

A GROUNDED THEORY OF INFORMATION QUALITY IN WEB ARCHIVES

Brenda Reyes

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

August 2018

APPROVED:

Jiangping Chen, Major Professor  
Oksana Zavalina, Major Professor  
Shawne Miksa, Committee Member  
Kathryn Masten-Cain, Committee Member  
Yunfei Du, Interim Chair of the  
Department of Information Science  
Kinshuk, Dean of the College of  
Information  
Victor Prybutok, Dean of the Toulouse  
Graduate School

Reyes, Brenda. *A Grounded Theory of Information Quality in Web Archives*. Doctor of Philosophy (Information Science), August 2018, 214 pp., 36 tables, 28 figures, references, 67 titles.

Web archiving is the practice of preserving websites as a historical record. It is a technologically-challenging endeavor that has as its goal the creation of a high-quality archived website that looks and behaves exactly like the original website. Despite the importance of the notion of quality, comprehensive definitions of Information Quality (IQ) in a web archive have yet to be developed. Currently, the field has no single, comprehensive theory that describes what is a high-quality or low-quality archived website. Furthermore, most of the research that has been conducted on web archives has been system-centered and not user-centered, leading to a dearth of information on how humans perceive web archives. This dissertation seeks to remedy this problem by presenting a user-centered grounded theory of IQ for web archives. It answers two research questions: 1) What is the definition of information quality (IQ) for web archives? and 2) How can IQ in a web archive be measured? The theory presented is grounded on data obtained from users of the Internet Archive's Archive-It system, the largest web-archiving subscription service in the United States. Also presented are mathematical definitions for each dimension of IQ, which can then be applied to measure the quality of a web archive.

Copyright 2018

by

Brenda Reyes

## ACKNOWLEDGMENTS

I would like to thank the members of my doctoral committee, who provided me with guidance and support throughout this long process: Dr. Jiangping Chen, Dr. Oksana Zavalina, Dr. Shawne Miksa, and Dr. Kathryn Masten-Cain.

I would also like to thank my family, without whom this dissertation would not be possible: my parents Santiago and Elba, and my brothers Joel and William. My husband Victor was a tirelessly encouraging presence since the beginning. And of course, to you, Yoshi, who were only here for a short while but provided so much light and love.

## TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS . . . . .	iii
LIST OF TABLES . . . . .	viii
LIST OF FIGURES . . . . .	x
CHAPTER 1. INTRODUCTION . . . . .	1
1.1. Statement of the Problem . . . . .	1
1.2. Goals and Research Questions . . . . .	2
CHAPTER 2. LITERATURE REVIEW . . . . .	4
2.1. Origins and Evolution of Web Archiving . . . . .	4
2.2. Disciplinary Connections . . . . .	7
2.2.1. Information Retrieval . . . . .	8
2.2.2. Web History/Historiography . . . . .	8
2.3. Web Archiving: a Problematic Term? . . . . .	11
2.4. The Concept of Information Quality: Paradigms, Theories, and Models . . . . .	12
2.4.1. Models of IQ in Information Science . . . . .	12
2.4.2. IQ in Philosophy . . . . .	31
2.4.3. IQ in Computer Science . . . . .	39
2.4.4. IQ in Other Fields . . . . .	42
2.5. IQ in Web Archiving . . . . .	45
2.5.1. Types and Severity of Quality Problems . . . . .	45
2.5.2. The Quality Assurance Process in Web Archiving . . . . .	49
2.5.3. Research on Quality in Web Archives . . . . .	51
CHAPTER 3. METHODOLOGY . . . . .	83

3.1.	The Grounded Theory of Glaser and Strauss . . . . .	83
3.1.1.	The Evolution of GT . . . . .	84
3.1.2.	Key Ideas, Processes, and Techniques . . . . .	84
3.1.3.	Grounded Theory vs. Logico-Formal theory . . . . .	90
3.2.	Lazarsfeld's Qualitative Mathematics . . . . .	91
3.3.	Process . . . . .	96
3.3.1.	Pilot study: Developing a Preliminary Model of IQ in Web Archives . . . . .	96
3.3.2.	Phase 1: Building a Substantive Theory of Quality in a Web Archive . . . . .	98
3.3.3.	Phase 2: Identifying How to Operationalize Dimensions of Web Archive Quality . . . . .	103
3.3.4.	Auditing the Dissertation Work . . . . .	105
3.3.5.	Timeline of the Dissertation . . . . .	105
CHAPTER 4.	FINDINGS: A GROUNDED THEORY OF INFORMATION QUALITY FOR WEB ARCHIVES . . . . .	108
4.1.	AIT Clients, their Roles, and Characteristics . . . . .	108
4.2.	Core Categories . . . . .	109
4.2.1.	Veering Away from Capture and Replay Issues as Categories . . . . .	111
4.2.2.	The Dimension of Correspondence . . . . .	113
4.2.3.	The Dimension of Relevance . . . . .	122
4.2.4.	The Dimension of Archivability . . . . .	127
CHAPTER 5.	FINDINGS: OPERATIONALIZING INFORMATION QUALITY FOR WEB ARCHIVES . . . . .	134
5.1.	Defining the Universe of Web Archiving . . . . .	134
5.2.	Operationalizing Correspondence . . . . .	136
5.2.1.	Operationalizing Visual Correspondence . . . . .	136
5.2.2.	Operationalizing Interactional Correspondence . . . . .	137

5.2.3. Operationalizing Completeness . . . . .	147
5.3. Operationalizing Relevance . . . . .	152
5.3.1. Topic Relevance . . . . .	154
5.3.2. Size Relevance . . . . .	160
5.4. Operationalizing Archivability . . . . .	162
CHAPTER 6. DISCUSSION AND CONCLUSION . . . . .	164
6.1. Research Questions . . . . .	164
6.1.1. RQ1: What is the Human-Centered Definition of Information Quality (IQ) for Web Archives? . . . . .	164
6.1.2. RQ2: How Can IQ in a Web Archive be Measured? . . . . .	165
6.1.3. Unexpected or Surprising Findings . . . . .	166
6.2. Limitations of the Study . . . . .	168
6.2.1. Methodological Issues . . . . .	168
6.2.2. Scope and Limitations . . . . .	169
6.3. Contributions of the Study . . . . .	171
6.4. Future Directions . . . . .	172
6.4.1. Applying the Operationalized Definitions of IQ . . . . .	172
6.4.2. Other Research Directions . . . . .	175
APPENDIX A. RESEARCHER AGREEMENT WITH ARCHIVE-IT . . . . .	177
APPENDIX B. A SAMPLE ARCHIVE-IT SUPPORT TICKET WITH XML TAGS . . . . .	179
APPENDIX C. ANONYMIZING THE SUPPORT TICKETS . . . . .	188
APPENDIX D. NVIVO CODEBOOK . . . . .	191
APPENDIX E. COMPLETENESS IN AN ARCHIVED WEBSITE USING THE JACCARD SIMILARITY . . . . .	202

APPENDIX F. MEASURING TOPIC RELEVANCE USING COSINE SIMILARITY . .	205
REFERENCES . . . . .	208



## LIST OF TABLES

	Page
Table 2.1. Rieh’s Facets of Information Quality . . . . .	15
Table 2.2. Bruce and Hillman’s Quality Measures . . . . .	17
Table 2.3. Stvilia’s Dimensions of Quality . . . . .	22
Table 2.4. Stvilia’s List of IQ Metrics . . . . .	25
Table 2.5. Quality Metrics and their Operational Definitions . . . . .	39
Table 2.6. Pattern Groups and their Coherence States . . . . .	57
Table 2.7. Pattern Groups, their Content Patterns, and their Coherence States . . . . .	60
Table 2.8. Facets of Archivability . . . . .	70
Table 2.9. The Results of Different VQI Comparisons and their Outcomes . . . . .	79
Table 3.1. Differences Between Grounded Theory and Traditional Approaches . . . . .	91
Table 3.2. Panel Analysis . . . . .	95
Table 3.3. Number of Tickets and Interactions About Information Quality Analyzed Per Year . . . . .	101
Table 3.4. Sample Panel Analysis . . . . .	104
Table 3.5. Timeline of Phase 1 of the Dissertation . . . . .	106
Table 3.6. Timeline of Phase 2 of the Dissertation . . . . .	107
Table 4.1. Differences Between AIT Clients and End Users of Web Archives . . . . .	109
Table 4.2. Dimensions of Information Quality in a Web Archive and their Frequencies	113
Table 4.3. Examples of Explicit Comparison to the Original Website Using Links . . . . .	115
Table 4.4. Examples of Explicit Comparison to the Original Website Without Using Links . . . . .	116
Table 4.5. Examples of Problems with Visual Correspondence . . . . .	118
Table 4.6. Examples of Problems with Interactional Correspondence . . . . .	119
Table 4.7. Examples of Problems with Completeness . . . . .	120

Table 4.8. Examples of Completeness Problems Caused by Robots Exclusions . . .	121
Table 4.9. Examples of Topic Relevance Problems . . . . .	123
Table 4.10. Examples of Seemingly Irrelevant Content that is Actually Relevant . . .	124
Table 4.11. Examples of General Size Relevance Problems . . . . .	125
Table 4.12. Examples of Size Relevance Problems Caused by Crawler Traps . . . . .	126
Table 4.13. Examples of Size Relevance Problems Caused by Problematic Content .	128
Table 4.14. Examples of Archivability Problems Caused by Websites Changing How it Delivers Content to Users . . . . .	129
Table 4.15. Examples of Archivability Problems Caused by Websites with Dynamic Content . . . . .	131
Table 4.16. Examples of Archivability Problems Caused by Websites Rendering Con- tent in Unique Ways . . . . .	133
Table 5.1. Sample of Network Requests from Library of Congress Website . . . . .	140
Table 5.2. Errors with the Archived Version of the UNT Campus Map . . . . .	143
Table 5.3. Summary of the Types of Web Archiving Functions and their Quality . .	150
Table 6.1. Dimensions of Information Quality in a Web Archive and their Corre- sponding Measures . . . . .	166
Table 6.2. Dimensions of IQ and the Levels to Which They are Best Applied . . . .	170

## LIST OF FIGURES

	Page
Figure 2.1. The general problem of quality composition . . . . .	38
Figure 2.2. The larger-the-better (L type) curve . . . . .	43
Figure 2.3. The smaller-the-better (S type) curve . . . . .	44
Figure 2.4. The nominal-the-best (N type) curve . . . . .	45
Figure 2.5. A good-quality archived version of the UNT Athletics site from 2007 . . .	46
Figure 2.6. An archived version of the UNT Admissions site from 2007, missing the styling of the original . . . . .	47
Figure 2.7. An archived version of the UNT Campus Map from 2004, almost unusable	48
Figure 2.8. An archived version of the CNN.com website, showing leakage . . . . .	49
Figure 2.9. An example of web archive incoherence . . . . .	55
Figure 2.10. How visual quality is calculated at the Swiss National Library. . . . .	81
Figure 3.1. The generalizability of substantive and formal theories . . . . .	87
Figure 3.2. Relationship between the predictive power of a theory and the degree of difference between groups . . . . .	87
Figure 4.1. The theory of IQ for web archives, visualized . . . . .	113
Figure 5.1. The function $f : O \rightarrow A$ maps the original website to the archived website	135
Figure 5.2. The Library of Congress website on January 2018 . . . . .	138
Figure 5.3. Network request for the Library of Congress website . . . . .	139
Figure 5.4. The campus map of the University of North Texas on January 2018 . . .	141
Figure 5.5. An archived version of the UNT Campus Map, missing its interactive features . . . . .	142
Figure 5.6. The bijective function $f : O \rightarrow A$ . . . . .	148
Figure 5.7. The injective function $g : O \rightarrow A$ . . . . .	148
Figure 5.8. The surjective function $h : O \rightarrow A$ . . . . .	149

Figure 5.9. Sample graph . . . . . 155

Figure 5.10. Site hierarchy for the Library of Congress website . . . . . 156

Figure 5.11. Link graph of a website . . . . . 156

Figure 5.12. Graph of topic relevance as a function of distance . . . . . 157

Figure 5.13. External links for the Library of Congress website . . . . . 158

Figure 5.14. The first condition for size relevance . . . . . 161

Figure 5.15. The second condition for size relevance . . . . . 161

## CHAPTER 1

### INTRODUCTION

In 1996, a small, non-profit organization called the Internet Archive was founded in San Francisco with the ambitious goal of building a universally accessible digital library. The Internet Archive began using a then-new technology known as a web crawler to periodically take snapshots of websites and store them in massive storage warehouses. Internet users could then access these archived websites using the Wayback Machine, a special piece of software developed by the Internet Archive. As the World Wide Web evolved, the pace at which websites changed their content and appearance accelerated dramatically: websites were redesigned or disappeared altogether, additional materials such as video and audio were added, and social media began to emerge. Often the Internet Archive's cache was the only record of how a website had evolved or that it had existed at all. By the dawn of the new millennium, the practice of "web archiving," as it became known, had spread beyond the Internet Archive. Organizations such as national libraries, government organizations, and universities began also to archive websites, for the purpose of preserving their digital heritage.

Though enormous strides have been made, web archiving today remains a complicated and technically-challenging endeavor. New web technologies emerge constantly, and web archivists struggle to keep up. Creating an archived website that is as close as possible to the original, live website remains one of the most difficult challenges in the field. Failing to adequately capture a website might mean an incomplete historical record or worse, no evidence that the site ever even existed. It is in the context of these challenges that this research takes place.

#### 1.1. Statement of the Problem

In the field of Web Archiving, there has been only one definition of Information Quality (IQ) in a web archive, put forward by Masanès (2006). He defined quality in a web archive

as having the following characteristics:

- (1) the completeness of material (linked files) archived within a target perimeter
- (2) the ability to render the original form of the site, particularly regarding navigation and interaction with the user (Masanès, 2006, p. 39)

This definition of quality is problematic because it is too centered on the technological tools needed to archive websites. Terms such as "target perimeter" refer to the configuration of web crawlers. If the web archive was created using alternative methods or if crawlers were replaced in the future by newer, more efficient tools, then Masanés' definition would become obsolete. Another problem is that it lacks a human element; one never finds out what quality might mean to the users and creators of web archives. This definition ignores the context in which a web archive exists and whether or not it meets the needs of its users.

Clearly a more robust definition of IQ in web archives is needed, one that is both independent of the technology currently in use to create web archives and that incorporates a human element. The goal here is to create a theory of IQ that counteracts the weaknesses of Masanés' original definition by being more abstract and grounded in research with actual users and creators of web archives.

The lack of a proper definition of quality is indicative of a larger problem in the field of web archiving. The technical developments in the field have far outpaced the development of proper theoretical tools or models. Over two decades into its history, web archiving still lacks a theoretical underpinning. Essentially, we have technological tools to build web archives, but no conceptual tools to understand them.

## 1.2. Goals and Research Questions

The goal of this research is to build a theory of IQ for web archives that is grounded in user-centered empirical data and in research with users. This goal leads to the following research questions:

**RQ 1:** What is the human-centered definition of information quality (IQ) for web archives?

**RQ 2:** How can IQ in a web archive be measured?

By clarifying the notion of quality for web archives, the resulting theory will begin the work of establishing a much-needed theoretical groundwork for the field. Practitioners in the field of web archiving will also benefit from this theory and its accompanying operational definitions of quality. Knowing which aspects of web archive quality can be measured will allow web archiving professionals to improve the Quality Assurance processes for their organizations.

## CHAPTER 2

### LITERATURE REVIEW

#### 2.1. Origins and Evolution of Web Archiving

Web archiving as a field is only a few years older than the web itself, and can be said to have arisen due to fears that digital information would not withstand the passage of time, and thus important pieces of the historical record would eventually be lost. In a conference for the International Federation of Library Associations and Institutions, Terry Kuny (1997) referred to the future as a “Digital Dark Age,” an “era where much of what we know today, much of what is coded and written electronically, will be lost forever” (Kuny, 1997, p. 1). He warned that enormous amounts of digital information had already been lost and that “Digital history [could not] be recreated by individuals and organizations [could not] recreate a digital history” (Kuny, 1997, p. 2).

Kuny’s notion of a Digital Dark Age proved popular and enduring, and continues to inform many scholars’ opinions about the dangers of losing digital information. In an interview in 2010, researcher Kirsten Foot expressed a similar opinion about the web:

There is a significant collective consciousness that is heading to a dark ages where we aren’t writing anything down; in fact, we are writing lots down on the web, but then we are writing over what we just wrote. It will be very hard for future scholars even in 5 years, 10 years to understand what kinds of political and social and cultural moments or phenomena retrospectively without key aspects of the web.

(Foot as cited in Dougherty and Meyer, 2014, p.2196)

The loss of the digital historical record can be seen to have technological causes. Obsolescence is a major contributor; file formats become obsolete when the software necessary to read them disappears or stops being produced. In many cases, the file format might still



be read, but the media on which the information is stored becomes unreadable. This was the case with floppy disks and ZIP drives, where the hardware to read them is no longer a standard feature on computers. However, some authors have seen the loss of the digital historical record as a problem of culture rather than technology. In their analysis Foot and Schneider identify two factors that contribute to this loss: a “technological determinist ideology” and “historical amnesia”: “far more of the web is over-written, erased, or deleted than captured, due to the dominant ideology of perpetual technological innovation and the widespread cultural impulses to revise or forget (web history)” (Foot & Schneider, 2010, p. 67).

By the time Kuny expressed concern about a Digital Dark Age, efforts had already begun to address the problem. A year earlier in 1996, Brewster Kahle founded the Internet Archive with the mission of creating a universally accessible digital library. As its website states, the Internet Archive works “to prevent the Internet - a new medium with major historical significance - and other ‘born-digital’ materials from disappearing into the past” (Internet Archive, 2012). In order to achieve this goal, the Internet Archive periodically takes snapshots of the entire World Wide Web and stores copies of the captured sites in massive storage warehouses. Users can then access these websites using the Wayback Machine, a special piece of software developed by the Internet Archive. Due to its size and scope, the Internet Archive has remained the largest and most ambitious web archiving project to date.

Many organizations soon followed in the footsteps of the Internet Archive by launching web archiving programs of their own. In 1996, the National Library of Australia inaugurated the first-ever web archiving program by a national library, an effort to capture portions of Australia’s national domain (.au) as well as other web resources deemed significant to Australia’s history and culture. In 2000, the Library of Congress began its Minerva Project (now the Library of Congress Web Archives), a web archiving effort aimed at creating topical collections of materials relevant to American history and culture. Many national libraries soon

followed suit, such as the British Library, which began its UK Web Archive in 2004 and the National Library of France in 2006. Over the years, many universities have also begun web archiving projects (International Internet Preservation Consortium, 2016).

In 2003, the International Internet Preservation Consortium (IIPC) was founded with the mission of “improving the tools, standards and best practices of web archiving while promoting international collaboration, broad access and use of web archives for research and cultural heritage” (International Internet Preservation Consortium, n.d.). It has remained one of the few organizations addressing the needs of the web archiving field. The founding of the IIPC also coincided with a greater awareness of the importance of digital content and its relationship to cultural heritage. That same year, the United Nations Educational, Scientific and Cultural Organization (UNESCO) established its Charter on the Preservation of the Digital Heritage. In it, UNESCO recognized that “[cultural] resources of information and creative expression are increasingly produced, distributed, accessed and maintained in digital form, creating a new legacy - the digital heritage” (United National Educational Scientific and Cultural Organization, 2003, p. 1). UNESCO acknowledged that the world’s digital heritage was at risk of being lost and noted that “its preservation for the benefit of present and future generations is an urgent issue of worldwide concern” (United National Educational Scientific and Cultural Organization, 2003, p. 1). It postulated that efforts should be undertaken to prevent the loss of the digital cultural heritage: “Member States may wish to cooperate with relevant organizations and institutions in encouraging a legal and practical environment which will maximize accessibility of the digital heritage” (United National Educational Scientific and Cultural Organization, 2003, p. 2)

In 2016, the National Digital Stewardship Alliance (NDSA) conducted a survey of web archiving practices in the United States (Bailey, Grotke, McCain, Moffatt, & Taylor, 2016). They found over 100 American institutions that had web archiving programs in place, 63% were colleges and universities, 15% were federal, state, and local governments, 12% were

archives, and the rest were organizations such as private companies, historical associations, and museums (Bailey et al., 2016, p. 5). Of the respondents, 79% had formal web archiving programs, and the rest had programs that were either being planned or already in the pilot stage (Bailey et al., 2016, p. 7). The NDSA report also noted that web archiving was an activity undergoing increasing institutionalization, as evidenced by the increases in the number of active web archiving programs, the number of web archivists participating in professional groups, and a larger focus on preserving internal or institutional content (Bailey et al., 2016, p. 29).

## 2.2. Disciplinary Connections

Because of the prominent role of national libraries in the continuing evolution of the field, web archiving seems to have become the domain of libraries and archival institutions. The field is currently centered on solving very practical technical issues. Software development and collection management are at its core, with considerable effort being spent on developing and improving web crawlers and software to monitor the capture of resources and provide for their access. As Dougherty and Meyer explain:

Library and information science norms have been the basis for many developments in web archiving policy and infrastructure. The result is a strong focus on tools, an archival viewpoint, and traditional modes of collection development relying on broad notions for how web archives will eventually be used.

(Dougherty & Meyer, 2014, Discussion section, para. 1)

Despite the fact that the earliest adopters of web archiving were libraries, the field has not been truly integrated into Library and Information Science. Being a practice-oriented field, web archiving rarely or never incorporates theories or models from Information Science. The following sections describe two disciplines with important contributions to web archiving: Information Retrieval and Web History. Information Retrieval has provided the technologies

to make web archiving possible, while in the emerging area of Web History, scholars are making substantial efforts to place web archiving within a larger theoretical framework.

### 2.2.1. Information Retrieval

Technical innovations in Information Retrieval have contributed significantly to the development of web archiving as a practice. First and foremost amongst these is the software known as a crawler, which accesses a website and retrieves its content by recursively following each link. Web crawlers are commonly used by search engines to download copies of millions of pages; Google's crawler, GoogleBot, is an example of a well-known crawler.

The most popular crawler used by web archivists is called Heritrix. Developed by the Internet Archive in 2003, Heritrix is an open-source crawler that is highly customizable (Mohr, Kimpton, Stack, & Ranitovic, 2004). Heritrix has the ability to store captured websites in WebArchive (WARC) format, a file format able to store digital content into a single, highly-compressed file (ISO Technical Committee ISO/TC 46, 2009). Many institutions to date have used the Heritrix crawler to build their web archives.

The areas of search and indexing for Information Retrieval have also contributed significantly to web archiving. Once a web archive is built, it must be made accessible and searchable to users. Numerous institutions such as the Internet Archive, the Institut National de l'Audiovisuel (INA), and the Portuguese Web Archive have used popular software such as NutchWAX and Apache Lucene to provide search capabilities for their web archives (Wikipedia, n.d.).

### 2.2.2. Web History/Historiography

In the first decade of the new millennium, the web was increasingly being viewed as a valid object of study in academic circles. Slowly, researchers began to pay attention to the web and its evolution over time, giving rise to the term "web history." Though the number of "web historians" is still small, some have already made significant theoretical contributions

to the field. In a chapter in *Web History*, Brügger presents an overview of a conceptual framework that can be used in historical studies of a website.

Brügger specifically addresses “website history,” which focuses, not on the entire Web, but on a specific website as a historical artifact. Website history is a “discipline which aims at writing the history of the complex strategic situation in which the artifact is entangled” (Brügger, 2009, p. 33). This strategic situation is composed of three parts:

- (1) elements: entities which form part of either the sender, the medium, or the receiver
- (2) actors: elements that influence one or more elements(s) (not necessarily a person)
- (3) driving forces: an actor that has a “decisive” influence on the whole situation

(Brügger, 2009, p. 41)

As an example, Brügger explains the history behind the founding of [dr.dk](http://dr.dk), the official website of Denmark Radio (DR), in 1996. A history of this website would entail describing its strategic situation. For example, the World Wide Web (WWW) and Bulletin Board System (BBS) protocols could be viewed as elements of the medium (the website). Two of the actors in this situation are a) the Danish radio listeners who asked for the institution to establish an email address and b) rival broadcasters whose presence motivated the creation of the website. Driving forces in the situation include *Harddisken*, a popular radio program about new media, and *Projekt Internet*, an institutional project that sought to clarify the role of the Internet in DR’s future (Brügger, 2009, p. 48–49).

Web archiving has a logical connection to web history, as web archives constitute valuable source material for historians. By accessing archived versions of a site, a historian can examine the evolution of a website over time. For example, a historian might focus on the aesthetic changes of a website as represented by its use of color, typeface, and layout, while another historian might examine the text of an archived website to study changes in the rhetorical strategy of an institution. Despite the richness of archived material, Brügger points out that archived versions of websites are not sufficient as source material; they must

be supplemented by other sources commonly available to historians, such as news reports, interviews, and internal documentation.

