from the ideal is in site maintenance. Many areas of the website are either poorly maintained or not maintained at all. As the interviewees mention, maintaining their personal websites ranks fairly low on the list of priorities. To accomplish the ideal scenario, researchers need to be convinced about the value of making the sites an explicit focal point for the archiving of their work. Alternatively, tools utilized in the making of the site must support researcher activity while adding only minimally to the workload and enhancing the archivability of the content. Systems could be developed to handle much of the work of generating metadata. In the next section, I present a series of design recommendations to assist in building such systems.

## 5.3 Design Recommendations

---

### ***5.3.1 Translating Findings To Design Recommendations***

I have examined the impact of how researcher websites are currently being constructed and why they are being constructed in the way they are. The next step is to translate that into a set of recommendations that might be able to assist

researchers in creating their personal websites while providing better tools to enable archiving of their content. In this section, I will articulate a list of recommendations that could guide the development of these tools. These recommendations follow.

*5.3.1.1 The researcher's identity and distinctiveness needs to be maintained at all times.*

As discussed, one aspect of the researcher's website is that the site is seen as a means for communicating the researcher, his or her activities and interests, and most importantly the research to the outside world. That a significant number of the websites do not have any visual affiliation except to the researchers is evidence of this. Similarly the researchers I interviewed were unanimous in keeping only content related to their research on the site. Therefore, a system designed to support researchers' websites needs to provide researchers with the ability to customize sites to their specifications. Even if a large portion of researchers use a similar design and organization, the sense of distinctiveness is important.

The primary impact of this is that any system or tool built to support researchers' websites should allow for customization of the look and feel of the website. In most cases, researchers are not strongly attached to a given design per se but rather that it not reflect poorly on them as researchers. As well, the survey identified significant variances in how the content is organized on the site itself, and the variety of content types. Because of this, a system supporting the researcher should accommodate and integrate this variety of content types and organization. As a final point, most researchers seemed content to use an institutional URL of some type. However there was greater concern that their material should be easily findable. Therefore, to support locating the researcher, the URLs need to be easy to remember and persistent. More importantly, the system should create webpages that can be easily indexed by search engines to allow the researcher to percolate to the top of the search results list.

*5.3.1.2 Publications are the core content and most often updated. There should be a set of easy to use workflow tools that allow the researcher to update them.*

Throughout this chapter, the point that publications are the most important content from both a researcher behavior and an archival standpoint has been emphasized. Researchers have identified two key requirements to support the updating of their publications: First, that there is a specific formatting to the publication entry that they prefer. Second, while updating does not require substantial investment in time, they frequently emphasize that their current system makes it relatively easy to do. A new system should not make it any harder to do (and ideally easier) and should support those who only update once in a while.

One specific area that could be facilitated is the entry of metadata for the publications. As noted, one challenge encountered in the process of identifying researchers in the survey is that typically only the first initial and last name is used in the publication metadata and often misattribution occurs in the HCI bibliography as a result. A tool that could encourage researchers to add their full name and the full name of collaborators while allowing drag and drop in name entry on subsequent entries to simplify the creation of publication entries. This is one example of where a tool could improve metadata creation while reducing work for the researcher

*5.3.1.3 Publications are the core archival unit. There should be ways to link related items that comprise the publications as a unit.*

Related to the previous point, the publication is also the core archival unit—this means that it represents a cluster of content all related to the publication. The survey identified a number of sites with this kind of clustering, usually with the publication document, presentation files and videos. As noted in the discussion on impact, an automated harvesting of the website would likely gather them as separate files with no relational information. This would result in a significant loss of context for both the publication and for the related items. As well, losing the link

to the publication metadata may result in files with little metadata describing their purpose or context.

A system for supporting researcher websites should allow the related content to be added to a publication entry and displayed alongside the publication entry itself. At the same time, the system should automatically generate a way of associating the additional files to the publication itself, likely through supporting descriptive and/or relational metadata. This metadata could be encoded within the display itself or made available on harvesting through a machine interface. Given that researchers may have related files that they may not want general access to, the system might also allow researchers to suppress the display of those files depending on the user context while allowing authorized archiving systems to harvest the content. This might be extended to the publication itself in instances where the researcher does not have the right to make it available. At the same time, hidden links provided by the system could allow the researcher to e-mail the link to requesters of the publication without having to dig for the file.

*5.3.1.4 The publication list is needed for a number of purposes. There should be flexibility in presentation.*

As noted in the interviews, granting agencies and institutions often require different presentation schemes depending on the agency or the institution. As a few interviewees have noted, they maintain their CV and the online publication list in different formats to support this. A few have commented that having to do this twice can be annoying. The varieties of ways of presenting the publications that were encountered in the survey is also notable. In addition, a few of the websites have publication lists that support not just the researcher but the entire lab.

A system designed to support researcher websites needs to provide a variety of presentation styles for the publications. This may include the kind of information included in each publication, the sort order of the publications (ascending or descending), whether to have continuous numbering or section numbering of publications, the sort key used for ordering the publications and whether to have



sections based on chronology, publication type or venue. Given that the publication list has multiple uses, it makes sense to have this presentation be dynamic. Indeed, a number of existing websites allow the viewer to modify the presentation style on the fly so this should be supported as well. Also, some websites have an associated graphic for the publication that should be supported. Finally a few of the websites allow the viewer to search the list.

*5.3.1.5 Researchers should be encouraged to create more archivable content through easy to use tools like wizards and graphical user interfaces.*

Most of the recommendations have been about supporting what the researcher is already doing. One reason is the consistent reporting of the lack of time to do anything new. This lack would extend to applying greater effort to creating more metadata or adding significantly more content to the site. It would be unlikely that researchers could be convinced to do something that is neither mandated nor results in immediate personal benefit. Therefore it makes sense to focus on making things easier for researchers while using system conventions and hidden system automation to make the content more archivable.

However this does not mean that researchers should not be encouraged to add more content or take greater steps to create more archivable content. If the system provided an easy to use set of tools it might be possible to convince researchers to do so. For instance, it would make sense to have related publications associated in the descriptive metadata about each publication. An interface might present the researcher with the list of existing publications during the creation of a new publication. The researcher could then check off related publications. One benefit to the researcher is that it could enable the display of related items in the presentation of that publication to the viewer and thereby direct a viewer to newer, more up to date work. This would provide a feature with immediate, tangible benefits for the researcher while facilitating long term benefits in archiving. Such an approach might be used to provide a low cost means to populate project pages by automatically adding related items when an item is added to a project.

*5.3.1.6 Few researchers will start with little or no content. The ability to migrate from other systems or to other systems easily is a necessity as a result.*

One area of specific concern for the researchers interviewed is the challenge of moving from one system to another system. Most respondents concede that one of the reasons they still have a website constructed using just HTML and a web server is the effort to move to a new system. This stems from a number of problems. First is the challenge of setting up a new system and configuring/customizing it. Second is learning the new system. Third is moving the content from the old website to the new system. While some of the researchers admitted that their current system is a bit clunky and takes more effort to update than it should take, the cost in effort is measured in the hours whereas the cost to move to the new system is measured in weeks or months.

A system designed for supporting researcher websites should provide the ability to import information from an existing site into the new system. There are certainly examples of this outside of the sphere of academic tools. For instance, Posterous ([www.posterous.com](http://www.posterous.com)) allows the import of data from previous weblogs into their blogging system. In fact, the process is relatively painless and usually requires pointing to the Posterous system to the old website and it does the rest. One challenge for researcher websites is that few are using content management systems and much of their content is idiosyncratic in presentation. It is unlikely that a system could be built to accommodate all presentation styles. However, a template-based approach might facilitate the import where the researcher creates the template once and the content can then be automatically parsed.

Similarly, the system should provide the ability to export all content. This would reassure the researcher that when it comes time to move to the next system, the transition would be more painless than the initial transition from their HTML based or closed content management system (CMS). The ease of moving away from the system might encourage the researcher to adopt the system.

*5.3.1.7 Researchers generally trust institutional systems but prefer substantial control over the system. The design of the system should reflect that.*

When it comes to the backup of the content, researchers generally expressed confidence in the dependability of institutional systems. Similarly, none expressed concerns about the hosting of the webpages on institutional servers. But all expressed a desire to have control over how the system worked. As one interviewee noted, the more control (as expressed in chain of reporting), the quicker something was to get done.

For a system to gain acceptance of the researchers, it is clear that much of the control needs to remain in the hands of the researcher or his or her designate. The divide between what the researcher controls and what he or she can let go appears to be with the invisible and automatic. Subsystems like backup can remain firmly in the control of the system administrator. But changes to how the content is displayed or additional static pages need to remain in the hands of the researcher. Some websites have interactive elements or dynamically updated elements through RSS feeds that would have to be accommodated by the system. Whether this is something that the researcher could enable or only a system administrator could activate may impact the willingness of the researcher to adopt the system. Finally, the system's design needs to reassure researchers that their content is in their control at all times.

## 5.4 Evaluation of Current Systems

---

### ***5.4.1 Limitations of Current Approaches***

It is important to note that in general, researchers are content with their current approaches to hosting their website. When asked, few could identify things that could make the management of the site significantly easier. Therefore, while there are limitations, these limitations appear to be more annoyances than outright concerns. As a few researchers have acknowledged, updating can take longer than it should but since content only gets updated sporadically, this is not a significant concern.

The limitations of the current researcher approaches therefore fall more on the archiving community. The inability to easily harvest content and metadata, the inability to identify discrete units to archive, the inconsistency in the how publication metadata is formatted, the lack of content, these things mostly are problems at the institutional level. And in large part, these are only problems that arise long after the researcher is done with the material.

However this is not to say that researchers are opposed to change. Most admitted that there are likely better ways to manage their content. Even researchers who have implemented content management systems point out limitations. It is largely the complexity of moving and the minimal perceived gain that prevents them from doing so. Where current systems are limited in terms of things like binding associated content with core elements like publications, these are not seen as limitations.

#### ***5.4.2 Current Archiving Systems***

As it falls to archives to convince researchers to adopt systems more amenable to digital preservation activities, it is reasonable to look at how current digital archiving systems hold up against the proposed design recommendations. This would identify gaps between the current digital archiving systems and what would be required to satisfy researchers. I have mentioned institutional repositories already like Zentity (<http://research.microsoft.com/en-us/projects/zentity/>) and DSpace ([www.dspace.org](http://www.dspace.org)) whose goal is present the research output of an institution. I will return to DSpace later as the case study for the design recommendations. The majority of systems for digital preservation focus on back end management of the files and ensuring that the files are secure. Systems like LOCKSS ([www.lockss.org](http://www.lockss.org)), Fedora Commons ([www.fedora-commons.org](http://www.fedora-commons.org)) and IRODS ([www.irods.org](http://www.irods.org)) address the bit level preservation and management of files. Other systems focus more on the digital publishing workflow like the Open Journal System ([pkp.suf.ca/ojs/](http://pkp.suf.ca/ojs/)) or EPrints ([www.eprints.org](http://www.eprints.org)). A few focus on creating digital libraries like Greenstone ([www.greenstone.org](http://www.greenstone.org)).

The next section will analyze DSpace for the point of view of the design recommendations. One reason for focusing on DSpace is that it is the most common IR software with 949 institutions using DSpace (DSpace, 2010) at present. It is also the most mature of the dedicated IR systems. Given this, DSpace is typically the system that an organization like a library will have for digital preservation. The next section will compare the features and capabilities of DSpace against each of the recommendations, noting limitations and where DSpace does address the recommendation.

### 5.4.3 Analysis of DSpace

5.4.3.1 *The researcher's identity and distinctiveness needs to be maintained at all times.*

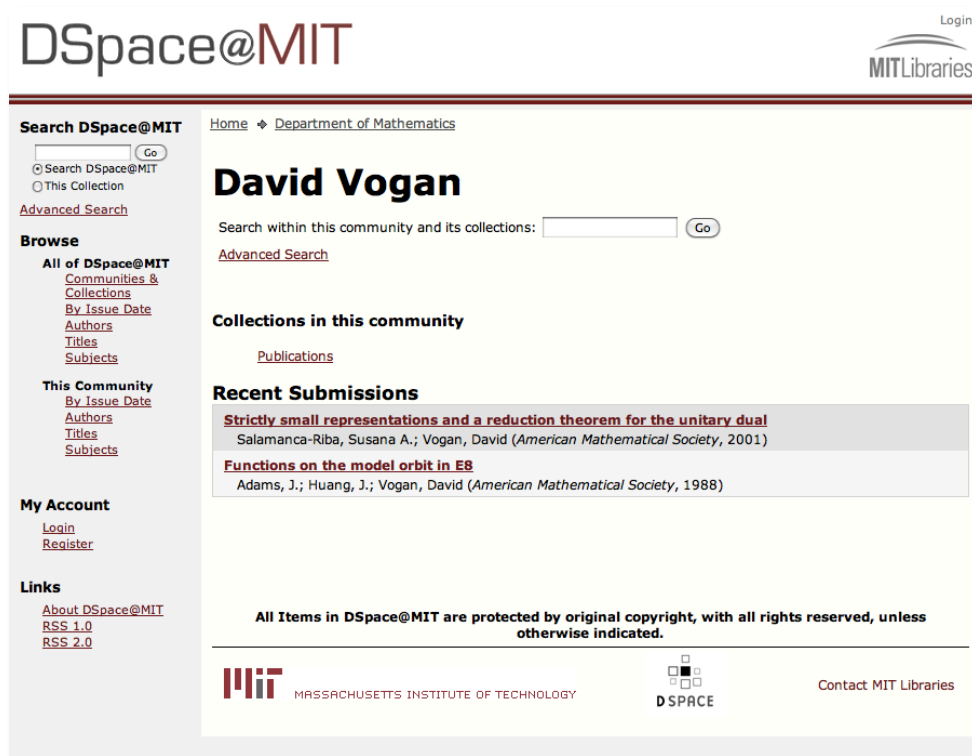


Figure 5-1: DSpace at MIT

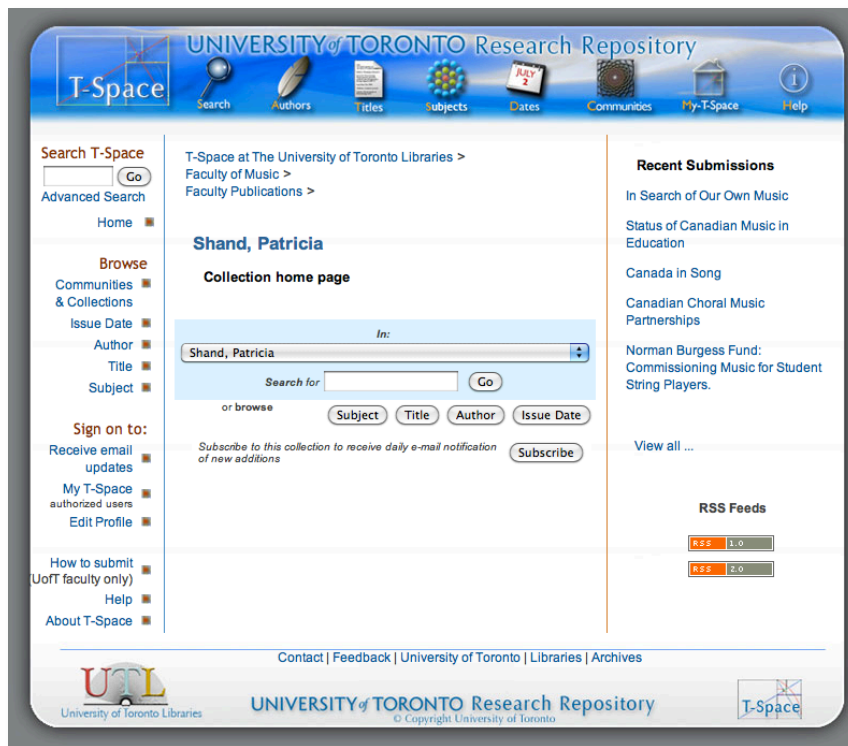


Figure 5-2: DSpace at the University of Toronto

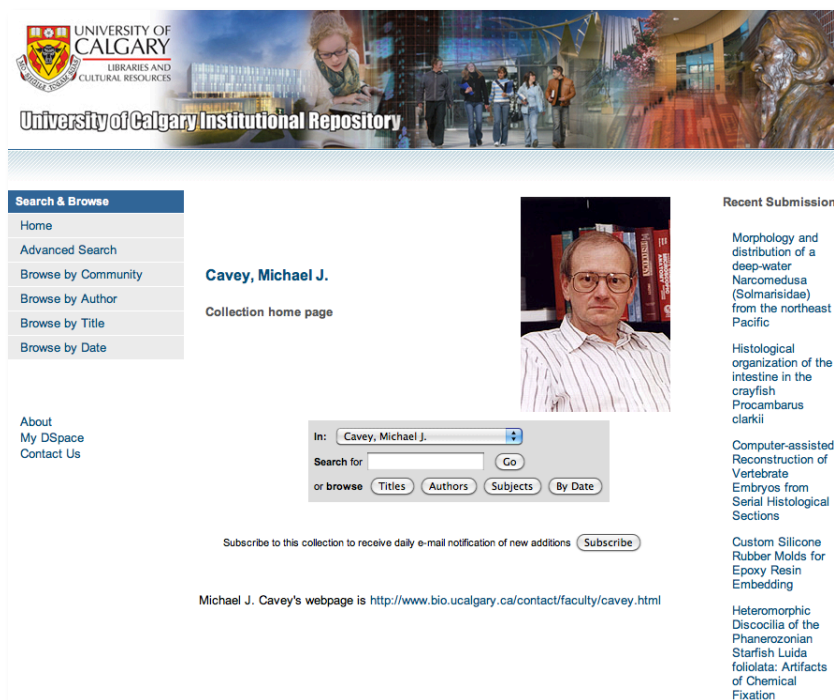


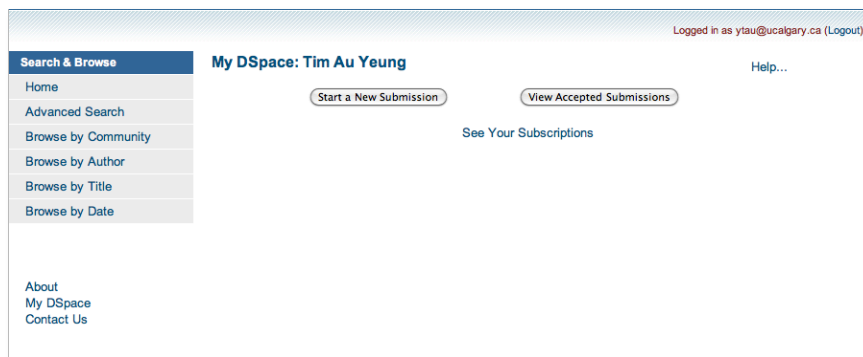
Figure 5-3: DSpace at the University of Calgary

Above are presented the “home” page for a researcher in three different implementations of DSpace: MIT (Figure 5-1), University of Toronto (Figure 5-2) and University of Calgary (Figure 5-3). In each implementation, the primary branding is that of the institution. This page, if integrated with the researcher’s website, would stand out as being foreign to the researcher website. Conversely if the researcher were to base their personal website on top of this implementation of DSpace, the branding would be that of the institution and not the researcher. Clearly this would not satisfy the requirement to present the researcher uniquely. Moreover, customization to add more researcher branding resides in the hands of the system administrator and not the researcher, which will be discussed further in the section on system control.

*5.4.3.2 Publications are the core content and most often updated. There should be a set of easy to use workflow tools that allow the researcher to update them.*

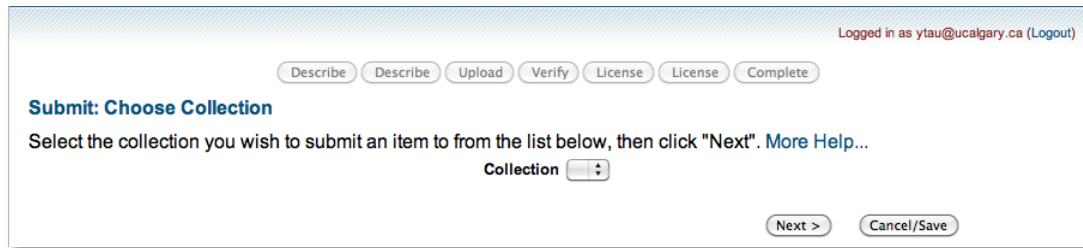
As noted in the description of DSpace and institutional repositories in general, the primary focus is on managing publications. It stands to reason that its handling of publications is reasonably solid. A multi-step wizard drives submission of a publication as follows:

#### 1. You start a new submission



**Figure 5-4: DSpace New Submission Screen**

## 2. Select the collection



Logged in as ytau@ucalgary.ca (Logout)