In “Object-Oriented Web Historiography,” Foot and Schneider (2010) present a rigorous object-oriented framework for studying web archives that is derived from activity theory. Within this framework, the subject (a human) has a particular need that must be met. The object (in this case, the web archive) is any entity through which a particular human need is pursued. They highlight several important motives that might lead researchers to create or study a web archive:

- (1) to preserve web phenomena that they (or their institutions) find meaningful
- (2) to preserve what others prefer to erase or expunge
- (3) to make sense of socio-cultural-political relations
- (4) to understand the evolution of the web

(Foot & Schneider, 2010, p. 67)

The authors also identify different approaches to the web as an object of study. A researcher might opt to study the web using any of the following methods:

- (1) a discursive or rhetorical analysis of a website: uses content analysis and treats a website essentially as a text
- (2) structural/feature analysis of a website: focuses on the structural elements of a site, such as its hierarchy and number of pages
- (3) sociocultural analysis of a website: examines the cultural context of a site, such as its relation to other sites and the aims and strategies of the website producers

(Foot & Schneider, 2010, p. 71–72)

The sociocultural analysis described by Foot and Schneider bears a close resemblance to the conceptual framework put forward by Brügger. Both approaches highlight the “situatedness” of a website by putting it in the larger context that takes into account the many elements that contribute to the content, appearance, and structure of a website.

### 2.3. Web Archiving: a Problematic Term?

The word “archiving” in “web archiving” is problematic because it seems to have been created somewhat haphazardly, without any regard to how the archiving discipline defines and uses the term. As a result, some traditional archivists might see the use of the term “archiving” as inappropriate to the practice of preserving websites for future use.

In his now-classic *Modern Archives Principles and Techniques*, Schellenberg (1975) describes the differences between an archiving institution and a library: “Archival institutions are *receiving* agencies, whereas libraries are *collecting* agencies” (Schellenberg, 1975, p. 19). Accordingly, they have different functions: “archival institutions do not collect materials; they receive them from only one source” (Schellenberg, 1975, p. 19).

According to Schellenberg, archives and libraries differ not only in their activities, but also in the type of material they collect:

one of the essential characteristics of archives [is] that they must have been produced or accumulated in direct connection with the functional activities of some government agency or other organization; and much of their significance depends on their organic relation to the agency and to each other. Their cultural values are incidental. Library materials, on the other hand, are produced in the first instance for cultural purposes (Schellenberg, 1975, p. 17).

Cook (1999), in *The Management of Information from Archives* adheres to Schellenberg’s characterization of archives and libraries, and also differentiates between the origins of archival records and the origins of library records by stating that, “Archives are information-bearing media which have been generated from *within* the organization; library and documentation materials are information-bearing media that were originally acquired from *outside* the organization” (Cook, 1999, p. 10).

In their work, both authors seem to characterize archives as passive and slow-moving,

and libraries as active and quick-changing. Archives cannot exist without a strong connection to the institution that authored their archival records, while libraries do not require this strong connection. Furthermore, archival collections grow organically; they receive materials as their agency creates them, while library collections are selected and curated.

From these statements, and the prior discussion of web archiving, it is clear that current web archives resemble library collections more closely than they do archival collections. Not only are the websites in a web archive actively selected and curated, they are collected precisely because of their cultural value. Perhaps a better term for the practice would be “web preservation” or “web collecting”; however, the term “web archiving” is already in widespread use and it would be difficult to change it.

## 2.4. The Concept of Information Quality: Paradigms, Theories, and Models

The second main focus of this dissertation is the notion of Information Quality (IQ). Information Quality has been studied widely across many fields and many scholars have attempted to define it. Most of the literature portrays IQ it as a multi-dimensional construct with facets such as accuracy, timeliness, and validity. IQ is also often described as highly subjective, dependent on both the context in which it is being applied and the audience that is viewing or utilizing the information. The following sections are intended as a survey of IQ across several disciplines.

### 2.4.1. Models of IQ in Information Science

The field of Information Science has produced several models of IQ, which are described here. Though none of the models cover web archives specifically, they are valuable contributions that can inform the creation of a theory of IQ.

#### 2.4.1.1. Taylor’s Value-Added Model

Taylor incorporated IQ as part of his value-added model. The value-added model describes the interaction between users and formal information systems. In it, the user is



the agent who actively seeks information from a formal system to achieve some objective. The user interface, which acts as the “negotiating space” between user and system, and the system itself is made up of a series of *value-added processes* (Taylor, 1986, p. 49). They are called this because they enhance or add value to the information being presented by the interface (Taylor, 1986, p. 51). For example, in a typical online library catalog, the system might implement a process to alphabetize results. The value-added process of alphabetizing will add the value of *browsing* to the interface. Users have internal criteria that they apply when responding to the information presented by the system. Taylor identified these criteria as ease of use, noise reduction, quality, adaptability, time-saving, and cost-saving (Taylor, 1986, p. 50).

As Taylor defines it, quality is “a user criterion which has to do with excellence or in some cases truthfulness in labeling” (Taylor, 1986, p. 62). Quality has the values of accuracy, comprehensiveness, currency, validity, and reliability as described below:

- (1) Accuracy assures an error-free transfer of data and information as it flows through the system and is eventually displayed to a user. Accuracy is a guarantee of a true copy, but is independent of the truth value of the information.
- (2) Comprehensiveness is the value added by the completeness of coverage of a particular subject or discipline. Comprehensiveness is especially valuable to historians, scholars, and lawyers.
- (3) Currency is the value added by the recency of the data acquired by the system and the capability of the system to reflect current modes of thinking in its access vocabularies. The optimal degree of currency varies depending on the user’s values and environment
- (4) Validity is the degree to which the information or data presented to users can be judged as sound. It is enhanced when these signals are presented

to the user, for example, by providing the assumptions behind the data or a critique of the research methodology.

- (5) Reliability is the trust a user has in the consistency of quality performance of the systems and its outputs over time. A system is reliable when it maintains an accepted level of accuracy, comprehensiveness, and currency. Taylor states that reliability is the summation of many aspects of quality.

(Taylor, 1986, p. 62–65)

Taylor derived his model after conducting a literature review of previous work on user criteria for information systems, examining an abstracting and indexing process, and summarizing the observations and experiences of skilled information professionals (Taylor, 1986, p. 54). In his book, Taylor goes on to describe the value-added processes that take place in libraries, abstracting and indexing services, and information analysis services such as the Congressional Research Service.

Though Taylor's description of IQ involved users' perceptions of quality, it was not derived from a study involving actual users. Also, the environments in which he applies the value-added model are all highly structured and hierarchical. Archives, libraries, and abstracting and indexing services all have firmly-established and standardized processes and procedures. This might not be the case in web archives, which are relatively new and have interfaces that are still in flux. The practice of web archiving is not yet so mature that it can be approached with the same rigor as Taylor's model requires.

Furthermore, Taylor's definition of currency could be problematic for web archives. If the purpose of a web archive is to preserve older websites for future use and study, as a type of historical record, then it is not so important that it contain the most up-to-the-minute information. Some users such as historians might regard a web archive to be more valuable the older its contents get. It seems that the notion of currency in a web archive is almost the opposite of what Taylor described.

#### 2.4.1.2. Rieh's Model

In 2002, Soo Young Rieh published a study that explores how users viewed the concepts of IQ and cognitive authority on the web. To address her research questions, she studied how users navigated web sites and how they judged the information quality of what they saw. She used a variety of instruments such as analysis of search logs, think-alouds, and follow-up interviews. In her results, five key aspects of information quality emerged: goodness, accuracy, currency, usefulness, and importance. These are summarized in Table 2.1.

Table 2.1

*Rieh's Facets of Information Quality*

<b>Facets</b>	<b>Keywords</b>
Good	Good job, bad, better, excellent, fine, nice, great, best, perfect, wonderful, incredible, cool, the state of the art, well kept site, well developed site
Accurate	Accurate, correct, right, precise
Current	Current, recent, up-to-date, out-of-date, old, timely
Useful	Useful, useless, hard to use, informative, helpful, doesn't help, it's not going to be of much use, didn't make good use
Important	Important

*Note.* Adapted from "Judgment of information quality and cognitive authority in the Web" by S.Y. Rieh, 2002, *Journal of the American Society for Information Science and Technology*, 53(2), p.152.

Rieh also found that the importance of each IQ facet varied with the task, for example, accuracy was the most important facet when users searched for medical information (Rieh, 2002, p. 152). Her definition of information quality differs from Taylor's in that it includes the

concepts of usefulness and goodness. Rieh defines usefulness as a subjective characteristic whose value is determined by the user, who judges whether or not the information is useful to her. The concept of goodness appears to denote “something in which the information excels or is superior” (Rieh, 2002, p. 157). People determine whether information is good by comparing the website to either their own expectations or to other websites. Their own knowledge and past experience is crucial to how they will evaluate the quality of the information: “judgments are not only based on external factors in terms of characteristics of information objects and sources but also on individuals’ own knowledge, which leads them to different predictions, expectations, and furthermore different evaluations” (Rieh, 2002, p. 157).

Rieh’s study is particularly important because, though it is informed by other theoretical models of IQ, her own model is derived from actual user research. This makes it more robust and verifiable than if it had not involved any users. Though her study was done on participants who surfed the Internet, a very similar study could be done to find out how users judge IQ for web archives. Researchers could give participants tasks to perform using web archives, such as: *find some web sites with good information about the recent elections*. As in Rieh’s study, users could be instructed to think aloud while they perform these tasks, and the experiment could involve a follow-up interview. Afterwards, researchers could analyze the navigation logs, think-aloud protocols, and the content of the interviews for clues about how users perceive IQ in web archives.

Though Rieh’s model could be applied to the study of IQ in web archives, some important differences might arise. The dimension of what is current in a web archive might be very different from Rieh’s definition of current. It is also important to note the differences in context between the open Internet in Rieh’s study and web archives. Because web archives are relatively new, most users will be unfamiliar with them, and so might have trouble using their prior knowledge to judge IQ. Some alternatives would be to debrief participants on the subject of web archives before the experiment, or to orient the study around academic

researchers (a possible future audience for web archives) and their perceptions of IQ. Though Rieh's model might not be entirely applicable to web archives, her methodology could inform future studies about their IQ.

#### 2.4.1.3. Bruce and Hillman's Guidelines for Metadata Quality

In their 2004 paper, *The Continuum of Metadata Quality: Defining, Expressing, Exploiting*, Bruce and Hillman (Bruce & Hillman, 2004) addressed the issue of metadata quality in Library and Information Science. They emphasized that quality is a quantifiable and measurable concept, and presented a list of quality measures and metrics that include completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility. For each of these measures, they specified criteria that librarians could use to assess the quality of metadata. It is important to note that Bruce and Hillman did not seek to create a theoretical model of IQ. Instead their aim was to establish a set of feasible guidelines for practitioners. The full list of quality measures and criteria is shown in Table 2.2

Table 2.2

#### *Bruce and Hillman's Quality Measures*

<b>Quality Measure</b>	<b>Definition</b>	<b>Quality Criteria</b>
Completeness	The element set describes the target object as completely as feasible and is applied to the target object population as completely as possible.	Does the element set completely describe the object? Are all relevant elements used for each object?

<b>Quality Measure</b>	<b>Definition</b>	<b>Quality Criteria</b>
Provenance	The persons who created the data, and their level of expertise, is known. Information about how the metadata was created, extracted, and transformed is included.	Who is responsible for creating, extracting, or transforming the metadata? How was the metadata created or extracted? What transformation have been done on the data since its creation?
Accuracy	The information provided in the values is correct and factual and lacks typographical errors, uses standard abbreviations, and so on.	Have accepted methods been used for creation or extraction? What has been done to ensure valid values and structure? Are default values appropriate, and have they been appropriately used?
Conformance to Expectations	The metadata contains those elements that the community would reasonably expect to find. It does not contain “false promises.”	Does metadata describe what it claims to? Are controlled vocabularies aligned with audience characteristics and understanding of the objects? Are compromises documented and in line with community expectations?
Logical Consistency Coherence	Elements are conceived in a way that is consistent with standard definitions and concepts used in the subject or related domains and are presented to the user in consistent ways.	Is data in elements consistent throughout? How does it compare with other data within the community?

Quality Measure	Definition	Quality Criteria
Timeliness	Metadata is in synchronization with the target object and has been recently reviewed and verified. The dissemination of metadata is synchronized with the dissemination of the object to which it applies.	Is metadata regularly updated as the resources change? Are controlled vocabularies updated when relevant?
Accessibility	Metadata can be read and understood by users.	Is an appropriate element set for audience and community being used? Is it affordable to use and maintain? Does it permit further value-adds?

In their guidelines, Bruce and Hillman present many of the same IQ facets seen in prior models, such as accuracy and completeness, though with some new information. For example, according to the authors, high-quality metadata has clear *provenance*, that is, it contains information about who created the metadata. Two other interesting aspects of quality are mentioned: Conformance to expectations (metadata must reasonably conform to community standards and expectation) and accessibility (metadata must be able to be read and understood by users). All three of these guidelines are useful when applied to metadata of any type; however, they are problematic when used to determine the IQ of a web archive.

Currently, it is often impossible to know the level of expertise of the person that created the web archive. It is also difficult to find out if a specific archived website has undergone any transformation since its creation. This makes provenance difficult to determine. Provenance as a dimension of IQ in web archives might be more applicable in small web archives that are purposefully created by a human curator. In the context of a large web archive, where

websites are automatically harvested by a crawler and automatically indexed and loaded into a replay mechanism, provenance is of limited value.

The measures of conformance to expectations and accessibility are subjective and depend entirely on the audience using the web archives and on the larger web archiving community. In a mature field such as Library and Information Science, metadata formats and controlled vocabularies have arisen that set standards for what a community can reasonably expect. Such expectations have yet to arise in the field of web archiving.

#### 2.4.1.4. Stvilia's Framework for Information Quality

In his dissertation, Besiki Stvilia (2006) developed a general IQ measurement and assessment framework based on his study of two large-scale collections of two large classes of information objects: Simple Dublin Core (DC) metadata records and Wikipedia articles. The framework was validated and refined by developing specific IQ measurement models for each collection. He defined three high-level categories of IQ and discussed how they could be measured:

- (1) Intrinsic: includes dimensions of information quality that can be assessed by measuring internal attributes/characteristics of information entities themselves in relation to some reference standard in a given culture. Examples include spelling mistakes (dictionary), conformance to formatting or representation standards (HTML validation), and information currency (age with respect to a standard index date, e.g. "today"). In general, intrinsic IQ attributes persist (as long as the reference culture does not change often) and depend little on context. Hence, these can be measured more or less objectively.
- (2) Relational/Contextual: measures relationships between information and some aspects of its usage context. One common subclass in this category includes the representational quality dimensions, which measure how well an information entity reflects (maps) some external condition (e.g., actual accuracy of addresses in an ad-



dress database) in a given context. Since related entities can change independently, relational/contextual characteristics of an information entity are not persistent with the entity itself. The usage context refers to the context of an activity system, which can change in time and space.

- (3) Reputational: measures the position of an information entity in a cultural or activity structure, often determined by its origin and its record of mediation.

(Stvilia, 2006, p. 209)

These distinctions are helpful for the researcher that is seeking to find ways to measure IQ because they help to identify those dimensions that will be most easily operationalized. Generally, intrinsic dimensions of IQ are the easiest to operationalize because they persist over time and usually have well-known and articulated reference standards. In contrast, relational and reputational IQ dimensions are context and user-dependent and usually vary over time, making them much more difficult to operationalize.

Each of Stvilia's categories comprise multiple IQ dimensions. Table 2.3 shows the IQ dimensions described by Stvilia and their definitions in the context of intrinsic IQ and relational IQ. Several dimensions, such as accuracy/validity, complexity, informativeness/redundancy, naturalness, precision/completeness, semantic consistency, and structural consistency have both intrinsic and relational definitions, while others, such as cohesiveness, currency, relevance, security, verifiability, and volatility belong to only one category. Unlike intrinsic IQ and relational IQ, reputational IQ has only one dimension, authority, defined as the "degree of reputation of an information object in a given community or culture" (Stvilia, 2006, p. 80).

Table 2.3

*Stvilia's Dimensions of Quality*

<b>IQ dimension</b>	<b>Definition in Intrinsic IQ</b>	<b>Definition in Relational/Contextual IQ</b>
Accuracy/Validity	extent to which information is legitimate or valid according to some stable reference source such as a dictionary, standard schema and/or set of domain constraints and norms (soundness)	degree to which an information object correctly represents another information object, process or phenomenon in the context of a particular activity and/or culture
Accessibility	N/A	speed, ease of locating and obtaining an information object relative to a particular activity
Cohesiveness	extent to which the content of an object is focused on one topic	N/A
Complexity	extent of cognitive complexity of an information object measured by some index/indices	degree of cognitive complexity of an information object relative to a particular activity
Currency	the age of an information object	N/A
Informativeness / redundancy	amount of information contained in an information object: the ratio of the size of the informative content (measured in word terms which are stemmed and stopped) to the overall size of an information object	extent to which the information is new or informative in the context of a particular activity/community

<b>IQ dimension</b>	<b>Definition in Intrinsic IQ</b>	<b>Definition in Relational/Contextual IQ</b>
Naturalness	extent to which an information object's model/schema and content are expressed by conventional, typified terms and forms according to some general purpose reference source	degree to which an information object's model and content are semantically close to the objects, states or processes they represent in the context of a particular activity (measured against the activity/-community specific ontology)
Precision / completeness	granularity or precision of an information object's model or content values according to some general purpose IS-A ontology such as WordNet	extent to which an information object matches the precision and completeness needed in the context of a given activity
Relevance (aboutness)	N/A	extent to which information is applicable and helpful/applicable in a given activity
Security	N/A	extent of protection of information from harm in the context of a particular activity
Semantic consistency	extent of consistency of using the same values (vocabulary control) and elements for conveying the same concepts and meanings in an information object	extent of consistency of using the same values (vocabulary control) and elements required or suggested by some external standards and recommended practice guides for conveying the same concepts and meanings in an information object

<b>IQ dimension</b>	<b>Definition in Intrinsic IQ</b>	<b>Definition in Relational/Contextual IQ</b>
Structural consistency	extent to which similar attributes or elements of an information object are consistently represented with the same structure, format and precision	extent to which similar attributes or elements of an information object are consistently represented with the same structure, format and precision required or suggested by some external standards and recommended practice guides
Verifiability	N/A	extent to which the correctness of information is verifiable and/or provable in the context of a particular activity
Volatility	N/A	amount of time the information remains valid in the context of a particular activity

*Note.* Adapted from “Measuring information quality” by B. Stvilia, 2006. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database: (Order No. 3223727).

For each of these dimensions, Stvilia proposed metrics that could be applied in order to measure IQ. Some of these metrics were already quite well-known and widely-implemented, while others were new and of his own devising. Table 2.4 shows Stvilia’s quality metrics. Discussing them all in detail is beyond the scope of this dissertation, therefore I focus only on those dimensions deemed most relevant to the field of web archiving: accuracy/validity, cohesiveness, complexity, precision/completeness, and relevance.

Table 2.4

*Stvilia's List of IQ Metrics*

<b>IQ dimension</b>	<b>Metric for Intrinsic IQ</b>	<b>Metric for Relational IQ</b>
Accuracy/Validity	Spelling Error Rate	Num. Broken External Links; Euclidean Similarity Distance
Accessibility	N/A	Throughput w/r of # Requests; Throughput w/r of Amount of data
Cohesiveness	IDF/AverageIDF; Cosine Angular Similarity Metric	N/A
Complexity	Cyclomatic Complexity Index; Flesch Reading Ease Score; Flesch-Kincaid Grade Level; Fog Index; Average Sentence Length; Average Word Length	Flesch Reading Ease Score; Flesch-Kincaid Grade Level; Fog Index; Average Sentence Length; Average Word Length
Currency	Currency	N/A
Informativeness / redundancy	Information Noise Metric; Distinct Elements Ratio; ContentSpecificity; Num. of Repeated Elements <sup>a</sup> ; Num. of Images <sup>b</sup> ; Diversity (# of Unique Editors/Total # of Edits) <sup>b</sup>	Kullback-Leibler Divergence; ID-F/AverageIDF
Naturalness	ContentSpecificity	Cosine Angular Similarity Metric
Precision / completeness	Completeness Ratio; Article length (in # of characters) <sup>b</sup>	Completeness Ratio; FRBR Completeness <sup>a</sup>
Relevance (aboutness)	N/A	NumberOfClicks; CitationCount (cited by); VectorSpaceModel; AuthorityAndHub; PageRank

<b>IQ dimension</b>	<b>Metric for Intrinsic IQ</b>	<b>Metric for Relational IQ</b>
Security	N/A	Break-ins Ratio
Semantic consistency	Semantic Consistency Index; Article Age (in days) <sup>b</sup> ; Admin. Edit Share (Num. of Admin Edits / Total Num. of Edits) <sup>b</sup>	Semantic Consistency Index
Structural consistency	Structural Consistency Index; Article Age (in days) <sup>b</sup> ; Admin. Edit Share (Num. of Admin Edits / Total Num. of Edits) <sup>b</sup>	Structural Consistency Index
Verifiability	N/A	Num. of URLs; Num. of References (citing)
Volatility	N/A	LinkRot

*Note.* Adapted from “Measuring information quality” by B. Stvilia, 2006. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database: (Order No. 3223727).

<sup>a</sup> The metric is specific to Stvilia’s analysis of metadata records.

<sup>b</sup> The metric is specific to Stvilia’s analysis of Wikipedia pages.

To measure intrinsic accuracy/validity, Stvilia proposed the use of the *Spelling Error Rate*, which is the number of spelling errors in a document divided by its size (Stvilia, 2006, p. 64). While this is useful for metadata objects and Wikipedia articles, it is not appropriate in a web archiving context. A high-quality archived website mirrors the appearance and functionality of the original, live site. If there are spelling mistakes in the original site, they should be reproduced in the archived version, with no corrections. Instead of Spelling Error Rate, his generalized notion of an accuracy measure is more appropriate and can more easily be adapted to web archiving, “a ratio of the number of valid or invalid values, or elements

over the length of the object, or the total number of elements in the object” (Stvilia, 2006, p. 64). For example, in an archived website, an “invalid” value could potentially be an element that does not look or behave like the original.

For relational accuracy/validity, Stvilia proposed using both the number of broken external links and the *Euclidean Similarity Distance*. In web archiving, the number of broken links is a potentially useful measure; however, web archives usually have specific, limited collecting scopes. Their mission is to preserve websites about a particular topic and not others, so a broken link in a web archive is not necessarily an indication of a quality problem, it might instead be a reflection of its collecting scope. The second accuracy measure is the Euclidean Similarity Distance, a well-known measure from the field of Information Retrieval, where there are several ways of measuring the similarity between two words, or between two documents. If using the standard vector model, where each word or each document is represented by a vector, a measure of similarity can be computed in a number of different ways. One such measure is the *Euclidean distance*, also known as the L2 norm, as seen in Equation 1.

$$(1) \quad ed(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^N (x_i - y_i)^2}$$

The Euclidean distance is a measure of the distance between two vectors. Though it can be used to calculate the similarity between two words, it is rarely used because it is very sensitive to extreme values (Jurafsky & Martin, 2008, p. 696). For the Euclidean distance to be truly useful, it should best be used in situations where the vectors being compared are of equal length (Sarkar, 2016, p. 278).

Cohesiveness, according to Stvilia, can be measured by using the *Cosine Angular Similarity* and the *AverageIDF* metrics from Information Retrieval. Cosine similarity is one commonly-used metric that is not sensitive to high-frequency words. Cosine similarity, as shown in Equation 2, measures the angle between two vectors. The values calculated by

cosine similarity range between 0, for vectors that do not share any terms, to 1, for vectors that are identical, to -1, for vectors that point in opposite directions (Jurafsky & Martin, 2008, p. 699).

$$(2) \quad k(x, y) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| * \|\mathbf{y}\|}$$

The formula for AverageIDF is shown in Equation 3. In it,  $n$  is the number of terms in the document,  $df(i)$  is the number of documents containing the  $i^{th}$  item and  $N$  is the total number of documents in the collection (Stvilia, 2006, p. 71). In the context of web archiving, both measures could be used to gauge the cohesiveness of a web archive that is focused on a single topic. However, these metrics could not be entirely successful if applied to web archives, because web archives often contain content that might seem off-topic at first glance, but actually help to make archived websites look and function a certain way. This is specially the case for elements such as JavaScript and CSS files, which on the surface are not topic-specific but are responsible for controlling the look and feel of a page. Metrics such as Cosine Similarity and AverageIDF are potentially useful for measuring cohesiveness in web archives, but have weaknesses that limit their applicability.

$$(3) \quad AverageIDF = \frac{\sum_{i \in d} idf(i)}{n} = \frac{\sum_{i \in d} (-\log_2 df(i)/N)}{n}$$

For the dimension of complexity, Stvilia proposed the *Flesch Reading Ease Score*, the *Flesch-Kincaid Grade Level* measure, the *Fog Index*, the *Average Sentence Length*, the *Average Word Length*, and the *Cyclomatic Complexity Index* for both intrinsic and relational IQ categories. The first five of these measures represent ways of measuring the linguistic complexity of a text and thus its readability. The last measure, the Cyclomatic Complexity Index, is a software metric that indicates the complexity of a program, as indicated by its



control flow (Stvilia, 2006, p. 65). None of these measures are truly appropriate for web archiving because they concern themselves only with textual content, while a website is made up of a myriad of components that include text, graphics, videos, links, and dynamic scripts. The readability of the text on a website has no bearing on its ability to be preserved for the future.