Describe Describe Upload Verify License License Complete

**Submit: Choose Collection**

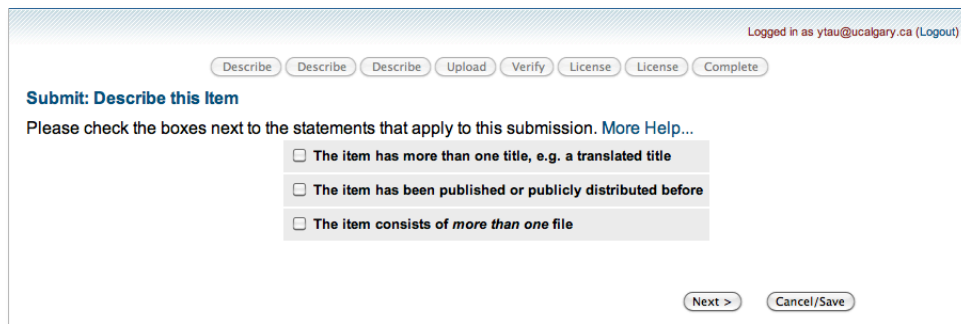
Select the collection you wish to submit an item to from the list below, then click "Next". [More Help...](#)

Collection

Next > Cancel/Save

**Figure 5-5: DSpace Choose Collection Screen**

## 3. Determine if your publication has more than one title, file or has been published before



Logged in as ytau@ucalgary.ca (Logout)

Describe Describe Describe Upload Verify License License Complete

**Submit: Describe this Item**

Please check the boxes next to the statements that apply to this submission. [More Help...](#)

- ☐ The item has more than one title, e.g. a translated title
- ☐ The item has been published or publicly distributed before
- ☐ The item consists of *more than one* file

Next > Cancel/Save

**Figure 5-6: DSpace: Describe Item Screen**

## 4. Assuming that none of the above questions are true, you are then passed to the page where you can enter the metadata for the publication



Describe Describe Describe Upload Verify License License Complete

**Submit: Describe this Item**

Please fill in the requested information about this submission below. In most browsers, you can use the tab key to move the cursor to the next input box or button, to save you having to use the mouse each time. ([More Help...](#))

Author(s)  Add More

Institution

Faculty

Department

Title

Publisher URL

Granting Agency

Grant Number

Refereed

Identifiers  Add More

Department ID  Add More

**Figure 5-7: DSpace: Describe Item Screen 2**

The above four steps represent the first three stages of creating a publication entry. There are five more as noted in the progress bar at the top of the screen shot in Figure 5-7 before the submission process is complete. However the description of the four steps should be sufficient to make the point that the process is tedious and a researcher would find it more time consuming. Moreover, if the citation had been created by the publication venue, the researcher could just copy and paste it to their current website. As well, while the metadata in Figure 5-7 is customizable on a collection-by-collection basis (which works if each researcher is treated as a collection), it is under the control of the system administrator.

5.4.3.3 Publications are the core archival unit. There should be ways to link related items that comprise the publications as a unit.

## Near-optimal bin packing algorithms

[Show full item record](#)

Citable URI: <http://hdl.handle.net/1721.1/57819>

**Title:** Near-optimal bin packing algorithms  
**Author:** Johnson, David S., 1945-  
**Advisor:** Michael J. Fischer.  
**Department:** Massachusetts Institute of Technology. Dept. of Mathematics  
**Publisher:** Massachusetts Institute of Technology  
**Issue Date:** 1973  
**Description:** Thesis (Ph. D.)--Massachusetts Institute of Technology, Dept. of Mathematics, 1973.  
Vita.  
Bibliography: leaves 398-399.  
**URI:** <http://hdl.handle.net/1721.1/57819>  
**Keywords:** Mathematics

### Files in this item

Files	Size	Format	View	Description
<a href="#">Preview, non-printable (open to all)</a>	13.61Mb	PDF	<a href="#">View/Open</a>	Preview, non-printable (open to all)
<a href="#">Full printable version (MIT only)</a>	13.61Mb	PDF	<a href="#">View/Open</a>	Full printable version (MIT only)

Figure 5-8: DSpace Item View Screen

DSpace does allow a publication unit to possess multiple files. It also allows for a description associated with each file to display the relationship of the file to the publication unit. However, it does not allow for the defining of relationships external to the item. So there would be no way of defining a project as an item and linking the documents. Nor would it allow linking across items that had a shared relationship.

5.4.3.4 The publication list is needed for a number of purposes. There should be flexibility in presentation.



Figure 5-9: DSpace: Browsing by Title Screen at U of C



Figure 5-10: DSpace: Browsing by Title Screen at U of T

As can be seen from Figure 5-9 and Figure 5-10, the way the publications can be presented can be altered to reflect some variances in formatting. However, neither display reflects the typical citation format nor does either resemble the majority of

the styles in which publication information is presented on researcher webpages. One core piece of information missing from these views is where it is published—one needs to go to the individual item view to see that information.

Similarly, while multiple items associated with a publication are displayed with visual clustering in the publication lists on researcher websites, this is not provided by the default DSpace views. Publication type (book chapter, article, poster, etc.) is also not available for either sorting or grouping. As with most aspects of DSpace, this can be customized via add-ons and custom programming but these tend to impact the entire system or require considerable effort to implement.

*5.4.3.5 Researchers should be encouraged to create more archivable content through easy to use tools like wizards and graphical user interfaces.*

Currently, there are two areas where DSpace encourages more archivable content: through additional metadata and explicit licensing. As can be seen from workflow screenshots above, DSpace adds more metadata than a researcher would typically add. Assuming that the presence of the blank fields encourages the researcher to add this metadata, it does serve to advance the archivability of the publications.

Similarly, DSpace administrators can set publications so that submitters are encouraged to attach a license to the work with the primary option being a Creative Commons license ([www.creativecommons.org](http://www.creativecommons.org)), which can encourage distribution of the publication. Presenting the option of having a license allowing for wider distribution encourages the researcher to think in terms of broad access that can result in a greater interest to preserve the work.

However, both of these affordances comes at the cost of the researcher being required to work through a longer process in order to create a publication entry. As noted above, this may impact the researcher's willingness to implement DSpace. Going beyond the individual publication, DSpace provides no facility to gather information at the level of the project nor does it provide affordances to construct relationships between related publications.

*5.4.3.6 Few researchers will start with little or no content. The ability to migrate from other systems or to other systems easily is a necessity as a result.*

The DSpace system allows for batch import of previous content into the system. However, there are two limitations to what DSpace provides. The first is that the process for a batch import requires constructing a complex set of descriptor files. The second is that the import tool is only usable at the command line, meaning it is in the domain of the system administrator and not available to the researcher.

Consequently, the DSpace import system represents all of the fears of researchers in terms of moving content from an old system to a new system. It is a complex system that is poorly documented. It requires considerable effort. And the import system is beyond their control in an institutionally hosted DSpace.

*5.4.3.7 Researchers generally trust institutional systems but prefer substantial control over the system. The design of the system should reflect that.*

One dominant theme emerging from this analysis of DSpace is that the majority of the customization and the special operations exist solely in the domain of the system administrator. If the researcher deploys DSpace as their personal repository system, many of these issues disappear. However, much of the value of having researchers use DSpace is that by hosting the system, the archive would have the ability to access the back end data structures and extract the contextual information. In order to address this recommendation, more of the control over the system needs to be shifted to the researcher without resorting to system administrator status.

## 5.5 Conclusion

---

The survey and interviews have provided good insight into the motivations of a researcher for creating a personal website and for the choices in both system selection and content. Relating it back to the digital preservation challenge, the findings have an impact on the identification of relevant archival units, where the researcher's activity falls into the context of the preservation lifecycle and finally

what metadata researchers are creating either deliberately or inadvertently as part of the process of constructing their website.

In analyzing how researchers create their websites, there are gaps between that and current institutional systems and approaches. These gaps can be ameliorated through systems designed more with the researcher in mind. To review, here are the design recommendations that could address those gaps:

1. The researcher's identity and distinctiveness needs to be maintained at all times.
2. Publications are the core content and most often updated. There should be a set of easy to use workflow tools that allow the researcher to update them.
3. Publications are the core archival unit. There should be ways to link related items that comprise the publications as a unit.
4. The publication list is needed for a number of purposes. There should be flexibility in presentation.
5. Researchers should be encouraged to create more archivable content through easy to use tools like wizards and graphical user interfaces.
6. Few researchers will start with little or no content. The ability to migrate from other systems or to other systems easily is a necessity as a result.
7. Researchers generally trust institutional systems but prefer substantial control over the system. The design of the system should reflect that.

**Table 5-1: Design Recommendations**

Current software being used by institutions focus on institutional needs and as such, there is a mismatch with researcher needs. The evaluation of a common system (DSpace) in use by institutions against the recommendations demonstrates this. DSpace does provide some things that researchers need but the analysis shows there are significant areas where DSpace does not serve the needs of the researcher.

In the next chapter, I conclude by reiterating the findings and identify areas where the researcher and the institutional archival communities are moving closer. As well, I propose new avenues where work could be done to bring the digital preservation and researcher communities closer together.

## Chapter Six: Conclusion

---

In this thesis, I started with the question: “Is it possible to use researchers’ personal websites to archive their work?” To make this broad question tractable, I broke it down into a set of three questions to find out what researchers currently do, how they feel about their current approach, and what could be done to make their efforts useful from a digital preservation perspective (see section 6.1 for the specific phrasings). To answer these questions, I surveyed existing websites of senior researchers in HCI to identify the content elements and how they are presented on these sites (chapter three), interviewed a subset of these researchers on how they approach creating the sites and whether the approach is adequate (chapter four), and from these derived a set of design recommendations suggesting how a digital preservation system needs to be built to meet researcher needs (chapter five). In this final chapter, I return to the original research questions (restated from section 6.1) review how their answers contribute to the literature (section 6.2) and suggest possible future work based on this research (section 6.3).

### 6.1 Research Questions

---

1. **What is the current state of personal websites for researchers?** There are currently no concerted efforts to study researcher websites. For this reason, it is important to understand what researchers are making available on their websites and how that can impact the preservation of their corpus.
2. **Could researchers be motivated to create preservable content?** At the moment we know researchers are creating websites. But we do not know whether they are creating them strictly for current needs or whether there is a sense of legacy that might motivate them to create preservable content. Similarly we do not know what are the kinds of motivators that would encourage them to change or adopt new tools and systems that institutions might provide.

3. **How can institutions facilitate making the content preservable?** Even if the content is preservable, it does not mean it is usable by the institutions. Moreover, even if it is usable, it may require significant work on the part of institutions. What can institutions do to facilitate the researcher process while encouraging the creation of easily usable content for the institution?

## 6.2 Thesis Contributions

---

1. *A Snapshot of Current Researcher Website Practises.* In chapter three, I answered question 1 by surveying a set of websites from senior researchers working in Human-Computer Interaction to get an idea of what they are putting on their websites and how they are presenting their content. The results of the survey argue strongly that researchers are creating sites that present themselves in a professional context. The bibliography of their work is the core content with secondary sections on their projects and research interests and their teaching. The sites contain almost no external content or personal information. I also identified a typology of researcher website consisting of basic professional, extensive, researcher/lab and organization template sites.
2. *Researcher Motivations for Website Choices.* In chapter four, I answered question 2 and added detail to question 1 by interviewing nine researchers to get a sense of why they create their websites, how they approach choosing tools and content and what they value about their website. I used Kaye et al.'s (2006) five values of a personal archive as the basis for trying to understand what the researchers are choosing to do with their sites. I identified that for the most part, researchers view their sites as communication tools even as they value the site as a complete record of their work. However, they viewed this record as a public projection rather than an archive. Researchers are also generally happy with their approaches to creating and managing the website content but acknowledge the possibility for improvement. This suggests that



institutions wishing to archive researcher work will have to convince the researcher to adopt tools with better archival attributes. As part of this work, I presented a set of 4 case studies, focusing each on a researcher who created one of the four types of sites identified in the survey. I did this to demonstrate commonalities and differences, and to highlight some of the more interesting results.

3. *Archival System Design Recommendations.* In chapter five, I answered question 3 by analyzing the results of the survey and interviews in the context of preserving the researchers' work. From the analysis, it is clear that there is currently a mismatch between what researchers are doing with their personal websites, and what would be needed to gather content from these personal sites into archival systems. However given that researcher websites may be the most complete index of the researcher's corpus, this strongly argues that further work should be done with researchers to provide what is needed. Given the lack of time researchers have to invest in this, institutions will have to accommodate the researcher. This has not been done to date. I proposed a set of design recommendations (Table 5-1) to build archival systems that might encourage researchers to provide content in a way could be archived easier. Finally I provided a critique of a current archival system (DSpace) against these design recommendations.

## 6.3 Future Work

---

The work done in this thesis barely scratches the surface of what is needed in this area. The divide between the institutional archival programme seeking to preserve research, its raw data and its final output vs. the researcher creating all of these artefacts is considerable. There are significant questions as to how this divide could be mitigated, who should be doing the archiving, and when it should be done. While not all of these questions can be answered here, one question that can be addressed is how to make this process easier.

At the onset of my work, I originally wanted to develop and evaluate a prototype archival system that would serve both the needs of the researcher and the needs of the institution. It became fairly obvious early on that building yet another system was both premature and counter-productive, so my focus shifted to understanding current practices and issues, and then considering design implications. The research in this thesis proves, if anything, that we need to understand the broader practises of the research community in greater detail before we can build appropriate tools. Talking with the researchers, one gets the sense that they have some inkling of the possible preservation issues coming in the future, but they do not yet have a full awareness of those implications.

If anything, my research demonstrates that a broad change in the behaviour of institutions and individual researchers is required; something that is unlikely to occur overnight by the delivery of some new system. This change will take time. However, with my research is a starting point, a number of concrete steps that might be undertaken to move us closer to that point. Specifically, I propose three concrete areas where further work could be undertaken. First, further work is needed to broaden the scope of survey and the interviews so as to better understand researcher goals and practises. Second, further work is needed to test the design recommendations and refine them against feedback from both curators and researchers. Finally, further work is needed to develop a prototype to provide evidence of the value of the design recommendations.

### ***6.3.1 Broadening the Surveys and Interviews***

The design of the surveys and interviews was intentionally limited so that it could be answered in the confines of a Masters thesis. As I note in both the methodology of the survey and the interviews, the community of even a single research discipline is challenging to address. Yet, this limits the external validity of my findings. There certainly is a question as to whether my results apply to other disciplines. For example, are we dealing with an anomalous community whose characteristics run counter to the wider research community, or is this community (given its expertise)

a reasonable predictor of what other communities will do in the near future?

Similarly, in selecting senior researchers, I chose a group that should have a larger amount of content to deal with. It is not an unreasonable assumption that most researchers will get there at some point but this is not guaranteed. There is also the question of whether the behaviours evinced by the researchers I surveyed and interviewed are a cohort effect, representative of all researchers in the area, or an effect of the senior phase in their career.

An obvious expansion of the research would be to survey a broader range of researchers within HCI. It would also be helpful to gather appropriate demographic information from the researchers to begin to do a more in-depth statistical analysis against factors like gender, age and seniority. A version of the interview done in questionnaire form could be used to broaden the scope without incurring the kind of resources necessary for the interviews and provide an easier to analyze data set.

Another expansion of the research would look at other disciplines using the same methodology to see if the kinds of behaviour and approaches to personal websites hold across disciplines. The HCI community could be unusual among academic disciplines for its use of videos, its heavy emphasis on conference presentations and development of software systems. Moreover, the HCI community does seem unusual in the amount of external press coverage it receives. Investigating another discipline could demonstrate that the findings only relate to preserving work in the HCI community. Or it could prove that there is enough commonality across the disciplines to warrant a multi-disciplinary approach to preserving individual researcher output.

### ***6.3.2 Design Recommendation Refinement***

The design recommendations I present in chapter five represent suggestions to the research and archival communities, based on the findings of the survey and the interviews. However, there needs to be a lot of work done to operationalize and to evaluate the design recommendations to ensure that they are actually useful. Some of the work needs to come from gathering more information based on the previous

section. Will the recommendations hold against the behaviours of other research disciplines or researchers at different stages of their careers?

In addition, the design recommendations can also be refined by feedback from both the researcher and the archival communities. As I noted earlier, developing a working software system is likely premature. However, using rapid and low fidelity prototyping approaches, mock-ups based on the design recommendations could be presented to curators and researchers in a focus group setting. The focus groups could be used to identify areas where the design recommendations could be operationalized and strengthened so as to reflect actual practise in both communities, with specificity added to guide software development.

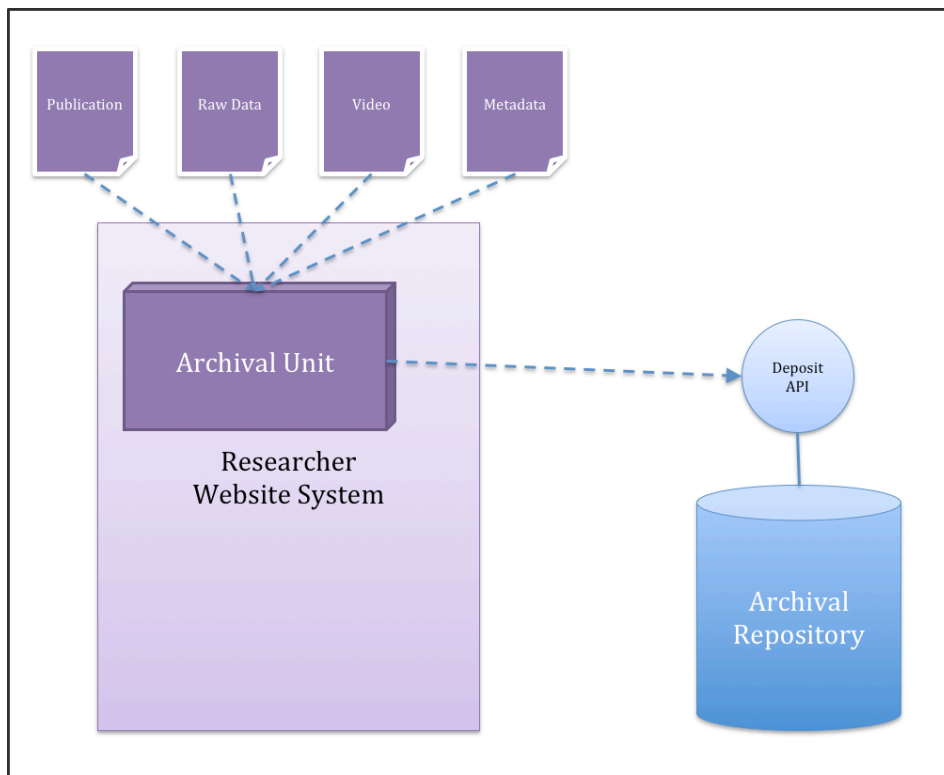
### ***6.3.3 Prototype System***

The final expansion to the research is to return to the original goal of developing a prototype software system. While I asserted that building a functional system was premature at the end of this phase of my research, further refinement and information gathering based on expanding the survey and interview programme and testing the design recommendations with appropriate communities could provide sufficient data to develop a prototype system.

One of the challenges of developing a new system would be how much information about user behaviour should be recorded. Would users be comfortable with a system with hooks built in to provide feedback and detailed statistics for the purposes of evaluating the system? Would it dampen the uptake of such a system? Yet having the system provide information about how the user interacts with the system would be vital for understanding where there are mismatches between the design and how researchers behave.

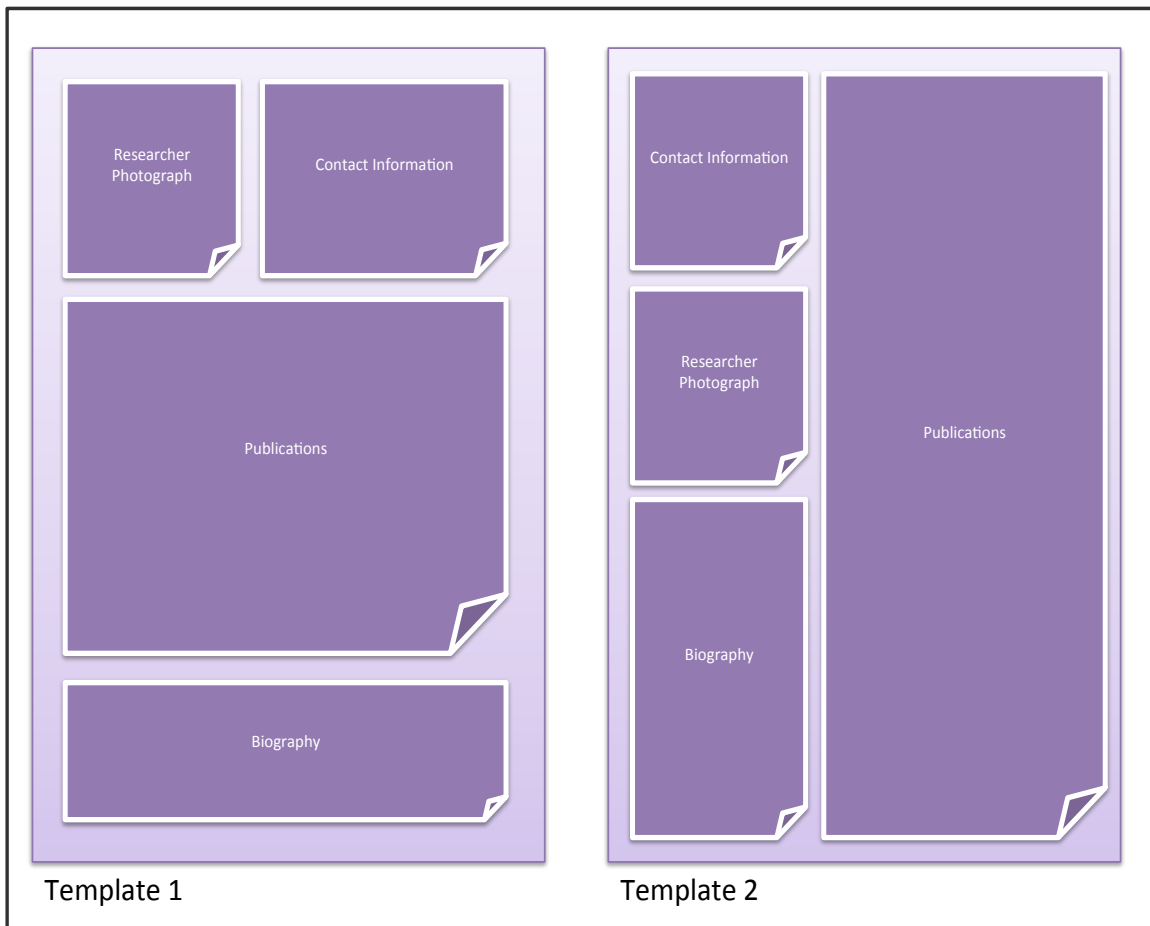
Setting aside the concerns about the kinds of feedback the system provides to external parties, it is clear that there are opportunities for building a system. As much as researchers like have control over their site, the 4 types of sites demonstrate that there is a commonality to their approach in site building.