To measure the intrinsic precision or completeness of an information object, Stvilia proposed the *Completeness Ratio*, defined as the “ratio of the number of ‘known’ incomplete elements ... over the total number of elements or the size of the object. The ‘known’ incomplete means empty elements (null values) or the elements with explicitly incomplete values” (Stvilia, 2006, p. 67). This metric could be usefully applied to web archiving, since a lack of completeness is a common problem for archived websites, as described in Section 2.5.1. There is one simple, but important caveat: there are always elements of a website, or even entire pages that web archivists do not know are missing. Stvilia’s definition operates within a *Closed World Assumption*, where we know all variables that exist and whether or not they have missing values. In contrast, web archiving usually operates within an *Open World Assumption*, where it is not actually known whether every variable exists or not. These philosophical concepts are further explored in Section 2.4.2.2.

Stvilia also proposes two, more specific metrics to measure relational precision and completeness: the length of an article in the case of Wikipedia, and the *Functional Requirements for Bibliographic Records (FRBR) Completeness* metric for metadata records. This last metric uses the Western States Dublin Core Metadata Best Practices (WSDCMBP) set of required elements as the gold standard that must be met. According to Stvilia, the completeness of a metadata record can be evaluated by asking how well it supports four important activities: Find, Identify, Select, and Obtain. He defines the FRBR Completeness as the product of the differences between the ideal number of elements necessary to support a task and the critical number of elements, as shown in Equation 4.

$$(4) \quad FRBR\_RC = \prod_{t=1}^n \frac{(e_{Ct} - e_{It})^2 - (e_t - e_{It})^2}{(e_{Ct} - e_{It})^2}$$

In the formula for FRBR Completeness,  $t$  is the number of tasks in the activity,  $e_{Ct}$  is the critical number of the relevant distinct elements for the task  $t$ ,  $e_{It}$  is the ideal number of the relevant distinct elements for the task  $t$ , and  $e_t$  is the number of the relevant elements for the task  $t$  in the record. The largest value ( $FRBR\_RC = 1$ ) is achieved when  $e_t = e_{It}$  for all  $t$ , and  $FRBR\_RC = 0$  when  $e_t = e_{Ct}$ . The values of  $FRBR\_RC$  range from 0 to 1. In the case of the FRBR activity the number of task/actions  $t$  equals 4 (Stvilia, 2006, p. 122). As a metric, the FRBR might possibly be adapted to web archiving.

For the dimension of relevance, which exists only within relational IQ, Stvilia proposed using metrics such as the *Vector Space Model*, the *NumberOfClicks*, the *CitationCount*, *AuthorityAndHub*, and *PageRank*. The Vector Space Model (VSM), an essential Information Retrieval paradigm introduced in 1975 by Salton, Wong, and Yang (1975), queries and documents are viewed as vectors of words and the relevance between the query and the document is calculated by the similarity (distance) function of their vector representations. Metrics such as Cosine Similarity (Equation 2) and Euclidean Distance (Equation 1) are built around the VSM. The VSM and its related metrics have one important assumption that usually goes unarticulated: that researchers have at their disposal a body of user-generated queries expressing clear information needs that can be compared to the documents present in an IR system. This is simply not the case in web archiving. Web archives are often created as part of a legal mandate by an institution, they are not always publicly accessible, have an undefined user base, and their collecting policies are often vague or not accessible. Retrieval experiments of the kind common in the IR field have yet to take place in web archiving.

Stvilia discusses using the number of user clicks as way to measure relevance, with higher number of clicks indicating greater popularity, and thus greater relevance to the users' information needs. He also mentions Kleinberg's Authority and Hub algorithm and Brin and

Page’s PageRank algorithm, which use different tactics to assign weights to retrieval results in order to rank them (Stvilia, 2006, p. 78). These metrics have the VSM as their base, and thus suffer from the same weaknesses when applied to web archiving.

One key characteristic of Stvilia’s work is that he sees IQ in terms of how it affects the final product. For him, an IQ problem only becomes critical when it affects the quality of an outcome (Stvilia, 2006, p. 96). For example, as mentioned before, a metadata record must support the activities of Find, Identify, Select, and Obtain. The value of quality then becomes “the value of the activity outcome with and without the quality” (Stvilia, 2006, p. 104). We cannot say we have improved the IQ of a metadata record if editing it does not result in a better outcome when the user or system attempts to find, identify, select, and obtain information. The effectiveness of a change in metadata quality can thus be expressed in terms of the change in value of the activity outcome, as seen in Equation 5.

$$(5) \quad E(\Delta QM(t+1, t)) = \frac{V(\Delta OA | \Delta QM(t+1, t))}{C(\Delta QM(t+1, t))}$$

In this equation, *OA* is the Activity Outcome, *QM* is Metadata Quality, *V* stands for Value, *E* is Effectiveness, *C* stands for Cost, and *t* stands for time. In other words, the effectiveness of a change in metadata quality is the change in the value of the outcome divided by the cost of the change itself. The cost of changing the metadata is inversely proportional to its effectiveness. At some point the cost might be too great to offset the (supposed) increase in effectiveness.

## 2.4.2. IQ in Philosophy

### 2.4.2.1. Floridi and the Concept of “Fit for Purpose”

In the past two decades, philosophers have been paying special attention to defining IQ. In an editorial in *Philosophy and Technology*, Luciano Floridi wrote about the current state of the concept of IQ. He highlights the fact that in the U.S and the U.K there are currently

several government programs aimed at establishing standards for information quality and similar efforts have also taken place in academia. However, Floridi dismisses these efforts as having had little impact because they have “failed to combine and cross-fertilise theory and practice” (Floridi, 2013, p. 2).

As in IS, many philosophers also see IQ as multidimensional and define it similarly, including factors such as accuracy and relevance. However, some have put forward an additional dimension of IQ known as “fit for purpose,” which denotes anticipating and meeting user requirements. Floridi accepts that IQ is multi-dimensional and is composed of facets such as accuracy, objectivity, accessibility, security, relevancy, timeliness, interpretability, and understanding. But his main point is that past work on IQ has misrepresented the concept of fit for purpose, which has sometimes been treated as a one-dimensional or absolute concept. He argues that the concept of fit for purpose is bi-categorical. High-quality information is:

- (1) Optimally fit for the specific purpose/s for which it is elaborated (purpose-depth)
  - (2) Easily re-usable for new purpose/s (purpose-scope)
- (Floridi, 2013, p. 4).

Floridi argues that there is an important tension between these two aspects. Often the better a piece of information fits its original and intended purpose, the less likely it can be reused for another purpose, and vice versa. To address this issue, Floridi proposes that traditional dimensions of quality such as accuracy and timeliness be measured along the concepts of purpose-depth and purpose-scope. Our concept of “fit for purpose” would then change. For example, a pre-Copernican book on astronomy would have low information quality if its purpose was to teach its audience about the nature of the galaxy, but it would have high information quality if its purpose was to teach us about the historical development of Ptolemaic astronomy (Floridi, 2013, p. 5).

Floridi’s concepts of fit for purpose, purpose-depth, and purpose-scope would be ap-

plicable to the study of web archives. In the web archiving community there seems to be some confusion as to the audience (real or potential) of a web archive. Some organizations aim to capture websites for a general audience, others focus on very specific audiences such as researchers, while others do not even specify an audience; information is simply captured and preserved. Having a clear idea of the purpose-depth and purpose-scope of a web archive might help web archivists improve their archives.

Ultimately, fit for purpose might be just another version of the “usefulness” dimension as described by Rieh, but Floridi’s definition is more detailed and nuanced. Fit for purpose might be used as a dimension to measure IQ in a web archive, particularly in a highly-specialized one. For example, if a web archive’s stated purpose-depth is to serve as an important resource for librarians studying Information Retrieval, but librarians found the archive to have little utility, then the archive’s IQ might be low. This would alert web archivists that perhaps the selection processes for the archive might need to be revised.

#### 2.4.2.2. Batini, Palmonari, and Viscusi’s Model

Other philosophers have added to and expanded on the notions of IQ. For example, Batini, Palmonari, and Viscusi (2012) make some very important points on the subject. They posit that humans evaluate IQ in two important ways:

**Method 1:** By using a reference version of the information

**Method 2:** By referring to the perceptual and/or technological characteristics of information.

These characteristics depend on the type of information representation

Batini, Palmonari, and Viscusi (2012, p. 8).

In other words, people evaluate the quality of different types of information in different ways. For example, a person might read an article that says the capital of Spain is Barcelona. If she consults an encyclopedia and finds that the capital of Spain is actually Madrid, she might judge the original article to have poor IQ (Method 1, IQ is judged by comparison to a

reference version). But if she looks at a photograph she might instantly judge it to have bad quality if she finds the image blurry or unfocused (Method 2, IQ is judged by perception).

The authors put forward their own definition of IQ, with the different dimensions clustered according to their perceived similarity:

- (1) *Accuracy/correctness/precision* refer to the adherence to a given reference reality.
- (2) *Completeness/pertinence* refer to the capability to express all (and only) the relevant aspects of the reality of interest.
- (3) *Currency/volatility/timeliness* refer to the information up-to-dating.
- (4) *Minimality/redundancy/compactness* refer to the capability of expressing all the aspects of the reality of interest only once and with the minimal use of resources.
- (5) *Readability/comprehensibility/usability* refer to ease of understanding and fruition by users.
- (6) *Consistency/coherence* refer to the capability of the information to comply to all properties of the membership set (class, category,...) as well as to those of the sets of elements the reality of interest is in some relationship.
- (7) *Credibility/reputation*, information derives from an authoritative source.

(Batini et al., 2012, p.16)

(Batini et al., 2012, p. 11)

The relative importance of these measures depends, again, on the type of information representation. Batini et al. (2012) distinguishes primarily between two: highly-structured data (such as the contents of a relational database or a geographic map) and unstructured data (such as a photograph or short story). They observe that “the less the information is structured, from a restricted domain to a totally unstructured domain, the more subjective measures prevail on objective measures.” (Batini et al., 2012, p. 18)

The model presented by Batini et al. (2012) might be the most comprehensive one yet. Like the other models, the dimensions of currency/volatility/timeliness in Batini et al.’s

model might mean something different in the world of web archives. However, this model also includes the dimensions of consistency and coherence, which are absent from the others. For a web archive to be of high quality, its components (the archived web sites) must have been consistently captured (no relevant websites were left out) and must replay consistently (an archived website that at times looks identical to the original, and at other times very different is not consistent).

Similarly, the individual archived web site must be coherent with the web archive as a whole. For example, a web archive on the topic of Chemistry that also contains large numbers of pornographic websites might be judged to be incoherent. Though this model has not been applied to studies of IQ in web archives, it could prove a useful starting point.

In a later book, *Data and Information Quality: Dimensions, Principles and Techniques*, Batini and Scannapieco (2016) further refined the model they had previously created. They focus specifically in defining data quality dimensions in the context of relational databases. The framework comprised the following dimensions (the item in italics is the representative dimension of the cluster, followed by other member dimensions):

- (1) *Accuracy*, correctness, validity, and precision focus on the adherence to a given reality of interest.
- (2) *Completeness*, pertinence, and relevance refer to the capability of representing all and only the relevant aspects of the reality of interest.
- (3) *Redundancy*, minimality, compactness, and conciseness refer to the capability of representing the aspects of the reality of interest with the minimal use of informative resources.
- (4) *Readability*, comprehensibility, clarity, and simplicity refer to ease of understanding and fruition of information by users.
- (5) *Accessibility* and availability are related to the ability of the user to access information from his or her culture, physical status/functions, and technologies available.

- (6) *Consistency*, cohesion, and coherence refer to the capability of the information to comply without contradictions to all properties of the reality of interest, as specified in terms of integrity constraints, data edits, business rules, and other formalisms.
- (7) *Usefulness*, related to the advantage the user gains from the use of information.
- (8) *Trust*, including believability, reliability, and reputation, catching how much information derives from an authoritative source. The trust cluster encompasses also issues related to security.

(Batini & Scannapieco, 2016, "A Classification Framework for Data and Information Quality Dimensions", para. 2)

After presenting their framework, the authors describe the dimensions in detail and propose ways of operationalizing them. They define accuracy as "the closeness between a data value  $v$  and a data value  $v'$ , considered as the correct representation of the real-life phenomenon that the data value  $v$  aims to represent" (Batini & Scannapieco, 2016, "Accuracy Cluster", para. 1). They identify two types of accuracy: syntactic and semantic. Syntactic accuracy is the closeness of a value  $v$  to the corresponding definition domain  $D$  (Batini & Scannapieco, 2016, "Structural Accuracy Dimensions", para. 2). Semantic accuracy is the closeness of a value  $v$  to the true value  $v'$  (para. 4). Semantic accuracy is often called correctness in other models. The authors state that syntactic accuracy is best measured with a distance function, while semantic accuracy is better measured with a binary value of yes/no or 0/1 (para. 5).

In their work, Batini and Scannapieco (2016) describe two latent aspects of completeness that have gone unexplored by researchers in other fields: the *closed world assumption* (CWA) and the *open world assumption* (OWA). They point out that in a database "a value can be missing [null] either because it exists but is unknown or because it does not exist at all or because it may exist but it is not actually known whether it exists or not" (Batini & Scannapieco, 2016, "Completeness of Relational Data", para. 2). The first case describes



the CWA assumption, where it is assumed that the values in a relational table  $r$  represent *all* the facts of the real world. The second case describes the OWA, where it is impossible to state whether values *not* represented in  $r$  are true or false. In a relational database with OWA and no null values, given the relation  $r$ , there is a *reference relation* called  $ref(r)$  that contains the objects in the real world. The authors then define completeness as the fraction of the tuples in  $ref(r)$  that are actually represented in  $r$ , as seen in Equation 6.

$$(6) \quad C(r) = \frac{|r|}{|ref(r)|}$$

Completeness in this case is measured in terms of size, that is, percentage of the the real world  $ref(r)$ , that is presented by the model  $r$ . From a web archiving perspective,  $ref(r)$  can be seen as the actual, live website that we seek to represent using an archived website  $r$ .

Batini and Scannapieco (2016) also explored the temporal dimension of completeness as it concerned web data. They acknowledged that web data is characterized by information that is continuously published and updated, and so completeness on the web also varies with time. The authors introduced the notion of *completability*, defined as an area  $Cb$  of a function that represents how completeness evolves between an instant  $t\_curr$  of observation and  $t\_max$ , the maximum time in the series (Batini & Scannapieco, 2016, “Completeness of Web Data”, para. 4). This definition is shown in Equation 7:

$$(7) \quad \int_{t\_curr}^{t\_max} C(t)$$

In this equation  $t \in [t\_pub, t\_max]$ , where  $t\_pub$  is the initial instant of publication of the data.

Batini and Scannapieco (2016) state that the dimensions of data quality are not independent of each other, that is, they are strongly correlated. If one dimension of quality is favored during some process, this may have negative consequences for other dimensions.

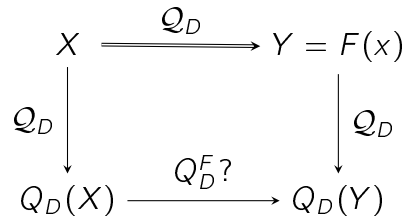


Figure 2.1. A graphical representation of the general problem statement for the definition of quality composition. Adapted from *Data and information quality: Dimensions, principles and techniques* [Kindle book] by Batini, C., & Scannapieco, M. (2016). Cham, Switzerland: Springer International Publishing.

For example, prioritizing the timeliness of data may lead to lower accuracy, completeness, or consistency. The reverse is also true, giving preference to accuracy, completeness, or consistency dimensions may cause delays and the resulting data will not be as timely (Batini & Scannapieco, 2016, "Trade-Offs Between Dimensions", para. 1).

The authors go on to define an algebra for defining information quality in terms of its dimensions, which they call the *information quality composition activity*. This definition is meant to be used in a variety of contexts involving relational databases, from e-business and e-government. It is designed to measure the overall quality of a composite information object which is made up of information elements from many sources. The model, shown graphically in Figure 2.1, has the following elements:

- $X$  is the set of information sources
- $F$  is a general composition function defined on a set of algebraic operators  $O = [o_1, \dots, o_k]$ , such as union, intersection, Cartesian product, etc.
- $D$  is an IQ dimension such as completeness or accuracy
- $Q_D^F$  is a function that evaluates the quality of an object for different hypotheses and different operators

According to the model, the function  $Q_D(X)$  calculates the value of the quality dimension  $D$  for the set of sources  $X$ . The value of  $D$  for the composite information object  $Y$  equals  $F(x)$ , or  $Q_D(Y)$ . In the figure, the function  $Q_D^F(X)$  calculates  $Q_D(Y)$  starting from  $Q_D(X)$  (Batini & Scannapieco, 2016, "Quality Composition", para. 3).

### 2.4.3. IQ in Computer Science

In their paper, Zhu and Gauch (2000) explored how quality metrics can be used to improve the performance of Information Retrieval systems. Their focus was on finding and using metrics that could be operationalized. The authors reviewed numerous quality metrics, and selected the ones they felt were amenable to automatic analysis. For their experiments, the authors operationalized the metrics as seen in Table 2.5.

Table 2.5

*Quality Metrics and their Operational Definitions.*

<b>Qualify Metric</b>	<b>Defined As</b>	<b>Operationalized as</b>
Currency	How recently a web page has been updated.	The time stamp of the last modification of the document
Availability	The number of broken links contained by the web page.	The number of broken links on a page divided by the total numbers of links it contains.
Information-to-Noise Ratio	Proportion of useful information contained in a web page of a given size.	The total length of the tokens (words) divided by the size of the document.
Authority	The reputation of the organization that produced the web page.	A score from the Yahoo Internet Life (YIL) reviews.

Qualify Metric	Defined As	Operationalized as
Popularity	How many other Web pages have cited this particular Web page.	The number of links pointing to a web page.
Cohesiveness	The degree to which the content of the page is focused on one topic.	How closely related the major topics in the page are (see Equation 9).

*Note.* Adapted from “Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web” by Zhu, X., & Gauch, S., 2000, In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p.289.

They defined the “goodness” of a site as its overall quality. Goodness can be defined as:

$$(8) \quad G_i = \bar{W}_i * (a_s'' * \bar{T}_i + b_s'' * \bar{A}_i + c_s'' * \bar{I}_i + d_s'' * \bar{R}_i + e_s'' * \bar{P}_i + f_s'' * \bar{C}_i)$$

where  $\bar{W}_i, \bar{T}_i, \bar{A}_i, \bar{I}_i, \bar{R}_i, \bar{P}_i$  are the means of information quantity, currency, availability, information-to-noise ratio, authority, and popularity of site  $i$  across topics relevant to the query,  $\bar{C}_i$ , is the cohesiveness of site  $i$ , and  $a_s'', b_s'', c_s'', d_s'', e_s'', f_s''$  are the weights representing the importance of each quality metric. (Zhu & Gauch, 2000, p. 291)

The cohesiveness of a site was defined as

$$(9) \quad C = \frac{\frac{N * (N - 1)}{2} * \frac{M * (M - 1)}{2} + \sum \sum P_{ij}}{\frac{N * (N - 1)}{2}} \quad (i, j = N, N - 1; i < j)$$

where  $N$  is the maximum number of top matching topics requested,  $M$  is the minimum of  $N$  and the number of matching topics returned, and  $P_{ij}$  is the length of the shared path between topic  $i$  and  $j$  divided by the height of the ontology (Zhu & Gauch, 2000, p. 290).

Zhu and Gauch (2000) evaluated the ability of the quality metrics to improve search effectiveness for three tasks: 1) query document matching (centralized search), query routing, and information fusion. They compared their results with those generated by using a more traditional IR system that did not incorporate quality metrics.

The authors found that incorporating currency, availability, information-to-noise ratio, and page cohesiveness metrics significantly improved search effectiveness for the first task. For site selection, search effectiveness improved significantly when availability, information-to-noise ratio, popularity, and cohesiveness were used. For the final task of information fusion, incorporating the popularity metric also resulted in significant improvements in search effectiveness. Zhu and Gauch (2000) concluded that overall, quality metrics can improve search effectiveness. They specifically singled out information-to-noise ratio as the most useful metric because it resulted in the greatest improvement in results (Zhu & Gauch, 2000, p. 294).

If we apply the definitions proposed by Zhu and Gauch to the practice of web archiving, several interesting points appear. As with the models described in previous sections, the notion of currency proposed by the authors does not apply to web archives. Also, metrics such as popularity and authority might be very different for web archives, since these are usually created by a human curator based on specific collections criteria.

However, the work of Zhu and Gauch also has important advantages over the definitions provided by other authors. Parallel quantitative metrics could be constructed for the specific case of web archives. Because their IQ definitions are quantitative, the process of measuring IQ for a web archive could be automated, rather than having to rely on a human evaluator to judge the IQ for every archived website. The authors also defined the goodness of a website in vector form, that is, as the sum total of the individual measures of goodness of each of its pages. By following this example, different formulas could also be created: one to measure the IQ of a single archived website, another to measure the IQ of an entire web

archive (which would include many websites).

But the most significant part of Zhu and Gauch's quantitative approach is the notion of weights. As can be seen in 8, each quality metric has a corresponding weight representing its importance; more important metrics are given more weight and less important metrics are given less weight. This flexibility would allow the equation to fit a variety of situations. For example, a web archiving institution that placed more emphasis on the metrics of availability and cohesiveness could set the weight of those aspects to be relatively high, and set the weights for the other metrics to be relatively low values.

The flexibility and practicality of a quantitative approach would be ideal for many institutions that practice web archiving. Many institutions compile large web archives, each containing hundreds or sometimes thousands of individual archived websites. The process of determining the IQ of archived resources usually falls on human evaluators, who are often pressed for time and lack adequate tools to complete the task. A set of carefully-created quantitative metrics would help alleviate the burden placed on human evaluators.

This section presented a history of the field of web archiving, explored the connections between web archiving and other disciplines, and gave an overview of IQ and how it is defined in several fields. Some of the ideas presented in the reviewed literature have informed the proposed research study.

#### 2.4.4. IQ in Other Fields

Fields such as Manufacturing, Management Science, and Business have long dealt with the problems of quality and have developed a variety of models and frameworks to describe it. The models in these disciplines lean heavily towards operationalization and validation. This section is not intended as a comprehensive review of IQ theories in those fields, but rather as an overview of IQ concepts that might be relevant or useful to operationalizing IQ in web archiving.

In the field of Industrial Engineering, Taguchi, Elsayed, and Hsiang (1988) saw quality

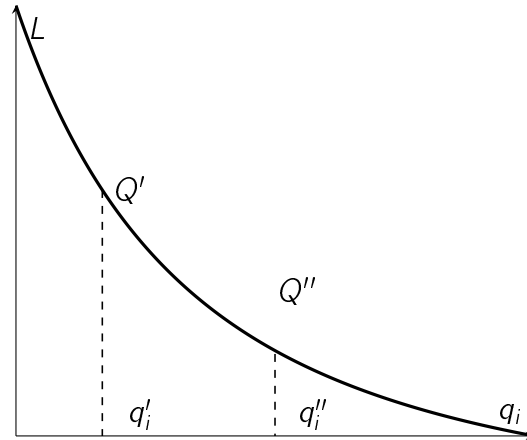


Figure 2.2. The larger-the-better (L type) relationship curve. Adapted from “Measuring information quality” by B. Stvilia, 2006. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database: (Order No. 3223727).

problems as inextricably tied to the loss of value of a product. If a product characteristic deviated from its target value, it would result in a loss incurred upon society, the “value loss”. The authors described three types of relationships between the deviation of a product from its target value, its quality, and its value loss: the *Larger-The-Better* (L type) relationship, the *Smaller-The-Better* (S type) relationship, and the *Nominal-The-Best* (N type) relationship.

In an L type relationship, the product characteristic has no pre-determined target value, but the larger this value is, the better it is. In manufacturing fields, this is the case with characteristics such as the strength of materials and fuel efficiency. Taguchi et al. (1988) operationalized the loss function for the L type as a relationship between the *tolerance* (the permissible variation of a product characteristic from its target value) and the amount of monetary loss incurred if the product characteristic is less than the lower tolerance limit,  $A$ . The loss function then becomes  $L(x) = \frac{A * \Delta^2}{x^2}$ , where  $x$  is the value of the characteristic and  $\Delta$  is the lower tolerance limit. In this case, the target or ideal value is  $m = +\infty$  (Taguchi et al., 1988, p. 34). This relationship, as pictured by Stvilia (2006) is shown in Figure 2.2. As can be seen from the graph, as the value of the characteristic increases, its quality increases, and the value loss decreases.

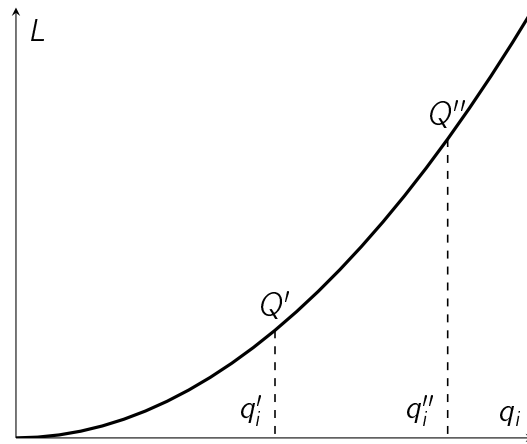


Figure 2.3. The smaller-the-better (S type) relationship curve. Adapted from “Measuring information quality” by B. Stvilia, 2006. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database: (Order No. 3223727).

In an S type relationship, a decrease in the value of a product characteristic causes the quality of a product to increase and its value loss to decrease. In this case, the ideal value of the product characteristic is zero. Examples of this relationship include characteristics such as wear, deterioration, and noise level. The loss function of an S type relationship is  $L(x) = (\frac{A}{\Delta^2})x^2$  where the target value is  $m = 0$  and  $\Delta$  is the upper tolerance limit (Taguchi et al., 1988, p. 33). Stvilia (2006)'s rendition of this relationship is shown in Figure 2.3. As the value of a product value increases, the value loss also increases, and its quality decreases.

The N type relationship describes a situation where a nominal value for a characteristic is preferred. Large deviations from the target value are undesirable and cause the value loss to increase and the quality to decrease. This is usually the case with characteristics such as dimension, clearance, and viscosity (Taguchi et al., 1988, p. 25). The loss in an N type relationship is defined as  $L(x) = k(x - M)^2$ , where the constant  $k$  is the loss  $A$  when  $y$  deviates from  $m$  by  $\Delta$ , or  $k = \frac{A}{\Delta^2}$  (Taguchi et al., 1988, p. 47). Figure 2.4 shows Stvilia (2006)'s rendition of the relationship. As can be seen, as the quality deviates from its target value, the loss increases.



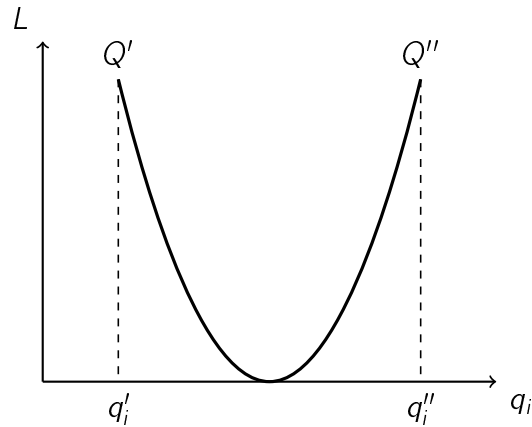


Figure 2.4. Nominal-the-best (N type) relationship curve. Adapted from “Measuring information quality” by B. Stvilia, 2006. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database: (Order No. 3223727).

## 2.5. IQ in Web Archiving

Problems with the quality of an archived website, or with an entire web archive, manifest themselves in a myriad of ways. The following section outlines the ways in which quality problems can negatively affect the appearance, functionality, and usability of archived websites. It also describes the Quality Assurance (QA) process that web archivists across many institutions undertake in order to detect and fix these quality problems. The problem of quality in web archives has been receiving an increasing amount of attention from research in the last few years, and the section finishes with an overview and discussion of research in this area.

### 2.5.1. Types and Severity of Quality Problems

The archived website shown in Figure 2.5 (University of North Texas, 2007b) is an example of what might be considered a “high-quality” archived website because it offers a good representation of what the original might have looked like. It contains all the visual elements of the original (colors, images, logos) as well as its intellectual elements (links, text, captions, etc). Furthermore, its functionality is much the same as the original. For example, clicking

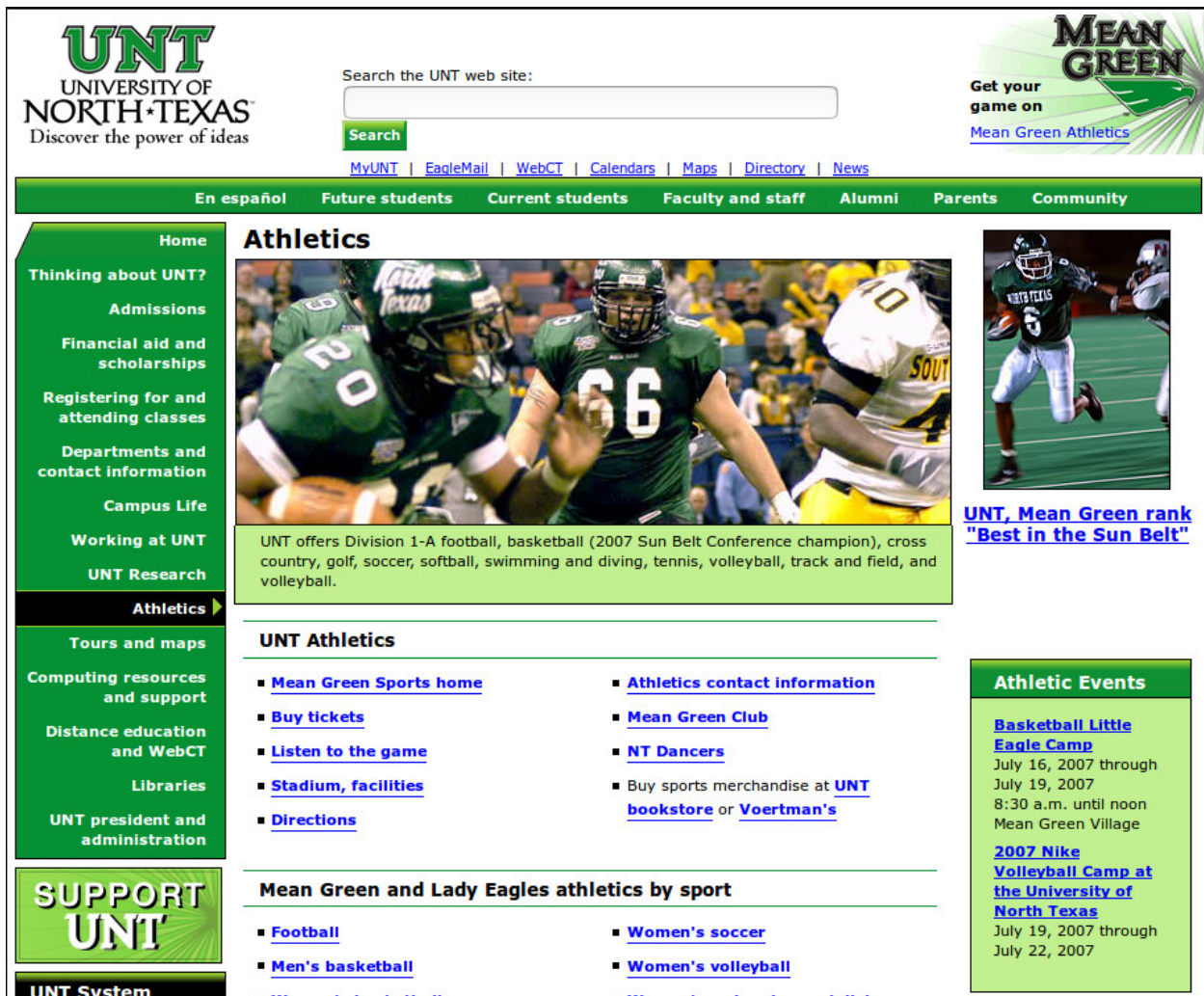


Figure 2.5. Screenshot of an archived version of the UNT Athletics site from 2007. The archived website seems to have reasonably good quality. Retrieved from <http://web.archive.org/web/20070716164831/https://www.unt.edu/athletics.htm>

on the “Buy tickets” link on the page leads users to the correct page containing ticket prices and information, while the “Athletics contact information” link leads to a page containing the address, email, and other contact information for the Athletics department.

The archived version of the UNT Admissions website on Figure 2.6 (University of North Texas, 2007a) is an example of a quality problem. The archived website is missing the visual elements of the original; it lacks the top banner, as well as the green menu on the left-hand side and other visual elements. Though the website might look different from the

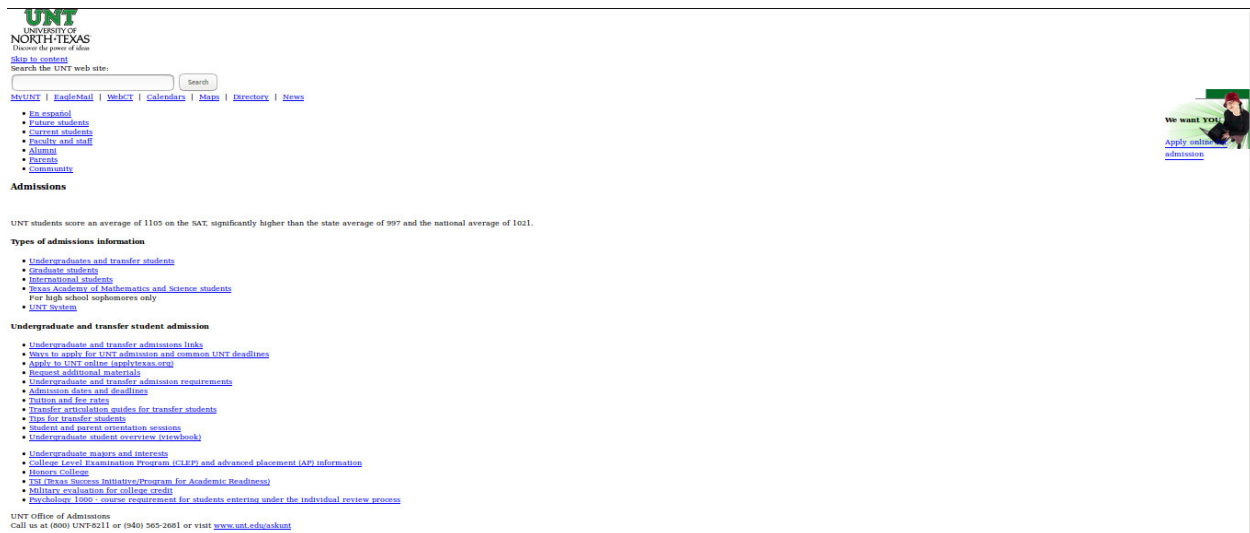


Figure 2.6. Screenshot of an archived version of the UNT Admissions site from 2007. The archived website is missing the styling of the original, but the intellectual information is still present. Retrieved from <http://web.archive.org/web/20070716164959/https://www.unt.edu/admissions.htm>

original, it retains most of its intellectual content in the form of text and links: clicking on any of the links will still lead the user to the corresponding page and its information. This archived website is still usable, though no longer a perfect copy of the original.

A more severe quality problem is shown in Figure 2.7 (University of North Texas, 2004). The archived website, from 2004, is supposed to contain an interactive map of the university campus; however, it is almost entirely blank. It contains no images, and clicking on the missing elements leads nowhere. In this example, the archived website is of such poor quality that it has been rendered virtually unusable. Arguably, a more severe quality problem than an unusable archived website is an entirely missing one, a not-uncommon occurrence in the world of web archiving. While a flawed archived website can be corrected or improved if the quality problem is caught in time, it is usually too late for a missing website.

Quality problems in an archived website are not always clearly visible to an end-user, but might take a more subtle form. One example is the serious problem of *leakage from*

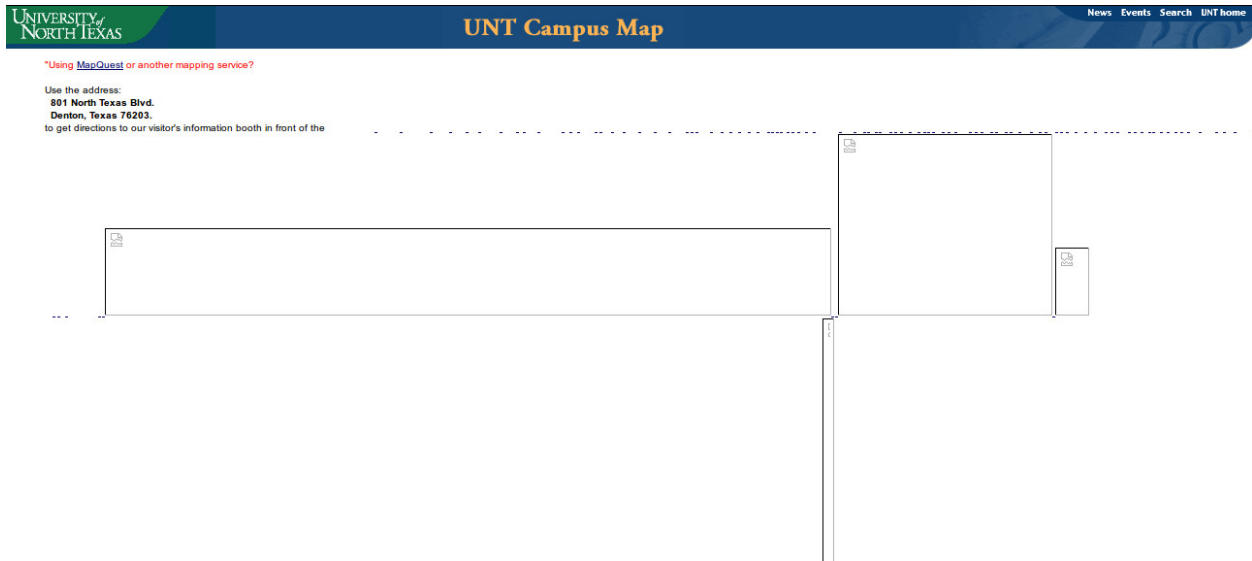


Figure 2.7. Screenshot of an archived version of the UNT Campus Map from 2004. The missing images negatively affect the quality of the archived map, rendering it unusable. Retrieved from <http://web.archive.org/web/20040722064240/http://www.unt.edu/pais/map/campusmap.htm> the live web, often referred to simply as *leakage*. Leakage occurs “when archived resources make requests to and include content from the live web when they should be accessing archived content only.” (Brunelle, Kelly, Weigle, & Nelson, 2015, p. 13). The end result is heterogenous website containing both archived content and content that is currently present on the live website. When this happens, the archived website ceases to become an accurate representation of a website as it was in the past. This phenomenon does not appear during the process of capturing a website, only during the replay process, when the archived website is being displayed in a client such as the Wayback Machine.

Though leakage is usually relatively benign, it can sometimes cause serious content incoherences. Figure 2.8 is one such example. It shows an archived web page from the CNN.com website in 2012. The article on the main page is about the 2008 Presidential Election; however, the content on the side bar, which is from a live web page, references the 2012 Presidential Election. Leakage usually goes undetected by the end users because they

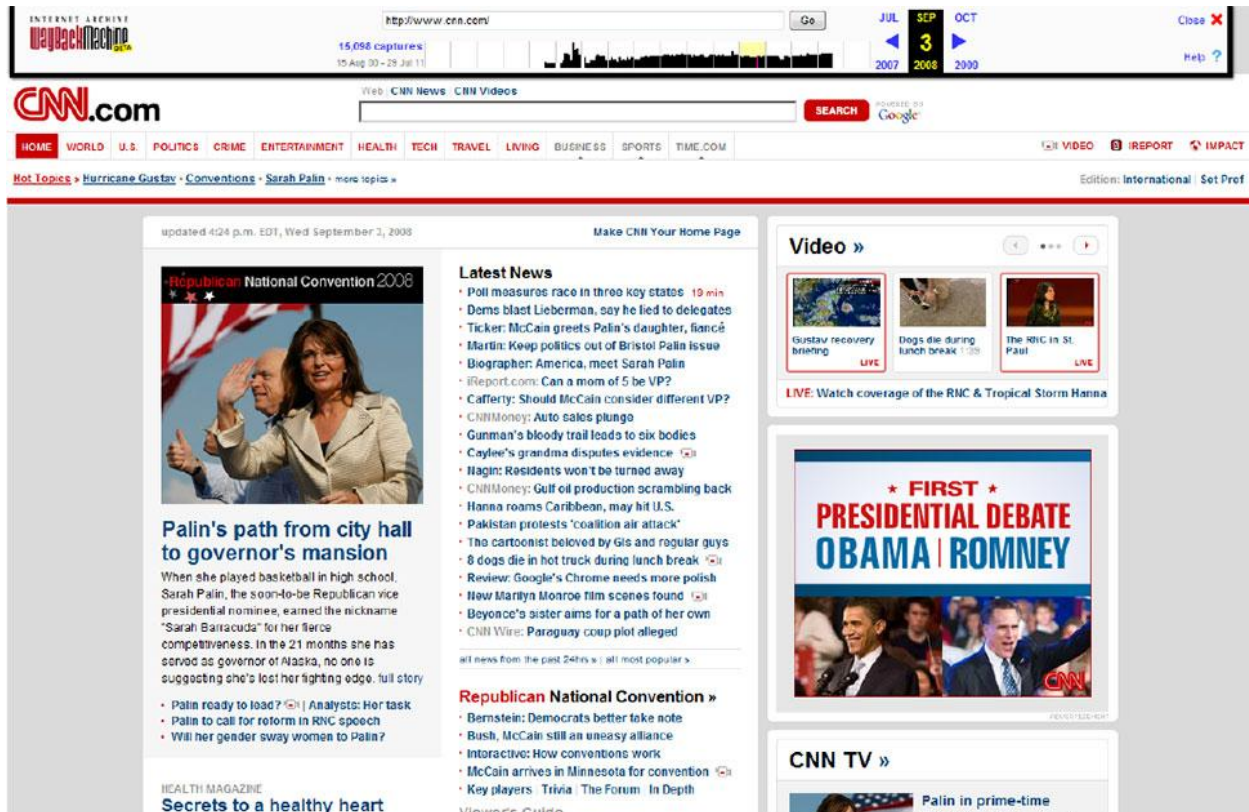


Figure 2.8. Screenshot of an archived version of the CNN.com website from 2012, showing leakage. The main article is about the 2008 Presidential Election; however, the content on the side bar references the 2012 Presidential Election. Adapted from “The impact of Javascript on archivability” by J. Brunelle, M. Kelly, M. Weigle, and M. Nelson, 2015, *International Journal on Digital Libraries*, 1-23. Retrieved from <http://dx.doi.org/10.1007/s00799-015-0140-8> doi: 10.1007/s00799-015-0140-8

are not familiar with the mechanics of web archiving. A trained web archivist can usually spot and prevent any leakage issues.

## 2.5.2. The Quality Assurance Process in Web Archiving

The process of archiving a website usually occurs in the following order:

- (1) Selection: During this phase, web archivists select the websites they are most interested in preserving.

(2) Acquisition/Capture/Harvest: A piece of software known as a “crawler” visits every resource to be captured, makes a copy of it, and stores it.

(3) Access: The institution provides access to the captured content.

(Reyes Ayala, 2013, p. 1)

In their survey of web archiving practices, (Reyes Ayala, Phillips, & Ko, 2014) identified quality as an important issue in web archiving and quality assurance (QA) as a process that almost all institutions undertake to ensure the high quality of their archived websites. The authors state that a “typical” QA process involves the following elements:

- QA is done after the sites are captured: QA is not a process that begins before the capture stage. Neither is it ongoing, rather, it is done once and at a discrete point in time, which is after the capture process.
- QA is done manually: This involves a person who looks at the archived version of the site and assesses its quality.
- View the site using the Wayback Machine: The most common method of assessing the quality of an archived website was by viewing it in the Internet Archive’s Wayback Machine.
- QA is done on every captured site. Also, the entire site is put through the QA process, not just the homepage or specific domains.
- Quality problems are noted, either in a spreadsheet or in another system such as a database.
- QA is done by the same person who implemented the crawl, such as a crawl operator or engineer. This suggests that web archiving teams throughout the world are small, and one person may be responsible for many different roles, such as determining what websites should be captured, launching the capture process, and checking the quality of a crawl. Relatively few institutions have dedicated QA staff.

(Reyes Ayala et al., 2014, p. 14)

The authors noted that the process of QA for web archives is an onerous one because it involves manually inspecting hundreds if not thousands of archived websites. This necessitates a significant time commitment from web archivists, and a specialized knowledge and skills (Reyes Ayala et al., 2014, p. 19). Other authors have pointed out additional difficulties. (Voorburg, 2010) stated that the definition of quality itself was confusing and that “it was difficult to define a ‘good enough’ level of quality for a captured resource”.

Some of these concerns were echoed by the results of the 2016 survey on web archiving in the United States, conducted by the NDSA. When asked what were their top concerns when developing a web archiving program at their respective institutions, 52% of participants cited quality as a top issue. Quality was the third most-cited concern for respondents, after cost and access and use (60% each) (Bailey et al., 2016, p. 13). Quality was also perceived by respondents as one of the areas where they had made the *least* progress in their programs over the last year. Based on these responses, Bailey et al. (2016) concluded organizations were increasingly concerned with the quality instead of with the volume of data, and that this indicated the field was beginning to mature.

### 2.5.3. Research on Quality in Web Archives

In the field of web archiving, a few researchers have recently begun to address the topic of quality for web archives. Some of the researchers have also attempted to operationalize individual aspects of quality and to create metrics to effectively measure it. Because finding out which aspects of web archive quality can be most successfully measured is one of the main goals of this dissertation, the research described in this section is covered in greater depth and detail than in other sections. Additionally, this section covers the methodologies employed by this research as much as it does the results. It should be noted that the research discussed in this section is very recent as of the time of the writing of this dissertation, and so its impact and influence has yet to be fully assessed.

### 2.5.3.1. The Notion of Coherence in a Web Archive

In their paper, Spaniol, Mazeika, Denev, and Weikum (2009) are primarily concerned with the data quality of web archives, specifically with the quality of a crawl, not with replay of the archived website itself. The authors introduce the concept of (temporal) coherence for a web archive. The contents of a web archive are considered to be coherent if they appear to be “as of” time point  $x$  or interval  $[x;y]$ . In a web archive, coherence defects can occur during the crawl, a process which can take anywhere from a few minutes to even weeks for large websites. Consider as an example a website with a hierarchical depth of 3. A crawler might begin by crawling the homepage of a website at time  $t_1$ , the first-level pages at time  $t_2$ , and the second-level pages at time  $t_3$ . However, by the time ( $t_3$ ) that the crawl concludes, the homepage of the website has changed its content, and so the final web archived website contains the most recent versions of the second and third-level pages, but an older version of the homepage. This is a coherence defect that can be particularly severe for large, constantly-changing websites such as news sites. Spaniol et al. (2009) explored ways to visualize coherence defects in a web archive, so that crawl engineers could detect them and adjust their crawling strategies accordingly.

In a later paper, Denev, Mazeika, Spaniol, and Weikum (2011) introduced the Sharp Archiving of Website Captures (SHARC) framework for data quality in web archiving. This framework included two measures of data quality for capturing websites: *blur* and *coherence*. Blur was defined as the expected number of page changes that a time-travel access to a site capture would accidentally see, instead of the ideal view of a instantaneously captured, “sharp” site. This value needed to be minimized in order to achieve a high-quality capture. The authors defined coherence as the number of unchanged and thus coherently captured pages in a site snapshot. Here, “unchanged” denotes pages that are definitely known to be invariant throughout some time window, ideally the entire crawl. Coherence needed to be maximized in order to achieve a high-quality capture.



Mathematically, these measures are defined in the following way below.

Let  $p_i$  be a web page captured at time  $t_i$ . The blur of the page,  $B$ , is the expected number of changes between  $t_i$  and query time  $t$ , averaged through the observation interval  $[0, n\Delta]$ :

$$(10) \quad B(p_i, t_i, n, \Delta) = \frac{1}{n\Delta} \int_0^{n\Delta} \lambda_i \cdot |t - t_i| dt = \frac{\lambda_i \omega(t_i, n, \Delta)}{n\Delta}$$

where the download schedule penalty:

$$(11) \quad \omega(t_i, n, \Delta) = t_i^2 - t_i n\Delta + \frac{n\Delta^2}{2}$$

Similarly, they defined the blur of an entire archived website (or an entire web archive) as the sum of the blur values of the individual pages. Let  $P = (p_0, \dots, p_n)$  be web pages captured at times  $T = (t_0, t_1, \dots, t_n)$ , then the blur of the website or web archive is:

$$(12) \quad B(P, T, n, \Delta) = \frac{1}{n\Delta} \sum_{i=0}^n \lambda_i \omega(t_i, n, \Delta)$$

Denev et al. (2011) presented several crawl strategies, including an algorithm, that if implemented would improve the quality of an archived website by minimizing blur and maximizing coherence.

The work of Ainsworth, Nelson, and Van de Sompel (2014) further expanded the notion of temporal coherence in a web archive. They pointed out that archived web pages are composite objects. Initially, a user might elect to browse an archived website (which they call the root resource) dating from November 1, 2010; however, because of the constantly changing nature of the web, many elements and pages from the archived website will have been collected before or after the November 2010 date. The final, archived website presented to the user via the Wayback Machine is often a patchwork collection of HTML pages, images,

and scripts from different dates and is thus temporally incoherent. An example of this phenomenon is shown in Figure 2.9, which shows an archived version of [wunderground.org](http://wunderground.org) from December 9, 2004. According to the authors' investigation, the components or embedded resources on this page are from different times: the logo on the upper left-hand corner was captured 15 hours before the page, the five-day forecast content was captured nine hours after the page, the "Nowcast" information was captured 20 days before the page, and the city conditions information captured 17 hours before the page. Most notably, the satellite image was captured *a full nine months after* the initial page. This creates not only a temporal incoherence but also a *content incoherence*, since the clear satellite image contrasts with the chance of rain and mostly cloudy conditions shown elsewhere on the page (Ainsworth et al., 2014, p.1).

The authors note two important things:

- (1) Even if captured within seconds of the root resources, embedded resources are not always temporally coherent.
- (2) Even if captured much later than the root resource, embedded resources are not necessarily incoherent.