Moreover, there are a number of core elements that are standard among the sites. These common elements can serve basic building blocks within a system. The widespread adoption of blogging software and wikis shows that people are willing to give up some control and flexibility for a simple to use system. As well, repository systems often have an application-programming interface to allow other systems to interact with them. A prototype system could be built to address primarily the researcher needs and then push the archival unit to the repository system like so:



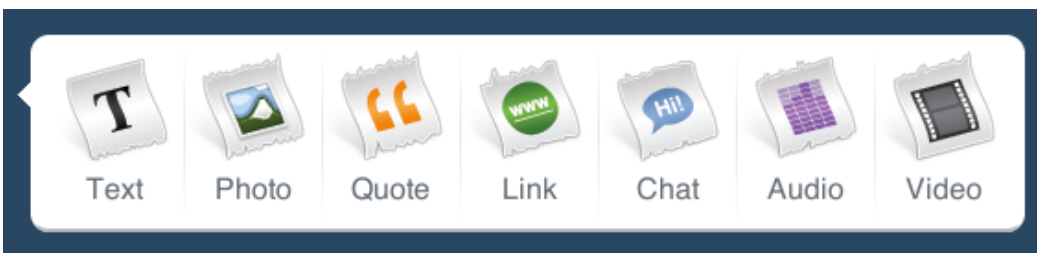
**Figure 6-1: Prototype System Data Flow**

The prototype system could be template-based with content types filling custom blocks. Each block could have its own data entry form based on the type of content and the blocks could be re-arranged depending on the preference of the researcher like so:



**Figure 6-2: Prototype System Templates**

Current systems like Tumblr (<http://www.tumblr.com>) allow for this kind of type-based approach to adding content:



**Figure 6-3: Tumblr Content Type Bar**

As Figure 6-1 suggests, researchers' websites need not be integrated with the archiving system. While managing all content with a single system has some advantages, the issue of control and system management identified in the design

recommendations means that it may be better in this case to have separate systems: one under the control of the researcher and one under the control of the institution. There are a number of approaches that could be taken to accomplish this. One possibility is to have the archiving system be the primary system that the researcher adds content, like publications, to. The archiving system would then provide a way to integrate that content back to the researcher's website; this could be done through a REST based interface for generating responses to web queries. Another approach might be to have the content structured on the researcher's website so that the archiving system could come in and retrieve the content on a periodic basis. This could be done using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) or similar protocol, allowing the archiving system to retrieve more metadata than would be available through the standard web interface.

Another area to consider with a prototype system is the inclusion of highly individual content types. Throughout this thesis, the goal has been to identify common elements in websites so that there would be a tractable problem with a solution that meets common needs. However, researchers tend to be idiosyncratic and often have unique needs. For instance, while little personal content was identified in the website survey, there were sites that did have personal content like family photos and hobby information and resources. A system that could support the inclusion of this kind of content would be more appealing to a researcher. Such content need not be tagged for long-term preservation and in fact, could be flagged for removal under certain conditions (like the death of the researcher).

In fact, the idea of utilizing a flag for the status of the researcher might also open the door to the addition of other content related to the researcher. For instance, researchers affiliated with the primary researcher may be able to add supplementary material; this might address a problem that Marshall (2008) notes where given the nature of collaborative research, a researcher may not "own" all the constituent parts of a research project. Another possibility is to draw in content like comments springing from community review processes being suggested in the

context of open access and other “open” movements within the transformation of scholarly communication umbrella.

## 6.4 Final Comments

---

Archivists invest a tremendous effort to collect, describe and organize research so that future generations can access this research. Often, the work that archivists do seems to parallel and duplicate efforts by researchers to present their research to their communities and to the organizations that house and fund them. It only makes sense to try to bring these two groups together to ensure that research is being effectively preserved for future generations. But the challenge is: how can we bring these two groups together?

Right now, archivists are hoping that researchers will deposit their research into repository systems. But the current experience is that researchers are not doing this. In this thesis, I put forward the proposition that we should instead use the researchers’ personal websites as a starting point. Based on surveying researcher websites and interviewing researchers about their motivations for how they create their websites, there appears to be promise that researchers are doing some initial organizing that could be leveraged. But there is also a significant mismatch between the intent of the researchers and the goals of archivists to preserve their work. Based on the interviews, it is clear that much of the impedance is not about researcher resistance to archiving, but a lack of motivation, time and resources to do so. Researchers are using approaches that do not lend themselves to easy extraction of the information. Yet switching to (current) institutional repository systems would interfere too much with a researcher’s processes.

My design recommendations are initial suggestions. We need to think harder about how repository systems could support researchers, while still gathering preservation information. My suggestions are not meant to be an end point, but rather the starting of a dialogue. Hopefully, the recommendations will be seen as an



opening between the communities to find common ground to build systems that support the aims of both groups.

## References

---

1. Alexander, J. (2002). Homo-Pages and Queer Sites: Studying the Construction and Representation of Queer Identities on the World Wide Web. *International Journal of Sexuality and Gender Studies* 7(2/3): 85-106.
2. Alperin, J. P., Fischman, G.E. & Wilinsky, J. (2008). "Open Access and Scholarly Publishing in Latin America: Ten Flavours and a Few Reflections." *Liinc em Revista* 4(2). Last viewed November 1, 2010.
3. Asirvatham, A. and K. Ravi (2001). "Web Page Classification based on Document Structure." from IEEE National Convention.
4. Barjak, F., Li, X. & Thelwall, M. (2007). Which Factors Explain the Web Impact of Scientists' Personal Homepages? *Journal of the American Society for Informaton Science and Technology* 58(2): 200-211.
5. BBC (2002). Digital Domesday Book Unlocked. *BBC News World Edition*. <http://news.bbc.co.uk/2/hi/technology/2534391.stm>. Last viewed November 1, 2010.
6. Beagrie, N. (2005). Plenty of Room at the Bottom? Personal Digital Libraries and Collections. *D-Lib Magazine* 11(6).
7. Beagrie, N. & Jones, M. (2001). *Preservation Management of Digital Materials: A Handbook*. London, British Library.
8. Bell, G. (2001). A Personal Digital Store. *Commuications of the ACM* 44(1).
9. Bernstein, M., van Kleek, M., Karger, D. & schraefel, m.c. (2007). Wicked Problems and Gnarly Results: Reflecting on Design and Evaluation Methods for Idiosyncratic Personal Information Management Tasks Southampton. University of Southampton: <http://eprints.ecs.soton.ac.uk/14668/>. Last viewed November 1, 2010.
10. Besser, H. (2000). Digital Longevity. *Handbook for Digital Projects: A Management Tool for Preservation and Access*. M. K. Sitts. Andover, Massachusetts, Northeast Document Conservation Center.

11. Boardman, R. & Sasse, M. A. (2004). Stuff Goes into the Computer and Doesn't Come Out. *CHI Letters* 6(1): 7.
12. Bortree, D. S. (2005). Presentation of Self on the Web: an ethnographic study of teenage girls' weblogs. *Education, Communication & Information* 5(1).
13. Bruce, H., Jones, W., & Dumais, S. (2004). Keeping and re-finding information on the web: What do people do and what do they need to do? *Proceedings of ASIST 2004*.
14. Buten, J. (1996). Personal Home Page Survey. from <http://www.asc.upenn.edu/USR/sbuten/phpi.htm>. Last viewed November 1, 2010.
15. Cedars Project (2002). *Cedars Guide to Preservation Metadata*. Leeds, U.K., University of Leeds.
16. Center for Research Libraries and OCLC. (2007). *Trustworthy Repositories Audit & Certification: Criteria and Checklist*. Chicago, Center for Research Libraries.
17. Chandler, D. (1998). Personal home pages and the construction of identities on the web. from <http://www.aber.ac.uk/media/Documents/short/webident.html>. Last viewed November 1, 2010.
18. Consultative Committee for Space Data Systems (2002). *Reference Model for an Open Archival Information System (OAIS)*. Internet: <http://public.ccsds.org/publications/archive/650x0b1.pdf>. Last viewed November 1, 2010.
19. Dominick, J. (1999). Who Do You Think You Are? Personal Home Pages and Self-Presentation on the on the World Wide Web. *Journalism and Mass Communication Quarterly* 76(4): 646-658.
20. DSpace (2010). Who's Using DSpace." from [http://www.dspace.org/index.php?Itemid=151&option=com\\_formdashboar](http://www.dspace.org/index.php?Itemid=151&option=com_formdashboar)d. Last viewed November 1, 2010.

21. Dumont, K. & Frindte, W. (2005). Content analysis of the homepages of academic psychologists. *Computers in Human Behavior* **21**: 73-83.
22. Foster, N. F., Gibbons, S., Bell, S. & Lindahl, D. (2007). *Institutional Repositories, Policies and Disruption*. Rochester. Rochester: University of Rochester.
23. Gemmell, J., Bell, G., & Lueder, R. (2006). MyLifeBits: A Personal Database for Everything. *Communications of the ACM* **49**(1): 7.
24. Granger, S. (2000). Emulation as a Digital Preservation Strategy. *D-Lib Magazine* **6**(10).
25. Gray, A. (2009). Personal Webpages of Academics: Design and Content. from Personal Communication.
26. Greenhow, C., Robelia, B., & Hughes, J.E. (2009). Learning, Teaching, and Scholarship in a Digital Age. *Educational Researcher* **38**(4): 246-259.
27. Haynes, D. (2004). Metadata for Information Management and Retrieval. London: Facet Publishing.
28. Henderson, S. (2004). How Do People Organize Their Desktops? *CHI '04 Extended Abstracts on Human factors in Computing Systems*. Vienna, Austria.
29. Higgins, S. (2007). Draft DCC Curation Lifecycle Model. *International Journal of Digital Curation* **2**(2): 5.
30. Hodge, G. (2000). Best Practises for Digital Archiving. *D-Lib Magazine* **6**(1).
31. Hodge, G. & Frangakis, E. (2004). Digital Preservation and Permanent Access to Scientific Information: The State of the Practice. A report sponsored by International Council for Scientific and Technical Information (ICSTI) and CENDI US Federal Information Managers Group.
32. Hodkinson, P. & Lincoln, S. (2008). Online journals as virtual bedrooms?: Young people, identity and personal space. *Young: Nordic Journal of Youth Research* **16**(1): 27-46.
33. Ishii, K. (2000). A Comparative Study of Personal Web Pages. *Internet Development in the Asia Pacific*. Singapore: International Association for Media and Communication Research.

34. Jensen, K. B. & Helles, R. (2005). Who do you think we are?  
*Interface://Culture–The World Wide Web as Political Resource and Aesthetic Form*. K. B. Jensen. Frederiksberg, Denmark and
35. Jung, T., Youn, H. & McClung, S. (2007). Motivations and Self-Presentation Strategies on Korean-Based “Cyworld” Weblog Format Personal Homepages. *CyberPsychology & Behavior* **10**(1): 24-31.
36. Kaye, J., Vertesi, J., Avery, S., Dafoe, A., David, S., Onaga, L., Rosero, I. & Pinch, T. (2006). To have and to hold: exploring the personal archive. *Proceedings of the SIGCHI conference on Human Factors in computing systems*. Montreal, Quebec, Canada, ACM: 275-284.
37. Kenney, A. R., McGovern, N. Y., Entlich, R., Kehoe, W.R. & Buckley, E. (2004). Digital Preservation Management: Implementaing Short-term Strategies for Long-term Problems. from <http://www.icpsr.umich.edu/dpm/index.html>. Last viewed November 1, 2010
38. Kim, H. & Papacharissi, Z. (2003). Cross-cultural differences in online self-presentation: A content analysis of personal Korean and US home pages. *Asian Journal of Communication* **13**(1): 19.
39. Knight, G. (2006). A Lifecycle Model for en E-Print in the Insitutional Repository. London: SHERPA DP.
40. Kretschmer, H. & Aguillo, I. (2004). Visibility of collaboration on the Web. *Scientometrics* **61**(3): 405-426.
41. Lamb, R. & Davidson, E. (2002). Social Scientists: Managing Identity in Socio-technical Networks. *35th Annual Hawaii International Conference on System Sciences (HICSS'02)*.
42. Lavoie, B. & Dempsey, L. (2004). Thirteen Ways of Looking at...Digital Preservation. *D-Lib Magazine* **10**(7/8).
43. Lawrence, G. W., Kehoe, W. R., Rieger, O.Y., Walters, W.H. & Kenney, A.R. (2000). *Risk Management of Digital Information: A File Format Investigation*. Washington: Council on Library and Information Resources.

44. Lawrence, S., Coetzee, F., Glover, E., Pennock, D., Flake, G., Nielsen, F., Krovetz, B., Kruger, A. & Giles, L. (2000). Persistence of Web References in Scientific Research. *IEEE Computer* **34**(2): 26-31.
45. Library of Congress (2002). Preserving Our Digital Heritage: Plan for the National Digital Information Infrastructure and Preservation Program. Washington, D.C., Library of Congress.
46. LIFE Project (2005). LIFE: Life Cycle Information for E-Literature. from <http://www.life.ac.uk/>. Last viewed November 1, 2010.
47. Lupovici, C. & Masanes, J. (2000). *Metadata for Long-Term Preservation*. <http://nedlib.kb.nl/results/D4.2/D4.2.htm>. Last viewed November 1, 2010.
48. Marco, M.J.L. (2002). A Genre Analysis of Corporate Home Pages. *LSP and Professional Communication* **2**(1): 41-56.
49. Marcus, B., Machilek, F., & Schutz, A. (2006). Personality in Cyberspace: Personal Web Sites as Media for Personality Expressions and Impressions. *Journal of Personality and Social Psychology* **90**(6): 1014-1031.
50. Marshall, C. C. (2007). How People Manage Personal Information over a Lifetime. *Personal Information Management*. Jones and Teevan, eds. Seattle, Washington: University of Washington Press.
51. Marshall, C. C. (2008). Rethinking Personal Digital Archiving, Part 1. *D-Lib Magazine* **14**(3/4).
52. Marshall, C. C. (2008). Rethinking Personal Digital Archiving, Part 2. *D-Lib Magazine* **14**(3/4).
53. Marshall, C. C. (2008). From Writing and Analysis to the Repository: Taking the Scholars' Perspective on Scholarly Archiving. *Proceedings of JCDL'08*, Pittsburgh, PA, USA.
54. Marshall, C. C., Bly, S., & Brun-Cottan, F. (2007). The Long Term Fate of Our Personal Digital Belongings: Toward a Service Model for Personal Archives. *Proceedings of Archiving 2006*. Ottawa, Canada: Society for Imaging Science and Technology.

55. Martyniak, C., Nadal, J., Ryder, B., Frangakis, E., Blood, G., Brown, K., Bynrnes, M. & Mielkle, S. (2007). Definitions of Digital Preservation. from <http://www.ala.org/ala/mgrps/divs/alcts/resources/preserv/defdigpres0408.cfm>. Last viwed November 1, 2010.
56. Mock, K. (2001). An Experiemental Framework for Email Categorization and Management. *SIGIR 2001*, New York: 392-393.
57. NINCH Working Group on Best Practices (2002). *The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials*. Online, The National Initiative for a Networked Cultural Heritage. <http://www.nyu.edu/its/humanities/ninchguide/>. Last viewed November 1, 2010.
58. OCLC/RLG Working Group on Preservation Metadata (2002). Preservation Metadata and the OAIS Information Model: A Metadata Framework to Support the Preservation of Digital Objects. Dublin, Ohio: OCLC.
59. Peterson, K. (2006). Academic Web Site Design and Academic Templates: Where Does the Library Fit In? *Information Technology and Libraries* (December 2006): 217-221.
60. Piwowar H.A., Day R.S. & Fridsma D.B. (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLoS ONE* 2(3): e308. doi:10.1371/journal.pone.0000308 .
61. Planets Project (2006). About Planets. from <http://www.planets-project.eu/about/>. Last viewed November 1, 2010.
62. PREMIS Editorial Committee (2008). *PREMIS Data Dictionary*. Washington, D.C.
63. Quisbert, H., Korenkova, M. & Hägerfors, A. (2007). Towards a Definition of Digital Information Preservation Object. *Proceedings of the 2nd International Conference on Metadata and Semantics Research*, Corfu, Ionian Academy.
64. Reich, V. & Rosenthal, D. S. H. (2001). LOCKSS: A Permanent Web Publishing and Access System. *D-Lib Magazine* 7(6).

65. Rick, J. (2007). AniAniWeb: a wiki approach to personal home pages.  
*Proceedings of the 2007 international symposium on Wikis*. Montreal, Canada, October 21-23, 2007.
66. Rick, J. (2007). *Personal home pages in academia: The medium, its adopters, and their practises*. College of Computing, Georgia Institute of Technology.  
**Doctor of Philosophy.**
67. RLG-OCLC Digital Archive Attributes Working Group (2002). *Trusted Digital Repositories: Attributes and Responsibilities*. Dublin, Ohio: OCLC.
68. Rosenthal, D. S. H., Robertson, T., Lipkis, T., Reich, V. & Morabito, S. (2005). Requirement for Digital Preservation Systems. *D-Lib Magazine* **11**(11).
69. Rothenberg, J. (1999). Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation. Washington, DC: Council on Library and Information Resources.
70. Schutz, A. & Machilek, F. (2003). Who owns a personal home page? A discussion of sampling problems and a strategy based on a search engine. *Swiss Journal of Psychology*.
71. Seeing Double Project (2004). Seeing Double: Emulation in Theory and Practice. from <http://www.variablemedia.net/e/seeingdouble/>. Last viewed November 1, 2010.
72. Staples, T., Wayland, R. & Payette, S. (2003). The Fedora Project: An Open-Source Digital Object Repository management System. *D-Lib Magazine* **9**(4).
73. Stern, S. (1999). Adolescent Girls' Expression on Web Home Pages: Spirited, Sombre and Self-Conscious Sites. *Convergence* **5**(22): 22-41.
74. Stern, S. R. (2004). Expressions of Identity Online: Prominent Features and Gender Differences in Adolescents' World Wide Web Home Pages. *Journal of Broadcasting & Electronic Media* **48**(2): 218-243.
75. Swan, A. & Brown, S. (2005). *Open access self-archiving: An author study*. Cornwall, UK: University of Southampton.  
<http://eprints.ecs.soton.ac.uk/10999/>. Last viewed November 1, 2010.



76. Thomas, C. & McDonald, R. H. (2007) Measuring and Comparing Participation Patterns In Digital Repositories: Repositories by the Numbers, Part 1. *D-Lib Magazine* **13**(9/10).
77. van der Hoeven, J., B. Lohman, & Verdegem, R. (2007). Emulation for digital Preservation in Practice: The Results. *International Journal of Digital Curation* **2**(2).
78. van Doorn, N. & van Zoonen, L. (2007). Writing from experience: presentations of gender identity on weblogs. *European Journal of Women's Studies* **14**(2): 143-159.
79. van House, N. A. (2007). Flickr and public image-sharing: distant closeness and photo exhibition. *CHI '07 extended abstracts on Human factors in computing systems*. San Jose, CA, USA, ACM: 2717-2722.
80. Walker, K. (2000). "It's Difficult to Hide It": The Presentation of Self on Internet Home Pages. *Qualitative Sociology* **23**(1): 99-120.
81. Waters, D. & Garrett, J. (1996). *Preserving Digital Information. Report of the Task Force on Archiving of Digital Information*. Washington, DC: The Commission on Preservation and Access.
82. Whittaker, S. & Sidner, C. (1996). Email Overload: Exploring Personal Information Management of Email. *CHI '96: Proceedings of the SIGCHI conference on Human factors in computing systems*: 276-283.
83. Whittaker, S. & Hirschberg, J. (2001). The Character, Value, and Management of Personal Paper Archives. *ACM Transactions on Computer Human Interaction* **8**: 20.
84. Wilkinson, D., Harries, G., Thelwall, M. & Price, L. (2003). Motivations for academic web site interlinking: evidence for the Web as a novel source of information on informal scholarly communication. *Journal of information Science* **29**(1): 49-56.