(Ainsworth et al., 2014, p.2)

They defined the *temporal coherence* of an archived website (which they call a memento) in the following way, "an embedded memento [is] temporally coherent with respect to a root memento when it can be shown that the embedded memento's representation existed at the time the root memento was captured" (Ainsworth et al., 2014, p.3). Also the *temporal spread* is the difference between the earliest and latest date-times in a composite memento. The authors presented five different temporal coherence states: Prima Facie Coherent, Prima Facie Violative, Possibly Coherent, Possibly Violative, and Coherence Undefined. These states have the following definitions:

- (1) **Prima Facie Coherent (C)**: The embedded memento existed in its

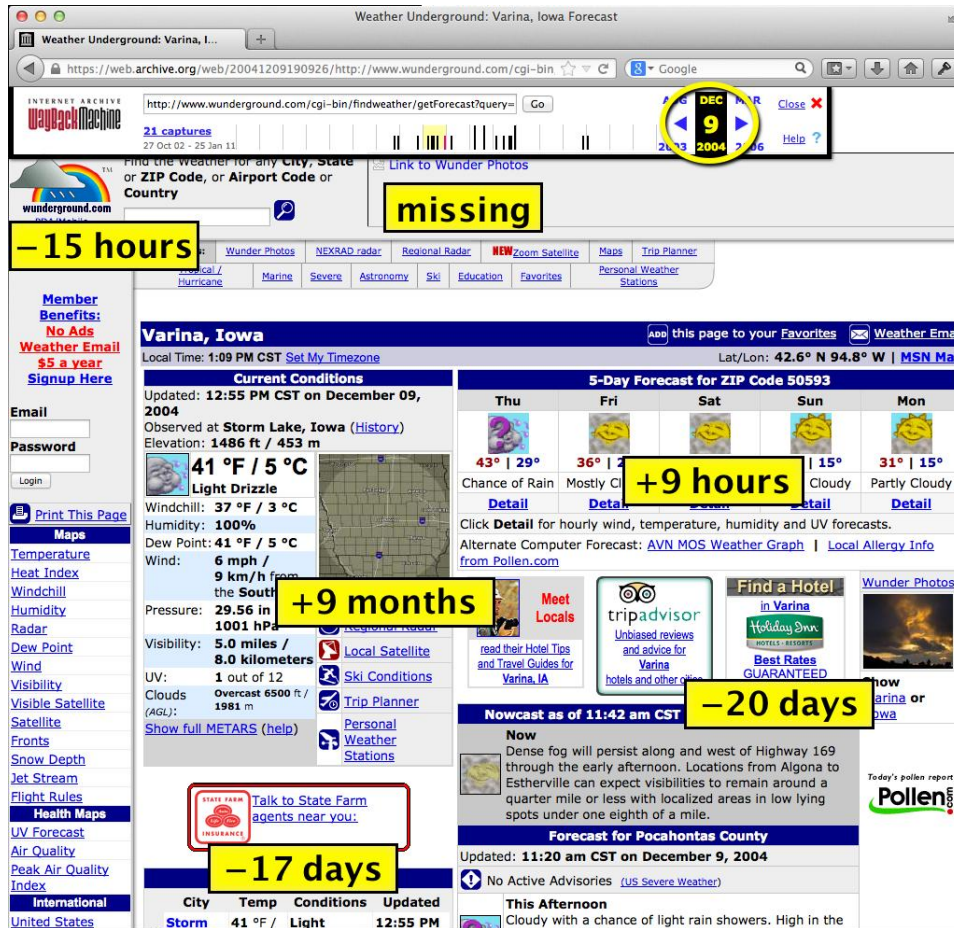


Figure 2.9. An archived version of wunderground.org from December 9, 2004. The main page itself was captured on that day; however, its embedded resources were captured hours and even days before and after the page itself. Adapted from “A Framework for Evaluation of Composite Memento Temporal Coherence” by S. Ainsworth, M. Nelson, and H. Van de Sompel, 2014, *Computing Research Respository (CoRR)*, [abs/1402.0928](http://arxiv.org/abs/1402.0928), p.1. Retrieved from <http://arxiv.org/abs/1402.0928>

archived state at the time the root memento was captured.

- (2) **Prima Facie Violative (V)**: The embedded memento did not exist in its archived state at the time the root memento was captured.
- (3) **Possibly Coherent (PC)**: The embedded memento could have existed in its archive state at the time the root memento was captured.

- (4) **Probably Violative (PV)**: The embedded memento probably did not exist in its archived state at the time the root memento was captured.
  - (5) **Coherence Undefined (CU)**: There is not enough information to determine coherence state.
- (Ainsworth et al., 2014, p.3)

Each possible condition is the result of a specific set of attributes. These are the Memento-Datetime (the date and time of capture) of both the root  $T_0$  and of the embedded memento  $T_{i,j}$ , as well as the Last-Modified datetime (the date and time it was last modified) of the root  $L_0$  and of the embedded memento  $L_{i,j}$ . Undefined data is denoted with an up arrow ( $\uparrow$ ), while a down arrow means the attribute is defined ( $\downarrow$ ). The set of these possible coherence states is shown in Table 2.6.

As can be seen from the table, Ainsworth et al. (2014) described in detail the many different conditions that occur during the complex process of web archiving. When addressing the undefined patterns from the table, Right Undefined Last-Modified and Left Undefined Last-Modified, the authors stated that missing Last-Modified datetimes normally indicated a dynamically-generated embedded element. These resources, such as result pages from queries and social media elements, do not exist on a web page until they are generated on demand, often by a user's action. Since dynamic elements do not exist at the time the root was captured, any associated patterns are classified as probably violative. In the case of Left Last-Modified, the existence of Last-Modified indicates the embedded memento was probably not dynamically generated, but this has no effect on the coherence, and so the pattern is classified as probably coherent.

Ainsworth et al.(2014) also specified an extension of their defined coherence states that involved calculating the similarity, or lack thereof, between two archived versions of the same website (or as the authors put it, between two mementos). This comparison, which they called a "content pattern", takes into account not just the time of archival (the Memento-

Table 2.6

*Pattern Groups and their Coherence States*

<b>Pattern Type</b>			<b>Definition</b>	<b>Coherence State</b>	<b>Predicate</b>
Right Bracket			The embedded memento's Memento-Datetime is after the root's and its Last-Modified datetime is on or before the root's Memento-Datetime.	C	$L_{i,1} \downarrow \wedge (L_{i,1} \leq T_0 \leq T_{i,1}) \Rightarrow C$
Right Modified	Newer	Last-	The embedded memento's Memento-Datetime and Last-Modified datetime are both later than the root's Memento-Datetime. This evidence indicates that the embedded memento was modified after the root memento.	V	$L_{i,1} \downarrow \wedge (T_0 < L_{i,1} \leq T_{i,1}) \Rightarrow V$
Right Modified	Undefined	Last-	The embedded memento's Memento-Datetime is after the root's but the Last-Modified is undefined.	PV	$L_{i,1} \uparrow \wedge (T_0 < T_{i,1}) \Rightarrow PV$

Continued on next page

**Table 2.6 – continued from previous page**

<b>Pattern Type</b>	<b>Definition</b>	<b>Coherence State</b>	<b>Predicate</b>
Left Last-Modified	The embedded embedded memento's datetime is before the root; however, the existence of Last-Modified indicates the embedded memento was probably not dynamically generated.	PC	$L_{i,n} \downarrow \wedge (T_{i,n} < T_0) \Rightarrow PC$
Left Undefined Last-Modified	The embedded memento's datetime is before the root, but the Last-Modified is undefined.	V	$L_{i,n} \uparrow \wedge (T_{i,n} < T_0) \Rightarrow V$
Simultaneous Capture	Embedded memento was captured simultaneously with the root.	V	$T_0 = T_{ij} \Rightarrow V$

*Note.* Adapted from “A Framework for Evaluation of Composite Memento Temporal Coherence” by S. Ainsworth, M. Nelson, and H. Van de Sompel, 2014, *Computing Research Repository (CoRR)*, *abs/1402.0928*, p.1. Retrieved from <http://arxiv.org/abs/1402.0928>

Datetime), but also the content of the two mementos in order to determine coherence. For example, two mementos can be compared:  $m_{i,j-1}$ , archived before root and  $m_j$ , archived after the root memento. For each comparison, the two mementos can be equal in content ( $m_{i,j-1} = m_{i,j}$ ), similar ( $m_{i,j-1} \sim m_{i,j}$ ), or not similar ( $m_{i,j-1} \not\sim m_{i,j}$ ). These coherence states are described in Table 2.7. It is important to note that according to the authors, the additional computational cost of calculating these comparisons “may render content patterns unsuitable for casual archive use or in restricted bandwidth conditions” (Ainsworth et al., 2014, p.6).

Table 2.7

*Pattern Groups, their Content Patterns, and their Coherence States*

<b>Pattern Type</b>	<b>Definition</b>	<b>Coherence State</b>	<b>Predicate</b>
Content Equal Bracket	$m_{i,j}$ has a Last-Modified datetime that is on or before the root's capture time. The two embedded mementos are equal.	C	$L_{i,j} \downarrow \wedge (T_{i,j-1} < L_{i,1} < T_0 < T_{i,1}) \wedge (m_{i,j-1} = m_{i,j}) \Rightarrow C$
Content Equal Newer Last-Modified	$m_{i,j}$ has a Last-Modified datetime that is after the root's capture time. The two embedded mementos are equal.	C	$L_{i,1} \downarrow \wedge (T_{i,j-1} < T_0 < L_{i,1} < T_{i,1}) \wedge (m_{i,j-1} = m_{i,j}) \Rightarrow C$
Content Equal Undefined Last-Modified	$m_{i,j}$ has an undefined Last-Modified datetime. The two embedded mementos are equal.	C	$L_{i,1} \uparrow \wedge (T_{i,j-1} < T_0 < T_{i,1}) \wedge (m_{i,j-1} = m_{i,j}) \Rightarrow C$
Content Similar Bracket	$m_{i,j}$ has a Last-Modified datetime that is on or before the root's capture time. The two embedded mementos are similar.	C	$L_{i,j} \downarrow \wedge (T_{i,j-1} < L_{i,1} < T_0 < T_{i,1}) \wedge (m_{i,j-1} \sim m_{i,j}) \Rightarrow C$

Continued on next page



**Table 2.7 – continued from previous page**

<b>Pattern Type</b>	<b>Definition</b>	<b>Coherence State</b>	<b>Predicate</b>
Content Similar Newer Last-Modified	$m_{i,j}$ has a Last-Modified datetime that is after the root's capture time. The two embedded mementos are similar.	C	$L_{i,1} \downarrow \wedge (T_{i,j-1} < T_0 < L_{i,1} < T_{i,1}) \wedge (m_{i,j-1} \sim m_{i,j}) \Rightarrow C$
Content Similar Undefined Last-Modified	$m_{i,j}$ has an undefined Last-Modified datetime. The two embedded mementos are similar.	C	$L_{i,1} \uparrow \wedge (T_{i,j-1} < T_0 < T_{i,1}) \wedge (m_{i,j-1} \sim m_{i,j}) \Rightarrow C$
Content Not Similar Bracket	$m_{i,j}$ has a Last-Modified datetime that is on or before the root's capture time. The two embedded mementos have different content.	C	$L_{i,j} \downarrow \wedge (T_{i,j-1} < L_{i,1} < T_0 < T_{i,1}) \wedge (m_{i,j-1} \not\sim m_{i,j}) \Rightarrow C$
Content Not Similar Newer Last-Modified	$m_{i,j}$ has a Last-Modified datetime that is after the root's capture time. The two embedded mementos have different content.	V	$L_{i,1} \downarrow \wedge (T_{i,j-1} < T_0 < L_{i,1} < T_{i,1}) \wedge (m_{i,j-1} \not\sim m_{i,j}) \Rightarrow V$
Content Not Similar Undefined Last-Modified	$m_{i,j}$ has an undefined Last-Modified datetime. The two embedded mementos have different content.	PV	$L_{i,1} \uparrow \wedge (T_{i,j-1} < T_0 < T_{i,1}) \wedge (m_{i,j-1} \not\sim m_{i,j}) \Rightarrow PV$

*Note.* Adapted from “A Framework for Evaluation of Composite Memento Temporal Coherence” by S. Ainsworth, M. Nelson, and H. Van de Sompel, 2014, *Computing Research Respository (CoRR)*, *abs/1402.0928*, p.1. Retrieved from <http://arxiv.org/abs/1402.0928>

In Table 2.7, all of the patterns that involve equality or similarity are classified as prima facie coherent (C). This indicates that equality and similarity of content between two mementos tends to override any evidence we may have about the Last-Modified datetime. In the paper it was also noted that similarity was a subjective term, “while the definition of equality is universal, the definition of similar will vary by application and user need” (Ainsworth et al., 2014, p.6). Though the authors use the notion of similarity, a clear definition of the term was never put forward. Additionally, the definitions put forward by the authors have significant limitations. Determining the coherence of two mementos requires a web archivist to have access to the crawl logs generated during the capture process, which contain technical details such as the time of capture and the time a website was last modified. This type of information may not be available to a web archivist, and in cases where an organization is using a paid subscription service such as Archive-It, it will almost certainly not be available.

Ainsworth and Nelson (2015) were also concerned with defining quality as meeting measurable characteristics. Their work elaborates on the notion of coherence put forward by Denev et al. (2011). They equate the completeness of a web archive to its coverage, in other words, a complete web archive does not have undesired or undocumented gaps. They adopted the definition of temporal coherence presented by Denev et al. (2011) and introduced a new characteristic of it: drift.

They defined *drift* as the difference between the target date-time originally required by the user and the actual date-time returned by an archive. Drift can be forwards or backwards in time, and occurs when a user navigates an archived website. Initially, a user might elect to browse an archived website dating from November 1, 2010 (the target date-time). However, because of the constantly changing nature of the web, many elements and pages from the archived website will have been collected before or after the November 2010 date. The final, archived website presented to the user via the Wayback Machine can become a patchwork collection of HTML pages, images, and scripts from different dates, thus losing its resem-

blance to the original. Ainsworth and Nelson (2015) found that during the browsing process the target date-time changes with each link followed and eventually “drifts” away from the date-time originally selected, they noted that “when browsing sparsely-archived pages, this nearly-silent drift can be many years in just a few clicks.” (Ainsworth & Nelson, 2015, p. 129).

Other researchers have addressed the notion of completeness in a web archive. Web archives do not contain complete and perfectly accurate copies of every single website they intend to capture; the dynamic nature of the web makes this almost technically impossible. However, as seen in Section 2.5.1, not all missing elements are created equal. Many archived websites are missing elements but still retain most of their intellectual content, while other archived websites, such as maps, are rendered unusable due to missing elements. Brunelle, Kelly, SalahEldeen, Weigle, and Nelson (2015) made precisely this point when they examined the importance of missing elements (which they call “resources”) and their impact on the quality of archived websites in their paper “Not all mementos are created equal: measuring the impact of missing resources” (p. 1-19).

When deploying crawlers to capture a website, some crawl engineers pay special attention to embedded resources. Embedded resources are files, such as images, videos, or CSS stylesheets, that are present and referenced in a website. In many cases, such as for CSS stylesheets, a user might not notice their presence, but embedded resources play a key role in ensuring the website looks and operates in the correct way. To this end, crawl engineers might calculate a percentage of missing embedded resources  $M_m$  in an archived website, and use it to estimate the overall quality of the site. Brunelle, Kelly, SalahEldeen, et al. (2015) showed that  $M_m$  is not always consistent with human judgments of the quality of an archived website and was thus not a suitable metric for measuring the “damage” to an archived website caused by missing embedded resources. Instead, the authors proposed a new metric to assess this damage that is based on three factors: the MIME type, size, and location of the embedded resource (Brunelle, Kelly, SalahEldeen, et al., 2015, p. 5).

$$\begin{aligned}
D_{[I|MM]} &= 1 + \frac{width * height}{Page\ Size(pixels)} \\
&+ (w_{horizontal} \iff Overlaps\ horizontal\ center) \\
(13) \quad &+ (w_{vertical} \iff Overlaps\ vertical\ center) \\
&w_{horizontal} = 0.25 \\
&w_{vertical} = 0.25
\end{aligned}$$

They define the set of all embedded resources  $R$  and all missing embedded resources  $R_r$  in Equation 15. The authors focus on three types of embedded resources: images, multimedia elements, and stylesheets, and calculate their importance in terms of the possible damage caused to the archived website if these were missing. The importance of both missing images  $D_I$  and missing multimedia elements  $D_{MM}$  is measured in terms of its size and centrality (positioning) in the original website.  $D_I$  and  $D_{MM}$  are defined in Equation 13.

Damage for missing style sheets:

$$\begin{aligned}
D_C &= 1 + w_{style} \iff \\
&(> 75 + (w_{tags} \iff \\
(14) \quad &(tags\ in\ the\ DOM\ without\ matching\ CSS) \\
&w_{style} = 0.5 \\
&w_{tags} = 0.5
\end{aligned}$$

Equation 14 shows the importance of missing stylesheets  $D_C$ . The variable  $w_{style}$  is a threshold based on calculations of background-colored and non-background-colored pixels on the archived web page. If more than 75% of the non-background-colored are in the left two-thirds of the page and a stylesheet is missing, then the authors assume the missing stylesheet was important and  $w_{style} = 0.5$ . Similarly for  $w_{tags}$ , a style threshold, “the presence of tags

on the page without a matching style suggests that the missing CSS contained the referenced formatting” (Brunelle, Kelly, SalahEldeen, et al., 2015, p. 6). For this case,  $w_{tags} = 0.5$ .

$$\begin{aligned}
 R &= \{All\ embedded\ resources\ requested\} \\
 R_r &= \{All\ missing\ embedded\ resources\} \\
 R_r &\subseteq R
 \end{aligned}
 \tag{15}$$

$$D_m = \frac{D_{m_{actual}}}{D_{m_{potential}}}
 \tag{16}$$

The authors define set of all embedded resources  $R$  and of all missing embedded resources  $R_r$ , as in Equation 15. After discussing the damage to the archived website caused by missing images, multimedia elements, and CSS stylesheets, they define  $D_m$  as the damage rating (or cumulative damage) of an archived website caused by missing embedded resources, expressed as the ratio of actual damage to potential damage. They define potential damage as the “cumulative importance of all embedded resources in the [archived website], while actual damage is only the importance of those embedded resources that are unsuccessfully dereferenced, or missing.”(Brunelle, Kelly, SalahEldeen, et al., 2015, p. 5). The formula for  $D_m$  is show in Equation 16.

$$\begin{aligned}
 D_{m_{potential}} &= \frac{\sum_{i=1}^{n_{[I,MM]}} D_{[I|MM]}(i)}{n_{[I|MM]}} + \frac{\sum_{i=1}^{n_C} D_C(i)}{n_C} \\
 \forall \{I = Images, MM = Multimedia, C = CSS\} \\
 n &\in R
 \end{aligned}
 \tag{17}$$

The potential damage  $D_{m_{potential}}$  is the sum of the importance of each embedded resource, as shown in Equation 17. The formula for the actual damage  $D_{m_{actual}}$  is the same as that for  $D_{m_{potential}}$ , with the exception that it is computed over the set of missing embedded

resources  $R_r$ . Brunelle, Kelly, SalahEldeen, et al. (2015) discovered that their formula for  $D_m$  was more consistent with human judgments of the quality of archived websites than the original metric  $M_m$ . Furthermore, they used their metric  $D_m$  to evaluate the performance of the Internet Archive's web archiving capabilities. According to the authors, most websites archived by the Internet Archive were missing few embedded resources (less than 10) and that the average yearly  $D_m$  dropped from 0.16 in 1998 to 0.13 in 2013 (Brunelle, Kelly, SalahEldeen, et al., 2015, p. 8). This suggested that the Internet Archive was doing a better job over time in reducing the number of missing embedded resources; however, the authors also found that the number of missing, *important* embedded resources was increasing over time. The Internet Archive is "missing an increasing number of embedded resources deemed important." (Brunelle, Kelly, SalahEldeen, et al., 2015, p. 9)

AlNoamany, Weigle, and Nelson (2015) also addressed quality problems that could affect the coherence of a web archive, such as off-topic web pages. Many web archives, such as those created using the Internet Archive's Archive-It service, are topic-specific, they collect and preserve many websites that cover a single topic or news event, such as Human Rights or the Arab Spring of 2010. Off-topic web pages are defined as those that have, over time, moved away from the initial scope of the page. This can occur because the page has been hacked, its domain has expired, or the service has been discontinued (AlNoamany et al. (2015, p. 226). The authors compiled three different Archive-It collections and experimented with several methods of detecting these off-topic webpages and with how to define threshold that separates the on-topic from the off-topic pages. This involved comparing the text (after pre-processing, stemming and stopword removal) of the archived website when it was first captured ( $URI - R@t_0$ ) with the text archived website that was captured at a later time ( $URI - R@t$ ). The methods tested were the following:

- (1) Cosine similarity: applied the cosine similarity formula to  $URI - R@t_0$  and  $URI - R@t$ .
- (2) Jaccard similarity coefficient: computed the size of the intersection between  $URI -$

$R@t_0$  and  $URI - R@t$ , divided by the size of their union.

- (3) TF-Intersection: compared the intersection of the top 20 most frequent terms of the  $URI - R@t_0$  with the top 20 most frequent terms of the  $URI - R@t$ .
- (4) Web-based kernel function: augmented the first five words of both  $URI - R@t_0$  and  $URI - R@t$  with additional terms from the web to increase the semantic context. Calculated the Jaccard similarity coefficient between these two new, expanded term lists.
- (5) Word count: compared the number of words in  $URI - R@t_0$  with the number of words in  $URI - R@t$ .
- (6) Change in size: compared the change in size, measured in bytes, of  $URI - R@t_0$  and  $URI - R@t$ .

(AlNoamany et al., 2015, p. 230)

According to their results, the cosine similarity method proved the best at detecting off-topic web pages, with an average accuracy of 0.983, and F-measure (harmonic mean of precision and recall) of 0.881, and an Area Under the Curve (AUC) measure of 0.961. The second-best performing measure was word count. The author also experimented with combining several similarity measures in an attempt to increase performance. The combination of the cosine similarity and word count methods yielded the best results, with an accuracy equal to 0.987,  $F = 0.906$ , and  $AUC = 0.968$  (AlNoamany et al., 2015, p. 234).

AlNoamany et al. (2015) never explicitly state what their definition of coherence is, but their approach differs markedly from other research that has been discussed in prior sections. The author are not concerned with temporal incoherence, as Denev et al. (2011) and Ainsworth et al. (2014) are, but with the topical coherence of archived web pages as compared to the rest of the web archive. As it is, their implicit notion of coherence is more similar to the notion of cohesiveness put forward by Zhu and Gauch (2000) and Consistency/coherence quality dimension described by Batini et al. (2012). Nevertheless, their work points out that

there might be more than one type of coherence in a web archive, for example:

- (1) Temporal coherence, as described by Denev et al. (2011) and Ainsworth et al. (2014).
- (2) Topical coherence, which is composed of:
  - (a) Archived web pages that were *once* coherent with the rest of the web archive, and have ceased to be coherent because of hacking, service discontinuation, etc, as described by AlNoamany et al. (2015).
  - (b) Archived web pages that were captured but were *never* coherent with the rest of the web archive. For example, a web archive on the topic of global climate change that contains large amounts of pornography. This case has yet to be explored in depth.

Furthermore, the research described in this section emphasizes coherence and completeness *only during the capture process*. However, as seen in Chapter 3 and 4, not all quality problems in a web archive occur during capture. Many quality problems arise as a result of the replay process because current technologies such as the Wayback Machine are unable to adequately render the archived website as it originally appeared. Since most organizations carry out Quality Assurance *after* capture has taken place (as seen in section 2.5.2), a quality problem might not even be detected at the time of capture, but much later.

#### 2.5.3.2. The Notion of Archivability

In their iPres paper “CLEAR: A Credible Method to Evaluate Website Archivability”, Banos, Kim, Ross, and Manolopoulos (2013) introduced the concept of website archivability. Archivability was defined as the “sum of the attributes that make a website amenable to being archived” (Banos et al., 2013, p. 1). The more easily it was to archive a website, the greater its archivability. The authors introduced a set of facets designed to determine the archivability of a website, termed the Credible Live Evaluation of Archive Readiness, or CLEAR, method. These facets were: standards compliance, performance, cohesion. and metadata us-



age. Later the authors expanded on their original work by introducing the CLEAR+ method, the incremental evolution of their original CLEAR+ method. According to CLEAR+, The archivability of a website is dependent on the following facets:

- Accessibility ( $F_A$ ): the ease with which a web crawler can visit a site, traverse its entirety and retrieve it via standard HTTP protocol requests. The website should provide resources so that a web crawler can discover and retrieve its different components (such as individual pages, images, and scripts). This facet also includes performance, or the speed at which a crawler can access the site.
- Standards Compliance ( $F_S$ ): the website and its individual components conform to common accepted technical standards. For example, its HTML pages, conform to the W3C standards for HTML. It is also important that the website provided content in open file formats, instead of closed, proprietary formats such as QuickTime and Flash.
- Cohesion ( $F_C$ ): the website does not have components that are dispersed across different locations on the web. For example, images, JavaScript files, and widgets.
- Metadata Usage ( $F_M$ ): the website contains descriptive metadata such as HTTP headers and HTML META headers. It is important to note that the authors do not commit to a specific metadata model, but recommend using widely-accepted metadata models such as the Dublin Core standards.

(Banos & Manolopoulos, 2015)

Each of these facets has several components, or criteria, each with its own significance. Criteria with high significance are more important to the archivability of a website, and if they are not met, can cause problematic web archiving results or even prevent the website from being archived at all. Medium-significance criteria are not critical but are still important, while low-significance criteria are considered minor issues. The full list of CLEAR+ archivability facets and their components is shown in Table 2.8.