## Appendix A: Researchers Websites Surveyed

# of Publications	Name	# of Publications	Name
515	Nielsen, J	75	Mynatt, E
202	Shneiderman, B	74	Dourish, P
196	Carroll, J	73	Baecker, R
155	Myers, B	73	Liu, Y *
125	Greenberg, S	73	Wright, P
118	Lee, J	73	Jacko, J
114	Salvendy, G	72	Norman, D
108	Marcus, A	72	John, B
106	Grudin, J	71	MacKenzie, I
105	Ishii, H	71	Fox, E
104	Rosson, M	71	Olson, J
100	Hudson, S	70	Abowd, G
99	Buxton, W, Buxton, B	69	Jones, M
98	Muller, M	69	Lee, S *
96	Pemberton, S	68	Plaisant, C
93	Dix, A	68	Wogalter, M
90	Stephanidis, C	67	Witten, I
89	Sutcliffe, A	67	Karat, J
88	Card, S	67	Paterno, F
87	Gutwin, C	67	Smith, J *
87	Chen, H *	67	Green, T
87	Perlman, G	66	Winograd, T
84	Balakrishnan, R	66	Vanderdonckt, J
83	Landay, J	65	Lee, C
82	Benford, S	65	Whittaker, S
82	Kim, J *	65	Nass, C
81	Sears, A	65	Scholtz, J
81	Monk, A	65	Edmonds, E
80	Kraut, R, Kraut, B	65	Mackay, W
80	Druin, A	65	Bederson, B
79	Croft, W	64	Malone, T *
79	Rodden, T	64	Moran, T
79	Johnson, P	64	Zhai, S
77	Wickens, C	63	Tscheligi, M
77	Chen, C *	63	Olsen, D
76	Brewster, S	62	Feiner, S
76	Czerwinski, M	62	Lewis, C *
76	Smith, M *	60	Foley, J

## Appendix B: Site Survey Fields

---

1. First Name
2. Last Name
3. Number of publications in HCI bibliography
4. URL
  - 4.1. URL Notes
5. Contact Information
  - 5.1. Directions to the researcher's office
6. Type of Page (Academic, Corporate, Institutional, Personal)
7. How much content is there? (Full, Brief)
8. Is the site a multi-page site?
9. Type/Content Notes
10. Design
  - 10.1. Type of design
  - 10.2. Presence of navigation bar
  - 10.3. Navigation bar items
11. Bibliography
  - 11.1. Bibliography distance from front page (in number of links)
  - 11.2. Bibliography distance notes
  - 11.3. How full is the bibliography? (Full, Selected, None)
  - 11.4. How is the bibliography organized? (sorting method: title, author, chronologically, publication venue, publication type, is it grouped?)
  - 11.5. Bibliography organization notes
  - 11.6. Is the bibliography hosted locally or in a remote database?
  - 11.7. What serves the bibliography? (plain HTML, database, content management system)
  - 11.8. Are there links in the bibliographical entry to the full text
  - 11.9. Are the links to the full text local or remote?
12. Project Pages

- 12.1. Where are the project pages located?
- 12.2. Do the project pages have related resources and files?
- 12.3. Do the project pages have related publications?
- 13. External Content
  - 13.1. Information about awards
  - 13.2. Links to external content
  - 13.3. Third party information about the researcher
- 14. Personal Content
  - 14.1. Photos related to professional work
  - 14.2. Photos not related to professional work
  - 14.3. Blog
  - 14.4. News page
  - 14.5. Biography
  - 14.6. Non-research related files
  - 14.7. Level of personal content revelation
- 15. Teaching
  - 15.1. Links to students' webpages
  - 15.2. Teaching resources like course pages
  - 15.3. Notes about teaching resources

# Appendix C: Interview Questions

---

1. Interview Metadata
  - 1.1. Subject
  - 1.2. Date
  - 1.3. Format
  - 1.4. Associated Files
  - 1.5. Notes
2. General
  - 2.1. Overall Notes
  - 2.2. Initial Questions to Invite Discussion
    - 2.2.1. Why do you keep a website?
    - 2.2.2. What do you get out of it? What have you gotten out of it?
    - 2.2.3. Tell me about your website?
  - 2.3. Who primarily maintains your website?
  - 2.4. How substantive would you describe the effort to maintain your website?
    - 2.4.1. Get a per week / per month estimate of time and resources invested in the maintenance
    - 2.4.2. Identify specific areas of frustration that the interviewee has experienced
  - 2.5. Would you be willing to allow a third party to maintain / organize / archive your website?
    - 2.5.1. What kind of third party would you trust to do this?
    - 2.5.2. If only parts of the site, find out which parts.
  - 2.6. What would make it easier to maintain your website?
  - 2.7. Have you ever tried to find out who links to your website?
    - 2.7.1. Is this something that would matter to you?
  - 2.8. Do you participate on any social networks?
    - 2.8.1. If so, which ones?
    - 2.8.2. What are your goals for participating in a social network?

2.8.3. Would you (if they haven't) put a link on your website to identify yourself on those social networks?

### 3. Finding It later

3.1. How often do you refer to your own site to access information?

3.1.1. If uncertain, prompt:

3.1.1.1. Every day

3.1.1.2. 3-4 times a week

3.1.1.3. Once a week

3.1.1.4. Once a month

3.1.1.5. Less than once a month

3.2. Do you have a section of your website for storing information only useful for yourself? If yes:

3.2.1. Is this section temporary? That is, do you clean it out on a regular basis?

3.2.2. Is this section private? Do you protect it with passwords or other security measures?

3.3. What kinds of things do you keep on your site just for your own information purposes?

3.4. How difficult is it to find things on your website?

3.5. Do you ever resort to other kinds of file sharing methods (e-mail, file sharing websites, IM file sending) because it's easier than putting it on your site?

3.6. Do you maintain a list of links to other websites you find useful?

3.7. How often do you cull the list for dead links?

### 4. Building a Legacy

4.1. What section of your website do you spend the most time cultivating for public consumption? Probe:

4.1.1. Find out why the section takes the most time

4.1.2. Find out why they view that section as the one that is most important for public consumption purposes

4.2. Tell me about your approach to maintaining your publications

- 4.2.1. Find out how the organize the publications and why they chose that method (convention, required, personal preference)
- 4.2.2. If they include full text in their listings
  - 4.2.2.1. Find out if there are any limitations to what they can include
  - 4.2.2.2. Find out if they've seek permission from the publishers or if they simply abide with the stated policies
- 4.2.3. If they do not include full text in their listings
  - 4.2.3.1. Find out if it's a personal preference or because of publisher limitations or other similar concerns
  - 4.2.3.2. Find out if they had permission to, would they?
- 4.2.4. Do you ever put a pre-print up on your site?
  - 4.2.4.1. Do you replace it with the final version or do you leave them side by side?
- 4.3. Have you ever made research data available on your site?
  - 4.3.1. Why or why not?
  - 4.3.2. Would you ever consider doing so on a permanent basis?
- 4.4. Would you make unpublished materials available on your site that expands on your published work?
- 5. Sharing Resources
  - 5.1. Do you use your website to share resources with colleagues? students? the public?
    - 5.1.1. If yes, what kind of resources do you share?
    - 5.1.2. Have you ever sought permission to share the resource or do you only share things that are freely available?
    - 5.1.3. For things you share that are hosted elsewhere, do you just share the link or do you grab a copy and place it on your site?
  - 5.2. If you're involved in a collaborative project, do you make a project page on your personal site for the project?
    - 5.2.1. If yes, do you take the project page down after the project is completed?

- 5.2.2. For projects with an external project site, have you ever requested a copy of the site to include with your personal site after the project is completed?
    - 5.2.2.1. Would you do that in the future?
  - 5.3. When you're teaching a course [for those who teach], do you make a course page on your personal site for the course?
    - 5.3.1. What kinds of things do you make available on the course page?
    - 5.3.2. Do you take the course page down after the course is completed?
  - 5.4. When you're collaborating with others, have you ever given someone else permission to maintain a portion of your site for the purposes of collaboration?
- 6. Fears of Loss
  - 6.1. Do you backup your entire website? Portions of it or specific content?
    - 6.1.1. How do you backup your site / content?
    - 6.1.2. Where do you store the backups?
  - 6.2. Have you ever replaced an older version of a file with a new version simply because the file format was out of date?
  - 6.3. Do you think of your website as a means of archiving important documents? Why or why not?
  - 6.4. Have you ever experienced a catastrophic loss?
    - 6.4.1. On your personal computers?
    - 6.4.2. Of your website?
    - 6.4.3. On institutional / third party servers?
  - 6.5. Do you save previous versions of the website?
    - 6.5.1. If so, do you live them online somewhere?
- 7. Identity Construction
  - 7.1. Use of Institutional Templates
    - 7.1.1. For those who follow an institutional template
      - 7.1.1.1. Did you have a choice in the use of the institutional template?



- 7.1.1.1.1. Yes: Why did you choose to use the template? (probe for a couple reasons)
- 7.1.1.1.2. No: If you had a choice, would you have done your own design? Why or why not?
- 7.1.2. For those who use their own design
  - 7.1.2.1. Did you have an option of an institutional template?
    - 7.1.2.1.1. Yes: Why did you choose to use your own design
    - 7.1.2.1.2. How much effort was invested in creating the current design?
- 7.2. How often do you change the design of your website?
  - 7.2.1. What motivates you to change the design of your website?
- 7.3. Do you think your website represents who you are as a researcher?
  - 7.3.1. What things did you specifically include to personalize the website?
  - 7.3.2. What aspects of the website do not reflect who you are? Why?