Table 2.8

*Facets of Archivability and their Components.*

<b>Facet</b>	<b>Components</b>	<b>Significance</b>
Accessibility	Percentage of valid vs. invalid hyperlink and CSS urls	High
	Presence of inline JavaScript code exists in HTML	High
	Presence of sitemap.xml file	High
	Max initial response time of all HTTP requests	High
	Usage of proprietary file format such as Flash and QuickTime	High
	Presence of "Disallow:" rules in robots.txt file	Medium
	Presence of "Sitemap:" rules in robots.txt file	Medium
	Percentage of downloadable linked media files	Medium
	Presence of HTTP Caching headers such as Expires, Last-modified or ETag	Medium
	Presence of RSS or Atom feeds in the HTML source code	Low
Standards Compliance	HTML source code complies with W3C standards	High
	Usage of proprietary file formats such as QuickTime and Flash	High
	Integrity and standards of images	Medium
	RSS feed format complies with W3C standards	Medium
	HTTP Content-encoding or Transfer-encoding headers are set	Medium
	Presence of HTTP Caching headers such as Expires, Last-modified or ETag	Medium
	The CSS referenced in the HTML source code complies with W3C standards	Medium
	Integrity and standards compliance of HTML5 Audio elements	Medium

**Table 2.8 – continued from previous page**

<b>Facet</b>	<b>Components</b>	<b>Significance</b>
	Integrity and the standards compliance of HTML Video elements	Medium
	Presence of HTTP Content-type header	Medium
Cohesion	Percentage of local vs. remote images	Medium
	Percentage of local vs. remote CSS files	Medium
	Percentage of local vs. remote script tags	Medium
	Percentage of local vs. remote video elements	Medium
	Percentage of local vs. remote audio elements	Medium
	Percentage of local vs. remote proprietary objects such as Flash and QuickTime files	Medium
Metadata	Presence of HTTP Content-type header	Medium
	Presence of HTTP Caching headers such as Expires, Last-modified or ETag	Medium
Usage	Presence of the metadata tags <i>robots noindex, nofollow, noarchive, nosnippet</i> and <i>noodp</i> in the HTML source code	Low
	Presence of the Dublin Core (DC) profile in the HTML source code	Low
	Presence of the Friend of a Friend (FOAF) profile in the HTML source code	Low
	Presence of the HTML meta description tags in the HTML source code	Low

*Note.* Adapted from “A quantitative approach to evaluate website archivability using the CLEAR+ method” by V. Banos and Y. Manolopoulos, 2015, *International Journal on Digital Libraries*. doi: 10.1007/ s00799-015-0144-4

Banos and Manolopoulos (2015) stated that a website's archivability (WA) can be computed by using the sum total of its score for each facet: accessibility, standards compliance, cohesion, and metadata usage. As shown in Equation 18, the value of each facet is the weighted average of its coordinates. The website has a score for each facet, represented as a tuple  $(x_1, \dots, x_k, \dots, x_N)$ . The value of  $x_k$  is either 0 or 1, which represents a negative or positive answer to a specific criterion. The components of a single facet are not weighted evenly, but are assigned a weight ( $\omega_k$ ) depending on their significance. For high-significance components,  $\omega_k = 4$ , while  $\omega_k = 2$  for medium-significance components and  $\omega_k = 1$  for low-significance components. These weighted scores are then divided by  $C$  (Equation 19) to average them.

$$(18) \quad F_\lambda = \sum_{k=1}^N \frac{\omega_k x_k}{C}$$

$$(19) \quad C = \sum_{i=1}^N w_i$$

Once the value for each facet has been calculated, the total archivability score for the website can also be calculated, as shown in Equation 20.  $F_A$ ,  $F_S$ ,  $F_C$ , and  $F_M$  represent the value of each facet with respect to accessibility, standards compliance, cohesion, and metadata usage.

$$(20) \quad WA = \sum_{\lambda \in \{A, S, C, M\}} w_\lambda F_\lambda$$

Banos and Manolopoulos (2015) created ArchiveReady, an evaluation system that implements the CLEAR+ model as a web application. To evaluate a website's archivability, a user can navigate to the ArchiveReady website, and type its URL (Banos, 2012). ArchiveReady

will then calculate the websites's archivability and present it to the user in terms of a percentage. For example, a website may be classed as having a 62% archivability rating. The application will also output the website's scores for accessibility, cohesion, metadata usage, and standards compliance.

Once they had created and implemented their metrics, the authors proceeded to evaluate their validity by investigating the correlation between websites' WA scores as computed by the CLEAR+ system and human experts' judgments of a website's archivability. They observed that the correlation between a WA scores and expert's rating was 0.516 or 51.6%, with  $F_A$ ,  $F_C$ ,  $F_M$ , and  $F_S$  having individual correlation scores of 0.38, 0.26, 0.28 and 0.18 respectively (Banos & Manolopoulos, 2015, p. 20). They also reported the results of a one-way Analysis of Variance test (ANOVA), which yield an F-value of 397.628 and  $p = 2.191e_{-54}$ . On this basis, Banos and Manolopoulos (2015) concluded that CLEAR+ is a valid, reliable method of evaluating website archivability.

An ANOVA test is used to determine if the means of three or more populations are equal. (Hinkle, Wiersma, & Jurs, 2002, p. 334). The author's ANOVA results show that the means of the four different facets ( $F_A$ ,  $F_C$ ,  $F_M$ , and  $F_S$ ) are different and statistically significant; however the authors never clarify which facets are most responsible fo this variance. The standard statistical procedure is to report a complete summary of the ANOVA test, including the sums of squares, degrees of freedom, and mean squares, which the authors do not do (Hinkle et al., 2002, p. 348).

The authors' efforts are laudable in that they carefully examined the many complexities of archiving websites, focused on an important aspect (website archivability), and formulated clear metrics to help measure it. They also implemented an easy-to-use system, ArchiveReady, that might be of great help to web archivists around the world. However, their methodological approach and interpretation of results has some errors. Examining the correlation coefficients for the individual facets, one can find that the strength of this correla-

tion comes mostly from  $F_A$ , accessibility, which has a 0.38 correlation with human judgment. Accessibility accounts for 74.4% of the correlation, with  $F_C$ ,  $F_M$ , and  $F_S$  all exhibiting very weak correlations.  $F_S$  (standards compliance) has a particularly weak correlation of 0.18.

These weak correlations point to fundamental problems in how the metrics were formulated and to issues with the underlying concepts themselves. This becomes more evident when taking a look at Banos and Manolopoulos (2015)'s definitions of the facets of website archivability, seen in Table 2.8. The component "Usage of proprietary file format such as Flash and QuickTime" is present in both the Accessibility and Standards Compliance facets. It is also mentioned in the Cohesion facet, though this time formulated as the percentage of remote vs. remote proprietary files. Though the duplication of this component shows that it can belong to one or more facets of archivability, when operationalized, it creates an issue of "measuring the same thing twice". The same issue occurs with the components "Presence of HTTP Caching headers such as Expires, Last-modified or ETag" and Presence of HTTP Content-type header. The former is present in the Accessibility, Standards Compliance, and Metadata Usage facets, while the latter is present in both the Standards Compliance and Metadata Usage facets. Because of their redundancy, the metrics proposed by Banos and Manolopoulos (2015) violate some important requirements for an IQ measurement model:

The model needs to be nonredundant to avoid a bias for certain dimensions.

If the effects of the conceptually same characteristics are counted more than once from different dimensions, an overall quality assessment may become dominated by those characteristics, and consequently unintentionally biased.

(Stvilia, 2006, p. 43)

The repetition of certain facets is likely to make the model biased and limit its usefulness.

Other researchers have also focused on the notion of archivability and attempted to operationalize it. In their paper "The impact of JavaScript on archivability", Brunelle, Kelly,

Weigle, and Nelson (2015) defined archivability as the ease with which a website can be archived, which is similar to the concept put forward by Banos and Manolopoulos (2015). The authors held that the current, live version of a website to be the ideal version. Thus, a perfectly archived website is one that replicates the original, live version in its entirety: “The web page in its live, native environment is the best version possible, and if an archival tool replicates the live web, it has perfectly captured and archived that resource” (Brunelle, Kelly, Weigle, & Nelson, 2015, p. 9).

However, obtaining a perfect copy of the original is an onerous process, made more difficult by the widespread use of the JavaScript programming language. The use of JavaScript, in the form of small pieces of code called *scripts*, has made websites more personalized and interactive. Its use offers enhanced browsing experiences for the user, such as the ability to share a link to a web page via Twitter or to set the location of a map in Google. JavaScript code is contained inside a website’s HTML markup and executed on the user’s own computer, and so it is called a *client-side* technology. Unfortunately, this rise in feature-rich and interactive websites has also made them more difficult to archive. As the authors state, today’s archival tools, such as the Heritrix web crawler employed by the Internet Archive, are unable to fully capture and render this complexity (Brunelle, Kelly, Weigle, & Nelson, 2015, p. 2).

A website that contains JavaScript, such as Google Maps, functions differently from a traditional, HTML-only website. Typically, a web browser requests a website from a server, then proceeds to load the basic elements, such as HTML code and images. After the initial page is loaded, the JavaScript code is executed, This code will then request additional components to be loaded onto the page, such as the panning and zooming functions of an interactive map or geographic location features. Brunelle et al (2015) define these type of websites as *deferred representations* because they are not “fully realized and constructed until *after* the client’s-side representation is rendered” (p. 3). When attempting to archive such a website, a crawler will usually capture the initial components that are loaded first, but will not

capture the other components that are loaded after the JavaScript code is executed. This is because crawlers such as Heritrix cannot execute JavaScript code (Brunelle et al, 2015, p.3).

To study the impact of JavaScript on archivability, the researchers compiled two sets of archived URLs: some taken from the social media platform Twitter and others from the Internet Archive’s Archive-It service (Archive-It, 2014). The archived URLs were from the period 2005 to 2012. Brunelle et al.(2015) studied the quality of the archived URLs and their use of the JavaScript language, and presented several metrics to measure their archivability.

Each URL had a specific number of client-side components (files which execute on the end user’s computer, such as JavaScript) and server-side components (files which execute on the server). The authors called these components *parameters* and defined them in Equation 21. The complexity of a single URL was measured as the arithmetic mean of its *depth* (number of levels down from the top-level domain) and the number of client-side and server-side parameters, as shown in Equation 22. According to this formulation, a URL such as <http://edition.cnn.com/2009/SPORT/football/06/11/ronaldo.real.madrid.manchester/index.html?eref=edition> would have a depth of 6 (since the site is six levels down from the homepage) and one server-side parameter (indicated by the `?eref=edition` portion of the URL). Its complexity would then be equal to 3.5.

$$(21) \quad F = \max(|client - sideparameters|, |server - sideparamters|)$$

$$(22) \quad UC = \frac{|Depth| + F}{2}$$

$$(23) \quad CC = \sum script\ tags \in HTML$$



$$(24) \quad \begin{aligned} & \text{Javascript} - \text{loaded resources} = \text{Number of resources loaded} \\ & \quad - \text{Number of resources in HTML tags and CSS} \end{aligned}$$

Brunelle et al (2015) also presented the *content complexity* metric for a URL, which is measured as the number of `<script>` tags present inside its HTML markup. These tags indicate the presence of JavaScript code, and so the more JavaScript a website contains the more complex it is. The researchers computed these metrics for their collections and analyzed their relationship to website archivability and completeness.

Unlike Banos and Manolopoulos (2015), Brunelle et al. (2015) thought of archivability not as a discrete measurement, but as a dynamic one that changed over time. They found that over half (54.5%) of the URLs in their collection used JavaScript to load embedded resources, an increase of 14.7% between 2005 and 2012 (p. 18). Similarly, JavaScript was responsible for 52.7% of all missing embedded resources during the same time period, an increase of 32.5% Brunelle, Kelly, Weigle, and Nelson (2015, p. 19). Based on these findings, they concluded that the archivability of websites was being negatively affected by the increasing use of JavaScript, and that in the future, the completeness of archived websites would also decrease as a result.

It is worthwhile to note that the research published by Brunelle et al (2015) focuses on specific, single URLs, not on an entire website, which can consist of dozens or even thousands of URLs. However, it would be reasonable to assume that, if a single web page becomes less archivable the more JavaScript it contains, the same would apply to a complete website and even an entire web archive. The more JavaScript a website contains, the less archivable it is, and the more JavaScript a collection of websites contains, the less archivable they are as a group.

Both Brunelle et al (2015) and Banos and Manolopoulos (2015) have a similar understanding of archivability at the conceptual level, but they differ sharply in the details. For

both groups of authors, a website is archivable if it is easy to preserve. However, Banos and Manolopoulos (2015) present a list of requirements that are not present in Brunelle et al (2015), such as the website's use of metadata records, the use of proprietary (as opposed to open-source) file formats, and the presence of remote (as opposed to local) website components. In choosing to focus on the use of JavaScript, Brunelle et al (2015) seem more concerned with being able to accurately replicate the original website's functionality for the end user, while Banos and Manolopoulos (2015) have a marked focus on standards and compliance.

#### 2.5.3.3. Web Archive Quality and Similarity

In the field of Information Retrieval, documents and queries are often represented in terms of a *Vector Space Model*. This model, introduced by Salton et al. (1975) represents each document in a system as a vector, with each word of the document representing a distinct element of the vector. For example document  $d$  might be represented as  $d = (w_1, w_2, w_3, \dots, w_n)$ . Similarly, queries are also represented as vectors. When a user issues a query for the system, the elements in the query vector are compared to the elements of the document vectors and the best matches are calculated and presented as results. Often, matches are found by calculating a degree of similarity between the query vector and a document vector, with higher degrees of similarity corresponding to higher-ranking, better results.

One of these similarity measures, the cosine similarity, was used by Stvilia as a metric for measuring cohesiveness and naturalness in metadata, as described in Section 2.4.1.4. Stvilia also used Euclidan similarity as a measure of accuracy/validity in metadata. Section 2.5.3.1 discussed how AlNoamany et al. applied cosine similarity to detect off-topic web pages in a web archive.

Table 2.9

*The Results of Different VQI Comparisons and their Outcomes*

Results of comparison according to the VQI			
<b>Comparing new crawl to...</b>	<b>Green</b>	<b>Orange</b>	<b>Red</b>
Live website	The comparison is OK	The live website shows some differences to the new crawl	The live website varies significantly from the new crawl
Most recent version of the archived website stored in the library	The comparison is OK	The archived version shows some differences to the new crawl	The archived version varies significantly from the new crawl
Next step	No significant changes to the website	The crawl engineer should analyze the differences.	There has been a complete redesign of the website. Quality control will be carried out by the web archivists

since the last snapshot was made.	The elements and colors might be arranged in a different way,
No action is needed	but there might not have been a complete redesign

*Note.* Adapted from *Visual quality indicator (VQI)* (Tech. Rep.) by Swiss National Library, 2015, Bern, Switzerland.

Clearly, there is some precedent for applying measures of similarity to the subject of IQ and web archives. In 2014, the Swiss National Library (NL) began implementing a system called the Visual Quality Indicator (VQI) to help their web archivists better conduct the QA process (Swiss National Library, 2015). The VQI system is deployed when an ongoing web crawl has reached a certain size. It compares the visual appearance of an archived website in an ongoing crawl to A) the live website and B) the most recent version of the archived website stored in the library. The results of these comparisons can be assigned to three types of statuses: green, orange, or red, each of which require a different action by the web archivists. Table 2.9 illustrates the different types of results that can be obtained from the VQI and the corresponding actions that must be carried out.

The part of the VQI responsible for the actual comparison works by creating screenshots of the different websites that are being compared. Each screenshot is divided into 25 different regions and the average RGB values for each region are calculated, as shown in Figure 2.10. A measure of similarity is then calculated, which produces the distance between the RGB values of each screenshot (Swiss National Library, 2015, p. 11). The greater the distance, the greater the difference between the two images, and thus, the greater the

## VQI calculation – Calculation of region vectors

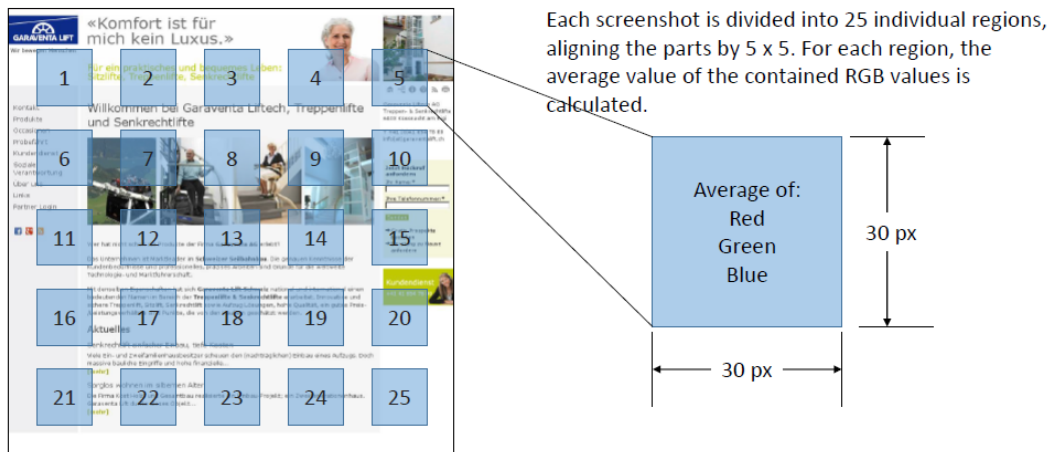


Figure 2.10. A visualization of how the Visual Quality Indicator (VQI) is calculated by system at the Swiss National Library. Adapted from *Visual quality indicator (VQI)* (Tech. Rep.) by Swiss National Library, 2015, Bern, Switzerland.

difference between the two websites.

The underlying algorithm to compare images is based on the Euclidean distance metric, which was covered earlier in Equation 1. The original code, written by Santos (2016), was adapted by the Swiss National Library to fit their web archiving needs. This simple application of the Euclidean distance metric shows it is possible to compare websites *visually*, instead of having to look at the data contained in crawl logs. This user-centered perspective would be useful in a context where a person does not have access to detailed crawl data.

As has been seen, the Euclidean and cosine similarity metrics have been already applied in some way to web archives. Another prominent measure, one that is to yet to be applied, is the *Jaccard similarity*, also known as the Tanimoto measure or min/max. Unlike cosine similarity, the Jaccard similarity was designed to measure of the overlap of two sets, as seen in Equation 25. The Jaccard similarity was originally designed for binary vectors (those vectors only containing values of 0 or 1), and that is the version given here (Jurafsky & Martin, 2008, p. 698). Ruge (1992) distinguishes between the cosine similarity and the Jaccard similarity

in terms of intention, “evaluating the relative position of two items in the semantic space [for the cosine similarity], and the overlap of property sets of the two items [for the Jaccard similarity]” (Ruge, 1992, p. 322).

$$(25) \quad k(X, Y) = \frac{\sum_{t \in X \cap Y} 1}{\sum_{t \in X \cup Y} 1} = \frac{|X \cap Y|}{|X \cup Y|}$$

Because the application of similarity measures to web archives is still in its infancy, it is yet to be determined which measure is the best one. It is possible that there is no single similarity measure that is most applicable to web archives. Instead the best measure would depend on the purpose and context in which it is applied.

## CHAPTER 3

### METHODOLOGY

This chapter introduces the main methodological approach that is used in the study — grounded theory — as well as some additional research methods.

#### 3.1. The Grounded Theory of Glaser and Strauss

In 1967, two prominent sociologists, Barney G. Glaser and Anselm L. Strauss, introduced an important new research approach: grounded theory (GT). Glaser and Strauss conceived GT as a reaction to the trends that were then prevalent in the field of sociology, notably the emphasis on verification of already-existing theories. According to Glaser and Strauss, the newest generations of sociological researchers were being trained to “master great-man theories and to test them in small ways, but hardly to question the theory as a whole in terms of its position or manner of generation.” (Glaser & Strauss, 1967/2009, “Verification and ‘Grand’ Theory”, para. 2)

The authors argued that the increased emphasis on verification had caused a dearth of theories in the field of sociology. In situations where a sociologist *had* generated a theory, she was often criticized because the theory had not yet been verified. Frustrated by these attacks, the researcher would often abandon the process of generating theory. In such situations,

[The] analyst’s confidence is destroyed because everyone involved fails to realize that accurate description and verification are not so crucial when one’s purpose is to generate theory. This is especially true because evidence and testing never destroy a theory (of any generality), they only modify it. A theory’s only replacement is a better theory. (Glaser & Strauss, 1967/2009, “Generating Theory”, para. 1)

This emphasis on verifying theory stemmed from sociologists’ desire to be seen as more “objective”, that is, aligned more closely with the pure and applied sciences. Glaser and

Strauss argued that, as a result of this approach, the field of sociology was depriving itself of new theories that would bring a fresh perspective to already-established ways of thinking.

In response to this situation, Glaser and Strauss created GT, which they defined as “the discovery of theory from data - systematically obtained and analyzed in social research” (Glaser & Strauss, 1967/2009, “The Discovery of Grounded Theory”, para. 1). For the authors, theory was not a perfected product that explains all facets of a phenomenon, but a process, an ever-developing entity (Glaser & Strauss, 1967/2009, “What Theory is Generated”, para. 2). GT is an inductive methodology. Working closely from the data, the researcher begins the work of generating a theory.

### 3.1.1. The Evolution of GT

Since the publication of *The Discovery of Grounded Theory* in 1967, GT has established itself as an important methodology for qualitative researchers; however, its original authors each took GT in different directions. In later works such as *Theoretical Sensitivity: Advances in the Methodology of Grounded Theory* (1970) and *Doing Grounded Theory: Issues and Discussions* (1998), Glaser remained consistent with his earlier elaboration of GT, while Strauss moved further towards verification (Charmaz, 2006, p. 8). In his later work, Strauss, in collaboration with Juliet Corbin developed new technical procedures for coding and data analysis, which were criticized by Glaser as an erosion of the original GT. Later, other researchers such as Kathy Charmaz and Judith Wuest crafted their own versions of GT. Glaser’s version of GT is sometimes called Classical Grounded Theory (CGT), while Strauss version is often called Straussian Grounded Theory (Evans, 2013). This work focuses on CGT as first introduced in *The Discovery of Grounded Theory*, and further clarified and expanded by Glaser.

### 3.1.2. Key Ideas, Processes, and Techniques

Glaser and Strauss established that a good GT must meet several requirements. It must:



- (1) Closely fit the substantive area in which it will be used.
- (2) Be readily understandable by laymen concerned with this area.
- (3) Be sufficiently general to be applicable to a multitude of diverse daily situations within the substantive area, not to just a specific type of situation.
- (4) Allow the user partial control over the structure and process of daily situations as they change through time.

(Glaser & Strauss, "Applying Grounded Theory", para. 1)

In other words, a good GT must closely fit the data and also be clear, readily-applicable, and flexible. This last requirement is especially important. A theory must be flexible enough that a user who applies the theory is able to adjust it and reformulate it, as she encounters new data and situations. Later, Glaser added another important aspect of a good GT: theoretical completeness. Theoretical completeness implies that the theory "explains with the fewest possible concepts, and with the greatest possible scope, as much variation as possible in the behavior and problem under study." (Glaser, 1978, p. 125). A complete grounded theory has taken the analyst as theoretically far as possible with the available data.

#### 3.1.2.1. Substantive Theory and Formal Theory

GT can be used to generate two types of theory: *substantive theory* and *formal theory*. A substantive theory is usually specific to the context in which it is created; it has less range and predictive power than formal theory, which is more general. According to Glaser and Strauss, a substantive theory is "developed for a substantive, or empirical, area of sociological inquiry", while a formal theory is "developed for a formal, or conceptual, area of sociological inquiry"(Glaser & Strauss, 1967/2009,"Substantive and Formal Theory", para. 1). As they describe it, a substantive theory hews closely to a specific context and situation, such as nurses caring for patients in a hospital ward (patient care) or teachers interacting with their students in a classroom (education). For a substantive theory to become a formal one, it must address general concepts that appear in a multitude of situations. For example, a researcher might

formulate a substantive theory of how nurses interact with their patients or how teachers interact with their students. A formal theory might generalize these situations, for example, by describing how persons of unequal status interact with each other in highly-structured environments with clear power structures.

Figure 3.1 illustrates the degree of generalizability of substantive and formal theories. The type of theory to be generated, whether substantive or formal, also impacts which groups the researcher chooses to compare in her study. When generating substantive theory, the research minimizes the differences between groups by comparing fairly similar groups. But when generating formal theory, the researcher must maximize the differences between groups by comparing more dissimilar groups (Glaser & Strauss, 1967/2009, "Which groups", para. 7-11). To continue with the previous example, a researcher can generate substantive theory by choosing to compare fairly similar groups, such as nurses in a hospital ward, nurses in a private medical practice, and nurses who work at a nursing home. A researcher might then decide to create a formal theory by comparing more dissimilar groups such as nurses and teachers.

These choices about comparison groups also affect the predictive power of the theory. Figure 3.2 illustrates this relationship. As can be seen from Figure 3.2, there is an inverse relationship between the predictive power of a theory and the degree of difference between comparison groups. As the degree of difference between groups increases, the predictive power of a theory decreases. In other words, highly predictive theories can only be formulated about small, very similar groups. As the comparison groups grow more diverse, and thus more dissimilar from each other, the theory's predictive power will suffer.

#### 3.1.2.2. Collecting and Coding the Data

GT differs from other approaches in that the phases of reviewing the literature, collecting the data, and analyzing the data take place simultaneously. It is an iterative process in which the analyst continually revises her theory as new data comes in and is interpreted.

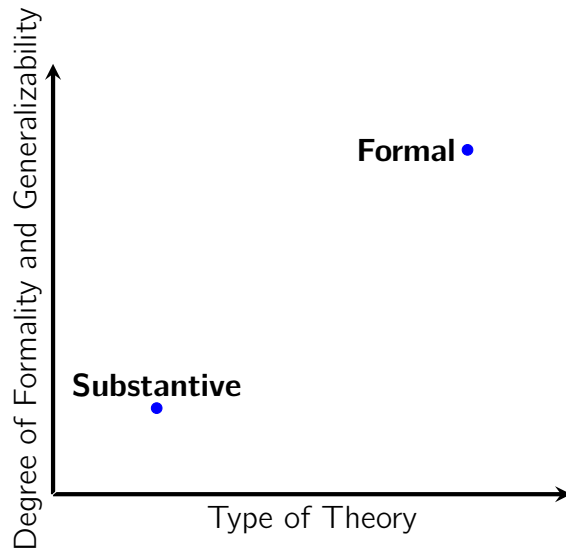


Figure 3.1. The generalizability of substantive and formal theories. Adapted from *The discovery of grounded theory: Strategies for qualitative research* by Glaser, B., & Strauss, A. (2009). [Kindle book]. Aldine Transaction. Retrieved from <http://amazon.com/o/ASIN/0202302601/> (Original work published 1967)

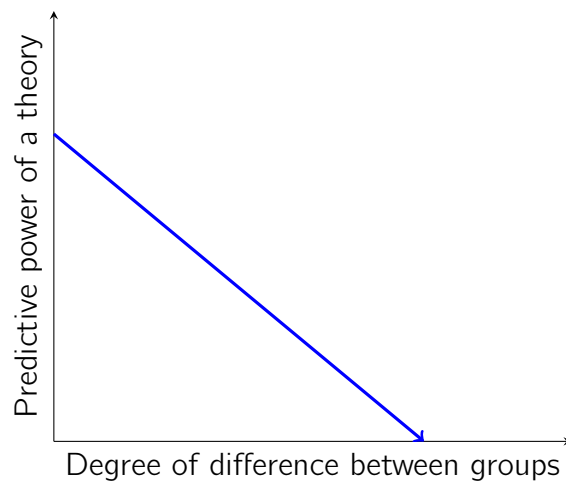


Figure 3.2. The relationship between the predictive power of a theory and the degree of difference between groups. Adapted from *The discovery of grounded theory: Strategies for qualitative research* by Glaser, B., & Strauss, A. (2009). [Kindle book]. Aldine Transaction. Retrieved from <http://amazon.com/o/ASIN/0202302601/> (Original work published 1967)

Glaser and Strauss called this process *theoretical sampling* and described it as “the process of data collection for generating theory whereby the analyst jointly collects, codes, and analyzes his data and decides what data to collect next and where to find them, in order to develop his theory as it emerges” (Glaser & Strauss, 1967/2009, “Theoretical Sampling”, para. 1). In this way, the process of generating the theory is driven by the data that is being collected, not the other way around. Additionally, by conducting the literature review during the data collection and analysis phase instead of before, the researcher avoids being unduly biased by concepts that might be best suited to a different area (Glaser & Strauss, 1967/2009, “Elements of the Theory”, para. 6).

Glaser and Strauss also broke with tradition in other ways. Although they accepted the traditional use of field notes and observations when generating theory, they also encouraged the use of alternative sources of data, such as documentary materials from the library (Glaser & Strauss, 1967, “New Sources for Qualitative Data”, para. 6). Contemporaries of Glaser and Strauss, influenced by traditional logico-deductive approaches, were careful to compare only groups that were deemed statistically comparable. Glaser and Strauss argued that the use of statistically comparable groups was fine if accurate evidence was the goal, but it hindered the generation of theory, where the “non-comparability” of groups was irrelevant (Glaser & Strauss, 1967/2009, “Which Groups?”, para. 4).

During data analysis, the researcher engages in coding, which involves “categorizing segments of data with a short name that simultaneously summarizes and accounts for each piece of data” (Charmaz, 2006, p. 43). Coding allows the researcher to discover what is happening in the data and to grapple with what it means. This process is called *open coding*

When coding the data, researchers using GT should use the *constant comparative* (or *comparative analysis*) method, which involves comparing several groups of data and “generating and plausibly suggesting (but not provisionally testing) many *categories, properties*, and hypotheses” (Glaser & Strauss, 1967/2009, “The Constant Comparative Method of Qualita-

tive Analysis”, para. 8). A category is a conceptual element of the theory, while a property, is a conceptual aspect or element of a category (Glaser & Strauss, 1967/2009, “Elements of the Theory”, para. 2). For example, in web archiving, a researcher could see characteristics such as accuracy, currency, and usefulness as being important properties of a web archive. These properties could then be classified under the category of “information quality in a web archive.” The categories with the most explanatory power are called core categories.

The process of generating theory using the constant comparison method is as follows:

- (1) Compare incidents applicable to each category. At first, the researcher begins by coding each incident of interest into as many categories as possible. While coding an incident for a category, the researcher should compare it with previous incidents in the same and different groups coded in the same category.
- (2) Integrate categories and their properties.
- (3) Delimit the theory. The researcher does this by formulating the theory with a smaller set of higher level concepts.
- (4) Write the theory.

(Glaser & Strauss, 1967/2009, “The Constant Comparative Method”, para. 1)

The output of the comparative analysis should be a theory that contains:

- (1) Conceptual categories and their conceptual properties.
- (2) Hypotheses or generalized relations among the categories and their properties

(Glaser & Strauss, 1967/2009, “Elements of the Theory”, para. 1)

Other characteristics of the GT approach include:

- The open coding process is iterative. Researchers know when to stop coding when they have reached *saturation*, that is, they can no longer extract anything new from the data (Grbich, 2012, p. 83).
- Theoretical memos: Memos are a “descriptive record of ideas, insights, hypotheses development, and testing” (Grbich, 2012, p. 87). The researcher creates these

theoretical memos after every coding session.

- Integration and model generation: This is the final step in the process and involves integrating the open coding data and the theoretical memos to create the final model (Grbich, 2012, p. 83).

For the GT to be successful, the researcher must extensively document the processes of data analysis, collection, and literature review. The coding approach and the theoretical memos are part of this documentation, which will be crucial when the researcher creates the final model.

### 3.1.3. Grounded Theory vs. Logico-Formal theory

The GT created by Glaser and Strauss is in many ways radically different from the traditional scientific approach to generating logico-formal theory. Not only is the process of generating theory very different, so are the underlying ideas behind it. Table 3.1 shows some of the differences between these two approaches.

The authors also emphasized the advantages of using GT over more traditional approaches that involved extensive verification:

Theory based on data can usually not be completely refuted by more data or replaced by another theory. Since it is too intimately linked to data, it is destined to last despite its inevitable modification and reformulation. (Glaser & Strauss, 1967/2009, "Grounded Theory", para. 4)

In traditional empirical approaches, a theory is refuted when new data is found to contradict it. Because GT is intimately linked to the data used to generate it, it can be modified and reformulated, but never outright refuted. Facts change quickly, but the general insights contributed by a grounded theory can have long-lasting impact.

Table 3.1

*Differences Between Grounded Theory and Traditional Approaches*

<b>Characteristic</b>	<b>Traditional Approach</b>	<b>Grounded Theory</b>
Literature Review	Takes place before data collection	Takes place throughout data collection and analysis
Method	Compare only "comparable" or "purified" groups	Compare any groups
Sampling	Statistical sampling	Theoretical sampling
Data	Researcher's own field notes, usually interviews and observations	Wide variety of materials, from fieldwork to library materials
Data Collection	Data is collected after theory is formulated	Data can be collected at any time
Purpose	To verify theory	To generate theory
Goal	To establish fact	To establish structural boundaries of fact
View of theory as	As a perfected product	As an ever-developing entity
...		

*Note.* Adapted from *The discovery of grounded theory: Strategies for qualitative research* by Glaser, B., & Strauss, A. (2009). [Kindle book]. Aldine Transaction. Retrieved from <http://amazon.com/o/ASIN/0202302601/> (Original work published 1967)

### 3.2. Lazarsfeld's Qualitative Mathematics

Glaser and Strauss assert that some aspects of qualitative data can be expressed mathematically, stating that "any concept can be operationalized in quantitative ways, but the sociologist should develop his concepts to facilitate this operationalization" (Glaser & Strauss, 1967/2009, "Qualitative vs. Quantitative Data", para. 11). In their chapter "Theoretical Evaluation of Quantitative Data", they discuss the process of using quantitative indexes to

express a theory derived from quantitative data. In the same chapter, the authors repeatedly reference the work of the prominent sociologist Paul L. Lazarsfeld, known as the founder of modern empirical sociology. Lazarsfeld pioneered the idea of using mathematical reasoning as an aid to theory building in the social sciences. According to Lazarsfeld, using mathematics does not lead to new findings, but it can clarify relationships: “The use of formalism in sociological data is not yet likely to lead to new findings. But it can disclose hitherto unnoticed implications or clarify the relation among propositions” (Lazarsfeld, 1959, p. 44).

One of the key concepts of Lazarsfeld’s work is the *variate* (also known as an *indicator*), which is “any classificatory or ordering device by means of which distinctions can be made among people or collectives” (Lazarsfeld, 1959, p. 46). A variate is an expression (sometimes an operationalization) of a characteristic, or trait, possessed by an individual or group. Examples of variates include the size of a city, the financial status of a company, or the IQ of an individual. Lazarsfeld’s variates are analogous to the concept of statistical *variables*.

There are two main types of variates, or indicators: expressive and predictive. Expressive indicators describe an underlying trait possessed by an individual or group, while a predictive indicator will predict the presence of that trait. Lazarsfeld observed that in research, variates often go from being predictive (more specific) to expressive (more general). (Lazarsfeld, 1959, p. 49–53). After the nature of variates is finalized, they are then combined to form an *index*. A theory, whether mathematical or not, is meant to express relationships between indices.

For example, if a researcher was exploring faculty success in a university setting she might choose variates such as whether or not a faculty member had written a dissertation, the number of scholarly articles she had published, and the number of conferences she attended. All of these variates might then be combined into a *productivity index*. Similarly, the researcher could also create an *index of honors* by combining variates such as the whether or not the



faculty member had won any awards, the number of research grants awarded to her, and whether or not she had held office in a professional society. She might then explore the relationship between faculty members' productivity and the honors they had received, which could then be expressed as a theory.

Another important contribution of Lazarsfeld's is the notion of the *interchangeability of indices*. Lazarsfeld (1959) noted, "the findings of empirical social research are to a considerable extent invariant when reasonable substitutions from one index to another are made" (p. 64). Simply put, when formulating the relationships between indices, the researcher will find that many indices are similar and lead to similar empirical results. Thus, substituting one index for another, or adding additional indices to the formula is unlikely to change the direction of the general relationship. The interchangeability of indices was Lazarsfeld's response to common tendencies in sociology, where, after a researcher would propose a set of indices, a critic would complain that the author had failed to catch the "whole meaning."

Lazarsfeld stressed that, when dealing with sociological data, the measurements and relationships expressed in a mathematical theory could never be absolute: Research questions could never be "answered unequivocally and absolutely, because [concepts such as] morale or status cannot be measured with the degree of agreement and precision with which weight or length of an object can be measured" (Lazarsfeld, 1959, p. 61). Glaser and Strauss seconded this view by stating that "for generating theory we are only looking for general relationships of direction — a positive or negative relation between concepts, and not either precise measurement of each person in the study or exact magnitudes of relationship" (1967/2009, "Concepts and Indices", para. 3). They argued that placing overt emphasis on building the perfect index might hamper the researcher's ability to create theory:

When generating theory, validation of a core index — demonstrating that the index measures the concept to a sufficient probable degree — need not be a special operation in which a theoretically relevant relation between two variables

is sacrificed from the substance of the analysis itself to prove the validity of the argument, as is typically necessary in verifications. If the index “works” —that is, if it is consistently related to a whole series of variables that, when put together, yield an integrated theory — this is validation enough of a core index. Integration of the theory is, in fact, a more trustworthy validation of an index than the standard method of merely showing that an obvious relationship exists between the index and another questionnaire item. (Glaser & Strauss, 1967/2009, “Concepts and Indices”, para. 5)

In GT, there is no need to verify that an index is a perfect measure of a concept. Mathematical approaches such as statistically significant correlations, coefficients of variation, or factor analysis need not apply. For Glaser and Strauss, if an index can lead to the creation of an integrated theory, it is sufficiently good.

According to Lazarsfeld, there are several ways of analyzing a group of variates. Lazarsfeld explored four primary ones, naming the entire process *Panel Analysis*. Table 3.2 presents the different types of analyses that comprise a panel analysis, and illustrates each one with an example using the faculty index of productivity that was mentioned earlier:

Lazarsfeld’s process of translating an original observation into an empirical index can be summarized in this way:

- (1) The researcher puts the original imagery, the intended classification, into words and communicates it by examples; She makes an effort to create definitions.
- (2) In the course of this verbalization, often called conceptual analysis, the researcher mentions several indicators, and these help to decide where a given concrete object (person or group or organization) belongs in regard to the new classificatory concept. As the discussion of the concept expands, the number of eligible indicators increases; the array of these is called the *universe of indicators*.
- (3) Usually this universe is very large, and for practical purposes, the researcher has to

Table 3.2

*Panel Analysis*

Type of Analysis	Sample Analysis using the Faculty Index of Productivity
1) The changes over time for each variate.	Does the number of published scholarly articles change over time? Does the number of conferences attended change over time?
2) Correlations between variates and their changes over time.	Does the number of published articles increase as the number of conferences attended increases?
3) Conditional relations, especially differences in (1) and (2) between subgroups that differ initially according to a specific variate, the qualifier.	Do the patterns in (1) and (2) differ for a) junior faculty, b) tenured faculty?
4) Concurrent changes of two or more variates.	Does the number of conferences attended increase <i>at the same time</i> the number of published articles increases?

select a *subset* of indicators which is then made the basis for empirical work.

(4) Finally the researcher combines the indicators into some kind of index.

(Lazarsfeld, 1959, p. 48)

In other words, the researcher begins by attempting to create a classification scheme. Then she finds all relevant indicators, selects the most important ones, and finally combines them to create an index.

### 3.3. Process

The following sections describe how I used the GT approach to generate a substantive theory of quality for web archives. It begins by describing some preliminary work I completed to prepare for the task, then describes the two primary phases of the research. The GT approach is optimal for this research problem for the following reasons:

- (1) There are no existing IQ models or theories in the area of web archiving. GT is appropriate for situations such as these where a field is relatively unexplored and there is a need for theoretical explanations and models. (Grbich, 2012, p. 79)
- (2) GT is user-centered. As its name implies, GT is heavily “grounded” in rich contextual data gathered from empirical research with actual persons.
- (3) GT is iterative. GT research involves the *constant comparison* method, which has the researcher constantly compare the emerging model/theory to the data. This allows the researcher to continually redefine a model and to become aware when no new information is emerging.(Grbich, 2012, p. 79)

#### 3.3.1. Pilot study: Developing a Preliminary Model of IQ in Web Archives

During the summer of 2014, I was an intern for the Internet Archive’s Archive-It team (AIT). Archive-It is a subscription-based web archiving service that helps organizations build and manage their own web archives. Archive-It is currently the most popular web archiving service, with over 300 clients (called “partners”) consisting of universities, state libraries and archives, museums, and national libraries in several countries (Archive-It, 2014). Before beginning the internship I met with a partner specialist and the head of the Archive-It service. They were particularly interested in two goals: gathering statistics about the tickets in the Archive-It support system, and creating a system of classification that would allow them to see the amount and type of issues they dealt with most often.

The accounts of Archive-It clients are managed by a team of partner specialists. When a client encounters a problem with the Archive-It service, she first opens a support ticket using

Zendesk, a popular customer-service platform. The ticket is received by a partner specialist, who is then responsible for addressing the issue. These initial tickets are part of the “Level 1” support. If the partner specialist determines that a problem is more serious or highly-technical in nature, the issue becomes a “Level 2” and a ticket is opened in JIRA, another issue-tracking platform. There is one support engineer who is responsible for addressing these Level 2 tickets. If he determines that the problem requires more extensive technical efforts, he will convert it to a “Level 3” ticket, which is then addressed by the software engineers at the Internet Archive.

I began my internship by examining the many tickets that had been submitted since the system began in 2011. The Zendesk system had over 600 tickets, while the JIRA system had over 400. At the time, they were not categorized. I was also given an analysis that an Archive-It intern had conducted several years prior. In it, she had created an ontology of ticket issues and labeled a large number of them using her categorization system. Despite its usefulness, this ontology was never implemented.

After examining the tickets and the prior work on them, I began to create a brand-new hierarchical classification system for Archive-It issues. After the first draft was completed, it was distributed to the Archive-It partner specialists, who provided feedback. I created a second draft, which I then tested by categorizing about 190 Zendesk tickets dating from 2013 to the present. This led me to change some categories, add new ones, and merge others to arrive at a final classification system, which is presented here. The I then made several important modifications to the ticket system in both Zendesk and JIRA, which allowed partner specialists to categorize their tickets using the ontology. I also used Zendesk Insights, an advanced reporting tool, to create automated reports that would supply statistics on the number of issues in each category over time.

It is important to note that, due to time constraints, most tickets were not classified. Zendesk does not allow users to edit tickets that have been closed, therefore restricting the

amount of tickets one could classify. However, partner specialists decided to use the system to classify future tickets. In the future, the use of this system would contribute important data that would illuminate what type of issues are most critical and need most attention from the Internet Archive.

The preliminary model of IQ for web archives that was developed based on the work summarized in this subsection is the following:

- (1) **Replay Quality** can be measured according to **Correspondence**: This is the dimension of quality that is most unique to web archives. Correspondence requires equivalence, or at least a close resemblance, between the original resource and the archived resource. In a traditional analog archive, there is a one-to-one correspondence (or at least the expectation of a one-to-one correspondence) between the original resource and the archived resource.
- (2) **Capture Quality** can be measured according to the following:
  - **Completeness**: The archived resource contains all its constituent elements.
  - **Coherence**: The archived resource is coherent if it integrates diverse elements in a logical and consistent manner.
  - **Integrity**: The data elements that constitute the captured resource are uncorrupted and error-free.

This model was a starting point for the dissertation study, the stages of which are described in the next sections.

### 3.3.2. Phase 1: Building a Substantive Theory of Quality in a Web Archive

#### 3.3.2.1. Data Gathering and Processing (Collection)

The first step in this phase was to obtain the Archive-It support tickets in order to analyze them. Since these tickets belonged to the Internet Archive, I negotiated a researcher agreement with the organization. A copy of the signed researcher agreement is provided in Appendix A. Among other conditions, the research agreement stipulates that the researcher

anonymize any personal or institutional information present in the tickets, as well as any other potentially identifying information. In order to comply with the terms of this agreement, all the information presented in this document has been anonymized: identifying elements such as personal names, names of institutions, and website addresses have been removed. For a more complete explanation of the anonymization process, please see Appendix C

The first batch of tickets was received in August 2016. This first batch was comprised of 129 AIT support tickets from the year 2012. In October 2016, a second batch of tickets was received, this one comprising 4,281 tickets from the years 2012 through 2016. It is important to note that this second batch included the original 129 tickets from 2012. They were in the form of a large file in XML format. A sample Archive-It support ticket, with the original XML tags, is included in Appendix B. This complicated XML formatting made the tickets difficult to read and analyze. In order to better analyze their content, they were put through extensive pre-processing in the form of several Python programs and Linux command-line scripts that I had written. The pre-processing steps are described below.

- (1) A Linux command-line script was used to split the large XML file into many smaller, separate XML files, each containing a single ticket.
- (2) A Linux command-line script was then used to analyze the content of the tickets and determine which year they belonged to. Tickets were then placed in their own folder by year, e.g, "tickets 2012", "tickets 2013", etc.
- (3) A Python program was used to remove the XML tags from the individual tickets.

After the tickets were cleaned, I wrote a Python program to randomly select 129 tickets each for the years 2013 through 2016. This randomization approach was taken to match the initial amount of 129 tickets for the year 2012, and also to minimize the selection bias that might have occurred if I had manually chosen which tickets to analyze. Also, the AIT platform has changed over the years: new features have been introduced while others have been dropped, the interface has been redesigned, and new, more sophisticated capture

technologies have been implemented. Choosing 129 tickets from each year ensured that the final dataset would not be biased by the strengths and weaknesses of a specific version of AIT in time. The final dataset of 645 tickets was then imported into the NVivo software package, a popular program for performing qualitative data analysis (QSR International, 2016).

### 3.3.2.2. Data Analysis

The tickets collected were Level 1 support tickets that had been submitted by AIT clients over the course of six months to a year, and included the initial question submitted by the client, the response given by the AIT partner specialist, and any subsequent communication between the two. As has been previously noted, Level 2 and Level 3 support tickets represent communication between the AIT support engineer and the team of software engineers. Because these tickets do not involve the AIT clients and are highly technical in nature, they do not contain the opinions of users and creators of web archives. Therefore, they were not considered relevant to the project and were not requested.

It is important to note that not all the AIT tickets deal with issues of quality in a web archive. Quite a few deal with collection management issues such as how to manage user accounts for a collection of web archives, storage limitations, and questions about the privacy or public access to archived content. This research focuses on tickets in which the client discusses a perceived flaw in an individual archived website or an entire web archive. From prior experiences, I had seen that these types of tickets are the most likely to deal with issues of quality. The following are some examples of AIT tickets that deal with quality issues:

- “We can’t figure out what we would need to do to capture all the images on these web pages (the vast majority of this website’s content is images).”
- “Only one page of the timeline is viewable. The live version loads earlier content as you scoll, which doesn’t happen in the crawled versions it just ends without any option to view earlier posts.”



Table 3.3

*Number of Tickets and Interactions About Information Quality Analyzed Per Year*

<b>Year</b>	<b>No. IQ tickets analyzed</b>	<b>No. interactions analyzed</b>
2012	74	478
2013	65	492
2014	67	540
2015	58	528
2016	41	506
Total	305	2544

- “I got quite a bit of info, but the stylesheets and/or layout is lacking, especially on the landing page.”
- “The site renders fine and you can hover over the progress bar for the videos and see that the frames are captured, but the video won’t play.”
- “The crawl took 12 hours and returned 103,173 documents and 3.1GB of data. This can not be correct. Crawling the whole \_\_\_\_\_ domain with my constraints yields 20,300 +- documents.”

Support tickets not pertaining to quality issues were classified as such and separated from the main data of interest. Each ticket analyzed consisted of the original ticket submitted by the client, the response sent by the AIT employee, and any subsequent interactions between them. Tickets could be quite brief, consisting of three interactions (the original client ticket, the employee’s response, and the client’s response), or they could have many interactions over time, spanning weeks or even months. Table 3.3 lists the number of tickets and interactions about IQ that were analyzed, which totaled 305 tickets and 2544 interactions.

These support tickets were analyzed using the grounded theory techniques of open coding and theoretical memos to identify the main concepts and categories present in the

data. The preliminary model created during previous work with the Internet Archive served as a guide, but most of it was modified or discarded altogether when it was found it did not fit the data. The following questions guided the analysis of each ticket:

- (1) Does this ticket deal with issues of quality in a web archive?
- (2) What is the flaw the client perceives in the archived content? How is it described?
- (3) What is the client's perception of a "good," or ideal archived website?
- (4) Is the client's idea of a good archived website different from that of the partner specialist? If so, how?
- (5) What specific language is used to describe a flawed archived website? What specific language is used to describe a good archived website?
- (6) Are any quantitative metrics used to describe any archived content, whether good or flawed?

Simultaneously, I also reviewed additional literature about GT and IQ.

According to the precepts of GT, after several rounds of coding, the researcher will reach *saturation*, a state when nothing new is being extracted from the data. Shortly after beginning to code the 2015 tickets, I reached saturation. Acting on the advice of the dissertation committee, I went back to coding, but began instead with the 2016 tickets and worked backwards to 2015. Though GT maintains there is no need to continue coding beyond saturation, this was done in an attempt to decrease bias, so the final theory would not be unduly influenced by data from a specific version of the AIT software. The additional coded data revealed nothing new, thus confirming saturation.

Appendix D contains the codebook used for coding the data in NVivo. It contains the codes(categories), their definitions, the number of tickets that contained that code, and the total number of instances of each code. The appendix contains *all* the categories that were coded for; however, not all the categories are present in the final theory. Per the guidelines of grounded theory, only the core categories (that is, the ones that explain most of the variation

in IQ) are part of the final theory.

### 3.3.3. Phase 2: Identifying How to Operationalize Dimensions of Web Archive Quality

Phase 2 of this study involved operationalizing the dimensions of quality present in the theory developed during Phase 1. That is, after having generated a multidimensional theory of IQ, the different dimensions can then be operationalized into mathematical definitions. An operational definition is the measure of a concept (Krathwohl, 2009, p. 141).