# Appendix D: Ethics Approvals

---



UNIVERSITY OF  
CALGARY

## MEMO

CONJOINT FACULTIES RESEARCH ETHICS BOARD  
c/o Research Services  
Main Floor, Energy Resources Research Building  
3512 - 33 Street N.W., Calgary, Alberta T2L 1Y7  
Telephone: (403) 220-3782  
Fax: (403) 289 0693  
Email: [rburrows@ucalgary.ca](mailto:rburrows@ucalgary.ca)  
Wednesday, April 22, 2009

**To: Timothy Au Yeung**  
Computer Science

**From:** Dr. Kathleen Oberle, Acting Chair  
Conjoint Faculties Research Ethics Board (CFREB)

**Re: Certification of Institutional Ethics Review:** Preserving the Individual Within the Institutional:  
Exploring the Gap between Personal Digital Archives of Researchers and Institutional Digital Preservation  
Efforts

The above named research protocol has been granted ethical approval by the Conjoint Faculties Research Ethics Board for the University of Calgary.

Enclosed are the original, and one copy, of a signed **Certification of Institutional Ethics Review**. Please make note of the conditions stated on the Certification. A copy has been sent to your supervisor as well as to the Chair of your Department/Faculty Research Ethics Committee. In the event the research is funded, you should notify the sponsor of the research and provide them with a copy for their records. The Conjoint Faculties Research Ethics Board will retain a copy of the clearance on your file.

Please note, an annual/progress/final report must be filed with the CFREB twelve months from the date on your ethics clearance. A form for this purpose has been created, and may be found on the "Ethics" website, <http://www.ucalgary.ca/research/compliance/ethics/renewal>

In closing let me take this opportunity to wish you the best of luck in your research endeavor.

Sincerely,

A handwritten signature in black ink, appearing to read 'Russell Burrows'.

Russell Burrows  
For:  
Kathleen Oberle, Ph.D., Faculty of Nursing and  
Acting Chair, Conjoint Faculties Research Ethics Board

Enclosures(2)  
cc: Chair, Department/Faculty Research Ethics Committee  
Supervisor: Saul Greenberg



UNIVERSITY OF  
CALGARY

MEMO

**Conjoint Faculties Research Ethics Board (CFREB)**  
Research Services Office  
Main Floor, Energy Resources Research Building  
Research Park  
Telephone: (403) 220-3782 or (403) 210-9863  
Fax: (403) 289-0693  
Email: [csjahrau@ucalgary.ca](mailto:csjahrau@ucalgary.ca) or [rburrows@ucalgary.ca](mailto:rburrows@ucalgary.ca)

**To:** Timothy Au Yeung  
Department of Computer Science

**Date:** April 15, 2010

**From:** Dr. Kathleen Oberle, Chair  
Conjoint Faculties Research Ethics Board

**Re:** Approval of Modification for: Preserving the Individual Within the Institutional: Exploring the Gap between Personal Digital Archives of Researchers and Institutional Digital Preservation Efforts  
Original Approval Date: April 22<sup>nd</sup>, 2009  
File No: 6033

The Certificate of Institutional Ethics Review issued on April 22<sup>nd</sup>, 2009 continues in force and extends to the modifications as set out in your email/memo dated April 6<sup>th</sup> and April 13<sup>th</sup> 2010. Your request to i) offer participants a choice in terms of whether they wish to be quoted and referenced either by participant number or by name, ii) associate participants quotes with snapshots of their website when they do choose to be named, and iii) use a consent form specifically approved for this purpose, is approved, as described.

You should attach a copy of the documentation you provided in order to request the modification, together with a copy of this memorandum, to the original Certification in your files.

Sincerely,

Kathleen Oberle, Ph.D.  
Chair, Conjoint Faculties Research Ethics Board

Cc: Dr. Saul Greenberg

## Appendix E: Consent Form

---



---

**Name of Researcher, Faculty, Department, Telephone & Email:**

Tim Au Yeung, Faculty of Science / Department of Computer Science, 403-220-8975, [ytau@ucalgary.ca](mailto:ytau@ucalgary.ca)

**Supervisor:**

Saul Greenberg, Faculty of Science / Department of Computer Science, 403-220-6087, [saul.greenberg@ucalgary.ca](mailto:saul.greenberg@ucalgary.ca)

**Title of Project:**

Impact of Creator Dissemination on Digital Preservation

**Sponsor:**

GroupLab, Department of Computer Science, University of Calgary

---

This consent form, a copy of which has been given to you, is only part of the process of informed consent. If you want more details about something mentioned here, or information not included here, you should feel free to ask. Please take the time to read this carefully and to understand any accompanying information.

The University of Calgary Conjoint Faculties Research Ethics Board has approved this research study.

**Purpose of the Study:**

Researchers in the GroupLab group in the Department of Computer Science at the University of Calgary, lead by Dr. Saul Greenberg are conducting a project on how scholars create personal online archives of their work in digital form. Among the issues that scholars are encountering is that as the creation and dissemination of scholarly output shifts from analog to digital, the traditional approaches to organizing and archiving their research materials is growing more complicated. Citations may not point to the same information six months later, file formats and media change so that old data sets can become unreadable in as short as a couple years. One of the most challenging aspects is the scholarly identity that the scholar presents to the outside world is getting lost as all of their work is being poured into giant databases that strip out any sense of evolution in thinking or changes in focus. This study looks at scholars' attitudes towards how their identity is projected into the online world through their publications and other materials made public and how the contextualization of the publications and related materials impacts how to preserve both the individual components of research as well as the greater context.

**What Will I Be Asked To Do?**

You will be interviewed either face to face or through telephone / videoconference about your approach to constructing your online scholarly identity through your personal scholarly website and through the presentation of your publication record. If you have an existing scholarly website or web identity (as part of an aggregator, social network or portal), you may be asked to describe the decisions made in terms of how things are presented and why they are presented that way. This interview should take between 1 to 2 hours and will be recorded for data gathering purposes only. The recordings will be audio only and done only to ensure that your statements are accurately recorded and representative of the discussion.

You may withdraw from participation at any time – if you do so, only information collected up to the point of withdrawal will be used as part of the study and only in anonymous form unless permission is given in this form.

**What Type of Personal Information Will Be Collected?**

There are several options for you to consider if you decide to take part in this research. You can choose all, some or none of them. Please put a check mark on the corresponding line(s) that grants me your permission to:

You may quote me from my interview in written form (using only the participant number): Yes: \_\_\_\_ No: \_\_\_\_

You may quote me from my interview in written form (using my name):

Yes: \_\_\_\_ No: \_\_\_\_

The reason we ask to identify you by name would be to add context by associating the quote with a snapshot of your website.

### **Are there Risks or Benefits if I Participate?**

There should be no risks that arise from participation in this study. If we discuss your website as part of a publication, there will be no association between it and any quotes or information that you provide unless permission is granted above to be quoted and referenced by name.

### **What Happens to the Information I Provide?**

Participation is completely voluntary and confidential. You are free to discontinue participation at any time during the study. No one except the researchers and his supervisor will be allowed to see or hear any of the answers to the interview or the audio tape. The interview and audio tape are kept in a locked cabinet only accessible by the researcher and his supervisor. Recordings will be kept for five (5) years after which point they will be destroyed. The results of the interview will be utilized in a Masters thesis and may also be used for academic presentations, and publication purposes such as academic journals and conferences.

---

### ***Signatures (written consent)***

Your signature on this form indicates that you 1) understand to your satisfaction the information provided to you about your participation in this research project, and 2) agree to participate as a research subject.

In no way does this waive your legal rights nor release the investigators, sponsors, or involved institutions from their legal and professional responsibilities. You are free to withdraw from this research project at any time. You should feel free to ask for clarification or new information throughout your participation.

Participant's Name: (please print) \_\_\_\_\_

Participant's Signature \_\_\_\_\_ Date: \_\_\_\_\_

Researcher's Name: (please print) \_\_\_\_\_

Researcher's Signature: \_\_\_\_\_ Date: \_\_\_\_\_

### **Questions/Concerns**

If you have any further questions or want clarification regarding this research and/or your participation, please contact:

Mr. Tim Au Yeung  
Department of Computer Science  
403-220-8975, ytau@ucalgary.ca

And

Dr. Saul Greenberg  
Department of Computer Science  
403-220-6087, [saül.greenberg@ucalgary.ca](mailto:saül.greenberg@ucalgary.ca)

If you have any concerns about the way you've been treated as a participant, please contact Bonnie Scherrer, Ethics Resource Officer, Research Services Office, University of Calgary at (403) 220-3782; email [bonnie.scherrer@ucalgary.ca](mailto:bonnie.scherrer@ucalgary.ca).

A copy of this consent form has been given to you to keep for your records and reference. The investigator has kept a copy of the consent form.

## Appendix F: List of Websites

Researcher	URL
Mark Weiser	<a href="http://sandbox.parc.com/weiser/">http://sandbox.parc.com/weiser/</a>
Doug Engelbart	<a href="http://dougengelbart.org/">http://dougengelbart.org/</a>
Peter Johnson	<a href="http://www.cs.bath.ac.uk/~pj/">http://www.cs.bath.ac.uk/~pj/</a>
Carl Gutwin	Personal: <a href="http://www.cs.usask.ca/~gutwin/">http://www.cs.usask.ca/~gutwin/</a> Lab: <a href="http://hci.usask.ca/">http://hci.usask.ca/</a>
Alan Dix	<a href="http://www.comp.lancs.ac.uk/~dixa/">http://www.comp.lancs.ac.uk/~dixa/</a>
Michael Muller	<a href="http://domino.research.ibm.com/cambridge/research.nsf/pages/michael_muller.html">http://domino.research.ibm.com/cambridge/research.nsf/pages/michael_muller.html</a>
Ben Shneiderman	<a href="http://www.cs.umd.edu/~ben/">http://www.cs.umd.edu/~ben/</a>
Saul Greenberg	<a href="http://pages.cpsc.ucalgary.ca/~saul/wiki/pmwiki.php">http://pages.cpsc.ucalgary.ca/~saul/wiki/pmwiki.php</a>
Ravin Balakrishnan	<a href="http://www.dgp.toronto.edu/~ravin/">http://www.dgp.toronto.edu/~ravin/</a>
Ben Bederson	<a href="http://www.cs.umd.edu/~bederson/">http://www.cs.umd.edu/~bederson/</a>
Jonathan Grudin	At Microsoft: <a href="http://research.microsoft.com/en-us/um/people/jgrudin/">http://research.microsoft.com/en-us/um/people/jgrudin/</a> Prior to Microsoft: <a href="http://research.microsoft.com/en-us/um/people/jgrudin/past/index.html">http://research.microsoft.com/en-us/um/people/jgrudin/past/index.html</a>
MIT DSpace	<a href="http://dspace.mit.edu/">http://dspace.mit.edu/</a>
University of Toronto DSpace	<a href="https://tspace.library.utoronto.ca/index.jsp">https://tspace.library.utoronto.ca/index.jsp</a>
University of Calgary DSpace	<a href="https://dspace.ucalgary.ca/">https://dspace.ucalgary.ca/</a>