These definitions can then be used to quantitatively measure the IQ of a web archive. It is important to note that there should be no expectation that *all* dimensions of IQ can be measured quantitatively. Some IQ dimensions put forward by other researchers, such as “usefulness,” are impossible to measure because they depend entirely on the user’s opinion. Stvilia (2006, p. 42) is of the same opinion when he states that, “not all quality dimensions can be measured objectively, especially those related to the user’s immediate cognitive state or the context of use”. No attempt should be made to operationalize these.

Furthermore, IQ can be measured at several levels in a web archive: at the webpage level, the website level, and at the level of the entire web archive. The level at which IQ is measured can affect the final judgment of quality, for example, a single, specific webpage might have high IQ, but the website which contains the webpage might have an overall low IQ. Similarly, a single website might have high IQ, but the larger web archive in which it is contained might have low IQ. This dissertation focuses on IQ at the webpage level; however, an effort will be made to generalize and abstract the findings for the website and web archive levels.

For this process I used GT, but employed Lazarsfeld’s panel analysis method to explore the relationships between aspects of IQ and operationalize them. To illustrate how panel analysis might apply to the notion of quality in a web archive, one can construct a quality index for web archives composed of two variates:

- (1) Does the archived site resemble the original site in look and feel? (similarity in look

and feel)

- (2) Does the archived site function the same as the original site? (similarity in functionality)

Table 3.4 presents a sample panel analysis using a the proposed quality index. This is only for illustration purposes and will likely differ from the actual variates in the final analysis.

Table 3.4

*Sample Panel Analysis*

<b>Type of Analysis</b>	<b>Sample Analysis using the Quality Index</b>
1) The changes over time for each variate.	Over time, does the similarity in look and feel between the archived site and the original site increase, decrease, or stay the same?
2) Correlations between variates and their changes over time.	Is the similarity in look and feel between the archived site and the original site correlated to the similarity in functionality? Does the similarity in look and feel increase as the similarity in functionality increase?
3) Conditional relations, especially differences in (1) and (2) between subgroups that differ initially according to a specific variate, the qualifier.	Do the patterns in (1) and (2) differ for a) text-heavy sites that are mostly HTML and b) media-heavy sites that include audio, video, and social media content?
4) Concurrent changes of two or more variates.	Does the similarity in look and feel change <i>at the same time</i> the similarity in functionality change?

During this phase, I reviewed literature on how to use mathematical methods in the Social Sciences and more recent articles on web archiving.

#### 3.3.4. Auditing the Dissertation Work

In June 2017, I attended the Joint Conference on Digital Libraries in Toronto, Canada, and was selected as a participant for the conference's Doctoral Consortium. The Doctoral Consortium is an event where doctoral students in the early stages of their dissertation present their work and are evaluated by a panel of academics, subject matter experts, and peers. It is an opportunity for students to receive valuable advice on their dissertation.

I presented an abbreviated version of my dissertation, along with preliminary research findings. Several prominent experts in the field of web archiving were present at the Doctoral Consortium as panel members. These included Dr. Michael Nelson and Dr. Michelle Weigle from Old Dominion University, who have both carried out extensive research in the field of web archiving. The panel members gave constructive criticism of my work and advised on how to proceed with the research, specifically with Phase 2.

Throughout the entire dissertation, several of the dissertation committee members were invited to audit the process. These audit sessions included a complete review of the codes and codebook used for the data, as well as the GT memos. Once the core categories emerged, the committee members audited the core categories and their sub-categories. The research results were also shared with employees of the Internet Archive's Archive-It service. The sequence of dissertation audits is seen in Table 3.5

#### 3.3.5. Timeline of the Dissertation

Tables 3.5 and 3.6 show the complete timeline for the dissertation.

Table 3.5

*Timeline of Phase 1 of the Dissertation*

Phase 1: Building a Theory of IQ for Web Archives		
<b>Tasks</b>	<b>Subtasks</b>	<b>Time Period</b>
	Initial receipt of materials	Aug. 2016
Data collection	Second receipt of materials	Oct. 2016
and preparation	Preprocessed the tickets to classify them by year, re-move the XML formatting, and split into separate files	Oct. 2016
	Imported tickets into NVivo software	Nov. 2016
	Classified tickets according to whether or not they dealt with quality issues. Separated the relevant ones	Nov. 2016
Data analysis	Round of open coding and memoing	Dec. 2016 - Jan. 2017
	Audit of codes and codebook (Dr. Oksana Zavalina)	Dec. 2016
	Audit of initial memos (Dr. Kathryn Masten-Cain)	Jan. 2017
	Audit of codes and codebook (Dr. Shawne Miksa)	Feb. 2017
	Audit of entire dissertation and preliminary findings (JCDL Doctoral Consortium)	June 2017
	Audit of core categories and preliminary findings (Dr. Kathryn Masten-Cain, Dr. Shawne Miksa)	Aug. 2017
	Finished coding and memoing	Sept. 2017
Literature review	Literature review focusing on web archiving. Integrated results into dissertation	Feb.-March 2017
Integration and model generation	Creation of core categories and sub-categories	Oct.-Nov. 2017
	Correction of initial theory	Dec. 2017

Table 3.6

*Timeline of Phase 2 of the Dissertation*

Phase 2: Operationalizing IQ		
<b>Tasks</b>	<b>Subtasks</b>	<b>Time Period</b>
Data analysis	Audit of dissertation progress (Dr. Oksana Zavalina)	Nov. 2017
Literature review	Literature review focusing on operationalizing and mathematics	Aug. 2017-Jan. 2018
Operationalizing the dimensions of IQ	Operationalizing completeness	Sept.-Oct. 2017
	Operationalizing size relevance	Oct-Nov. 2017
	Operationalizing archivability	Dec. 2017
	Operationalizing topic relevance	Jan. 2018
	Operationalizing interactional correspondence	Feb. 2018
	Operationalizing visual correspondence	March 2018

## CHAPTER 4

### FINDINGS: A GROUNDED THEORY OF INFORMATION QUALITY FOR WEB ARCHIVES

#### 4.1. AIT Clients, their Roles, and Characteristics

The tickets analyzed in this research came from clients of the Internet Archive's Archive-It (AIT) service, therefore the information and conclusions presented in this chapter are informed by their particular perspectives and needs. AIT clients create, maintain, and use web archives on a daily basis; however, they cannot be said to be part of the general public who might engage with web archives. According to the Oxford English Dictionary, a *user* is "a person or organization who makes use of a computer or system" ("User", 2017). In contrast, *end users* are "the persons ultimately intended to use a product, as opposed to people involved in developing or marketing it" ("End user", 2013). By these definitions, AIT clients can be described as users of web archives, but are not necessarily the end users of the web archives they create. These archives are created for many different audiences and purposes: some (such as the Internet Archive's) are intended as historical records to be accessed by the public, some are created for academic researchers in specific disciplines (such as thematic or event-based web archives), while others are created as a way to preserve institutional or organizational memory and are intended for internal use.

Consequently, AIT clients have some commonalities with end users, but also differ from them in notable ways. They are a varied group of people, comprising employees of national libraries, universities, archives, government agencies, and private companies. AIT clients possess various levels of technological expertise and experience with web archives. They range from clients that have used the Archive-It service for several years and have significant experience with and knowledge of web archiving, to new clients that are just getting started. Some of the clients come from technical backgrounds, and tend to delve deep into



Table 4.1

*Differences Between AIT Clients and End Users of Web Archives*

<b>Characteristic</b>	<b>AIT Clients</b>	<b>End users</b>
Technological expertise	Various levels	Various levels
Access to information	Possess detailed information (i.e crawl logs)	Do not possess detailed information
Interest	High	Low
Institutional role	Required to create, curate, and maintain a web archive	Do not have this requirement

the technical aspects of web archiving in their tickets, while others are more interested in the collection and curation aspects of web archiving. Though end users of web archives would also possess various levels of technological expertise, they would be unlikely to know much about the process of web archiving.

Furthermore, AIT clients possess a privileged level of access to information which is hidden from the view of end users. They have access to crawl logs, reports, institutional policies, and other information detailing exactly how and why a web archive was created. If they detect a quality problem in an archived website, they can troubleshoot it and prevent it from happening again. Their roles as creators also affect their level of interest and engagement with web archiving, thus an archived website with missing images or pages might be more of a cause for concern for an AIT client than for an end user. Table 4.1 summarizes some of the differences described here.

#### 4.2. Core Categories

The grounded theory presented here consists of three core categories: correspondence, relevance, and archivability. All three of them fit Glaser's requirements for core categories, which are the following:

- (1) It must be central, that is, related to as many other categories and their properties as possible and more than other candidates for the core category...It indicates that it accounts for a large portion of the variation in a pattern of behavior. It must *reoccur* frequently in the data...If it does not reoccur a lot, it does not mean the category is uninteresting. It may be quite interesting in its own right, but it just means it is not core.
- (2) It takes more *time to saturate* the core category than other categories.
- (3) It relates meaningfully and easily with other categories.
- (4) A core category in a substantive study, has *clear and grabbing implications for formal theory*.
- (5) Has considerable *carryover*...it does not lead to dead ends in the theory nor leave the analyst high and dry, rather it gets him through the analyses of the processes he is working on, by its relevance and explanatory power.
- (6) It is *completely variable*. Its frequent relations to other categories makes it highly dependently variable in degree, dimension, and type. Conditions vary it easily.
- (7) *A core category is also a dimension of the problem*. Thus, in part it explains itself and its own variation.
- (8) The criteria for a core category is so rich, "they tend to prevent two other sources of establishing a core" which are not grounded, but without grounding could easily occur: (1) sociological interest and (2) deductive, logical elaboration. These two sources can easily lead to core categories that do not fit the data, and are not sufficiently relevant or workable.
- (9) The core category can be *any kind of theoretical code*: a process, a condition, two dimensions, a consequence and so forth.

(Glaser, 1978, p. 96)

#### 4.2.1. Veering Away from Capture and Replay Issues as Categories

In the initial model described in Section 3.3.1, quality problems in a web archive were described in terms of two categories: capture issues and replay issues. Capture issues occurred when some content was missing from the web archive, such as an entire web page, an image, a video, or a script. Replay issues occurred when all content was present in a web archive; however, an archived website still failed to work as in the original. An example would be if clicking on the “Play” button of a video did not cause the video to begin playing, or the images in a slide show did not function as in the original.

The distinction between capture and replay issues is an important and useful one for web archivists because it allows them to pinpoint the root of a problem and determines how to solve it. For example, in the context of the Archive-It service, if a partner specialist was addressing a support ticket, one of the first steps she would take would be determining if it was a capture or replay issue. For a capture issue, the partner specialist might check the crawl logs (a detailed record that includes all the files captured during a crawl) to see if a specific file was captured or not during a crawl. For a replay issue, such as with a video, the partner specialist might look to see if there was a script that was failing to execute. Capture and replay issues are each addressed differently and each require different troubleshooting strategy.

Though the distinction between capture and replay is useful for web archiving practitioners, it is unnecessary for a theorist because it obscures the negative effects that capture and replay issues have on the overall quality of an archived website. It also obscures which quality dimensions are negatively affected by the problem. For example, here are two problems that might be classed as capture issues by an AIT employee:

- (1) “This crawl captured too much unnecessary content”
- (2) “This website only captures the first page”

For creators of web archives, too much content is as much of a problem as too little or

missing content. Designating the first problem as a capture issue obscures the fact that it is really a problem of the quality dimension “relevance”. The creator of the web archive is really saying that there is too much content that she deems irrelevant and not necessary for her web archive. The second problem is on the surface a problem of completeness: obviously the 2nd, 3rd, and other deeper levels of the archived website are missing from the web archive. But this is also a problem of correspondence, since the archived website does not resemble the original because it lacks these pages. In the final theory presented here, the distinctions between capture and replay have been discarded in favor of more abstract, theoretically-appropriate quality dimensions. These are the following:

- I. Correspondence
  - A. Visual correspondence
  - B. Interactional correspondence
  - C. Completeness
- II. Relevance
  - A. Topic relevance
  - B. Size relevance
- III. Archivability

Their frequencies are shown in Table 4.2. It is important to note that the frequency numbers in the sub-categories do not add up to the totals in their main categories. This is due to the fact that many interactions were coded as belonging to more than one core category and NVivo counts them as such.

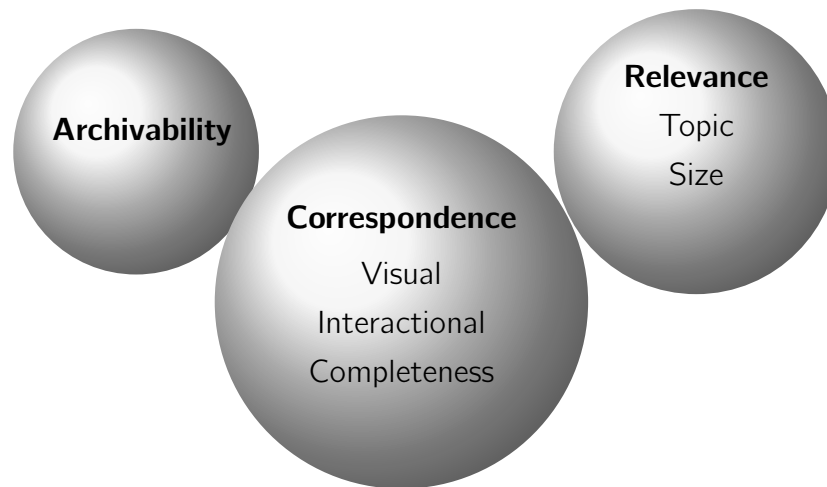
Figure 4.1 shows a visual representation of the dimensions and sub-dimensions of IQ in a web archive. The three core categories are represented as spheres whose size is proportional to their importance to the overall IQ.

Appendix D contains the entire codebook used for coding the data in Nvivo. All the data presented in this dissertation has been anonymized according to the guidelines shown in

Table 4.2

*Dimensions of Information Quality in a Web Archive and their Frequencies*

<b>Dimension</b>	<b>No. of Mentions</b>	<b>No. of Tickets</b>
<b>Correspondence</b>	852	226
Visual correspondence	160	91
Interactional correspondence	72	49
Completeness	478	157
<b>Relevance</b>	451	127
Topic relevance	93	54
Size relevance	351	107
<b>Archivability</b>	101	78



*Figure 4.1.* A visual representation of the theory of IQ for web archives

Appendix C.

4.2.2. The Dimension of Correspondence

When describing a quality problem in the tickets, AIT clients will often compare the archived website to the original website. They expect the archived website to provide the same interaction and user experience as the original. A problem occurs when a client's interaction

with the site is different from that of the original, unexpected, or deficient. AIT clients have a strong idea of what the archived website should look or behave like and are quick to report any discrepancies.

Clients express these comparisons in a number of ways. One way is by including a direct link to the original website in their tickets. This allows the partner specialist to make quick comparisons between the live site and the archived website and note the differences. Table 4.3 shows some examples of tickets where the clients made these explicit comparisons. In ticket 103, the client is reporting that she has been successful in capturing some YouTube videos; however she cannot view them using the Wayback Machine. The situation is also similar for tickets 75 and 3420. In ticket 33, the client notes a discrepancy: the archived website does not behave like the original. He points out how the archived website should look and behave (“Text next to the portraits should change as you scroll over the navigation bar”) and tells the AIT partner specialist to check the live website for the “proper” version (“how it should look”).

Many more tickets don't include the URL for the original website, but still explicitly compare it to the archived version. Some of these instances are shown in Table 4.4. In all three tickets, the clients compare the archived website to the original, live site and report on the differences between the two. For example, in ticket 1055, the client points out that the drop-down menus on the archived website do not behave as in the original and are missing content, in this case a video player and a link.

As can be seen, when describing a quality problem, clients will often unfavorably compare the archived website with the original website; however, they do not always do this explicitly. Often they simply describe an ideal state: what the website should look, behave, and what content it should include, and say there is a mismatch between this state and the actual archived website. Given the constant comparisons made between the original, live website and its archived counterpart, I determined that these are expressions of the dimension

Table 4.3

*Examples of Explicit Comparison to the Original Website Using Links*

<b>Ticket Name</b>	<b>Text of the Ticket</b>
tickets_2012/ ticket103	I have done a crawl of the following:  <a href="http://www.__.org/remembering/">http://www.__.org/remembering/</a> and the YouTube video display is problematic in Wayback on the pages. While the host report has the YouTube videos captured, they are not showing up on the web pages. See <a href="http://wayback.archive-it.org/yyhttp://www.__.org/remembering/life-work">http://wayback.archive-it.org/yyhttp://www.__.org/remembering/life-work</a> <a href="http://www.__.org/remembering/life-work">http://www.__.org/remembering/life-work</a> for how it should look.  <a href="http://wayback.archive-it.org/http://www.__.org/remembering/on-design">http://wayback.archive-it.org/http://www.__.org/remembering/on-design</a> <a href="http://www.__.org/remembering/on-design">http://www.__.org/remembering/on-design</a> for how it should look  <a href="http://wayback.archive-it.org/http://www.__.org/remembering/scotts-talks">http://wayback.archive-it.org/http://www.__.org/remembering/scotts-talks</a> and <a href="http://www.__.org/remembering/scotts-talks">http://www.__.org/remembering/scotts-talks</a> for how it should look.
tickets_2012/ ticket33	Treasures - Related items should display at the bottom of a treasure's page (see <a href="http://__.uk/roman-scrolls">http://__.uk/roman-scrolls</a> compared to <a href="http://wayback.archive-it.org/http://__.uk/roman-scrolls">http://wayback.archive-it.org/http://__.uk/roman-scrolls</a> )  Poets - Text next to the portraits should change as you scroll over the navigation bar. ( <a href="http://__.uk/">http://__.uk/</a> vs <a href="http://wayback.archive-it.org/http://poetry.__.uk/">http://wayback.archive-it.org/http://poetry.__.uk/</a> )  Byron - The theme pages should have an option for more/-less text (see <a href="http://poetry.__.uk/poems">http://poetry.__.uk/poems</a> vs <a href="http://wayback.archive-it.org/http://poetry.__.uk/poems">http://wayback.archive-it.org/http://poetry.__.uk/poems</a> )
tickets_2012/ ticket75	we are crawling the Governor's website and have captured the page where the streaming live videos appear ( <a href="http://www.governor.ne.gov/videos">http://www.governor.ne.gov/videos</a> ), but the Wayback version does not show the embedded video for that page.
tickets_2013/ ticket3420	URL <a href="http://focuspoint.dbt.ntu.edu/">http://focuspoint.dbt.ntu.edu/</a> in collection NTU Related News Publications Collection seems to have captured the videos but the webpage capture doesn't have the videos embedded.

Table 4.4

*Examples of Explicit Comparison to the Original Website Without Using Links*

<b>Ticket Name</b>	<b>Text of the Ticket</b>
tickets_2014/ ticket1055	From the site's homepage the drop down menu links are: Academics >Middle School >[missing video player / no videos link in header, in live site player plays a playlist of videos]. Academics >Upper School >[missing video player / no videos link in header, in live site player plays a playlist of videos]
tickets_2014/ ticket853	I've noticed that the quality of the video captured from [sic] Youtube in Wayback is lower than that displayed on the real YouTube site.I captured one video in my last crawl with collection MSU Social Media and the video quality is very poor ... When viewed on youtube its great.
tickets_2013/ ticket3319	The DOF crawls the site: www.obs.dof.jou regularly. However, the site owner reports discrepancies between archived & actual sites.

of quality defined here as *correspondence*. For web archives, good correspondence requires equivalence, or at least a close resemblance, between the original website and the archived website.

When assessing the quality of an archived website, AIT clients focus most on the following three flaws: mismatched appearance, mismatched behavior, and missing intellectual content. Mismatched appearance is a flaw that occurs when the archived website does not look like the original, in other words, it is a **lack of visual correspondence**. In the GT codebook, it is represented by the code "appearance of archived website", which occurs 160 times over 91 tickets. Mismatched behavior occurs when a user's interaction with the archived website is different from that of the original, unexpected, or deficient. This is termed as **lack of interactional correspondence**. It is represented by the code "user interaction is different", which occurs 72 times over 49 tickets. Missing intellectual content refers to a **lack of**



**completeness** in an archived website; the desired content is not present in the archived version. This flaw is represented by the code “completeness”, which occurs 478 times over 157 tickets.

Table 4.5 displays some examples of problems with visual correspondence. In these, AIT clients point out how the visual appearance of the archived website does not match that of the original. This is usually not stated explicitly, but the clients describe the archived website as being problematic: it lacks background images, it is “a bit off”, it “does not display properly”, or does not capture the “the look and feel” of the original. Similarly, examples of problems with interactional correspondence are shown in Table 4.6. When the clients attempt to interact with the archived website as they would with the original, they report unexpected behaviors: the text in the interactive floor plans does not display in the correct location, a page displays only very briefly and then redirects to another location, and text labels for images do not appear at all. Ticket 3428 is a special case of mismatched behavior. The client would like to recreate the search functionality available in the original. When clicking on the search box, he expects it will take him to a list of search results, which does not happen. Due to the technical constraints involved in web archiving, search functionality cannot be replicated. Table 4.7 displays examples of completeness problems, where the clients note that an archived website is missing content that assumed to be present in the original. They report missing search boxes, articles, and in some cases, even archived websites that are missing many pages.

It is important to note that these codes are not independent of each other. It is common for a low-quality archived website to have many problems, from missing pages to unexpected behaviors. Some quality problems straddle several categories. For example, ticket 260 from Table 4.5 is given as an instance of mismatched appearance, since the archived site does not include the background images as the original does. However, the same ticket can also be classified under the missing content node, since the site is missing images (intellectual

Table 4.5

*Examples of Problems with Visual Correspondence*

<b>Ticket Name</b>	<b>Text of the Ticket</b>
tickets_2013/ ticket260	On the new <a href="http://www.stateu.edu/academics">http://www.stateu.edu/academics</a> page we are not capturing the background images. I cannot figure out why since we are capturing other images from the same directory
tickets_2012/ ticket36	I also noticed that the display for your <a href="http://www.nzlibrary.edu">www.nzlibrary.edu</a> pages was a bit off.
tickets_2014/ ticket302	We're having some trouble with our Facebook site captures not displaying properly (or at all, really).
tickets_2013/ ticket3420	One thing related though, the page is not capturing its look and feel well... Any suggestions? It's missing the background and objects are not in the right locations.

content) that it should contain. In fact, many (though not all) archived websites that exhibit mismatched appearance and behaviors do so because they are missing important files that provide needed visual elements or functionality. Though the codes are separate, they are actually inextricably linked.

#### 4.2.2.1. Completeness as a Type of Correspondence

Completeness has already been described as the completeness of an archived website as it relates to the original. A perfectly complete archived website contains all of the components of the original. A completeness problem occurs when the original website's content has not been captured or is not present in the archive. Lack of completeness is caused by the absence of needed content. This section delves deeper into completeness and its causes.

Table 4.6

*Examples of Problems with Interactional Correspondence*

<b>Ticket Name</b>	<b>Text of the Ticket</b>
tickets_2012/ ticket33	the interactive floorplan isn't working as it should do - the text should appear over the map when you click on it, rather than in a list underneath.
tickets_2013/ ticket3284	When i click on it, it briefly flashes to the homepage and then it displays a URL with the nationalscience URL in it twice.
tickets_2013/ ticket3458	I would like to know if there is any way I can capture the search feature of the website, which is with the search box on the top right of the site attached. <a href="http://mishima.jp/">http://mishima.jp/</a>
tickets_2012/ ticket33	Poets - Text next to the portraits should change as you scroll over the navigation bar. ( <a href="http://__.uk/">http://__.uk/</a> vs <a href="http://wayback.archive-it.org/http://poetry.__.uk/">http://wayback.archive-it.org/http://poetry.__.uk/</a> )

## 4.2.2.2. The robots.txt file and its Role in Completeness

A *robots.txt* file is a short text file that is present in the home directory of many websites (such as [www.stateu.edu/robots.txt](http://www.stateu.edu/robots.txt)). It sets out rules, that crawlers (robots) should follow when crawling a site. Some websites utilize robots.txt files to specify that crawlers should crawl some directories, but not others, or block the crawler from crawling certain file formats, such as video or music. By default, AIT crawlers follow the rules set forward by a site's robots.txt file; however, in some cases a robots file can contain exclusions that can keep a crawler from capturing important content, as shown below. In these cases, the AIT client must enable the "ignore robots.txt" setting, which authorizes the AIT crawler to ignore the rules set out in a site's robots.txt file.

Table 4.8 presents some examples of completeness problems caused by rules in robots.txt

Table 4.7

*Examples of Problems with Completeness*

<b>Ticket Name</b>	<b>Text of the Ticket</b>
tickets_2012/ ticket33	there should be a Google search bar at the top of both websites.
tickets_2013/ ticket296	on all most every blog that we have captured from blogspot the Wayback Machine does not include the subsequent pages beyond the first.
tickets_2014/ ticket311	We're still having some trouble capturing the JavaScript menu at the top of the main page. I know that JS can be wonky, but is there anything we can do on our end to improve the chances that it will capture and display properly? Frustratingly, one of the implications of this is that the women's teams aren't being captured (or if they are, users can't navigate to them) because the only way to navigate to them in the live site (www.oursports.edu) is via the JS menu. There's a menu at the bottom that lists each sport, but the links only go to the men's teams.
tickets_2014/ ticket3117	The News pages (which are located under each individual sport) are being captured, but the actual articles that are listed and linked out are not.

files. As can be seen from the data, these types of exclusions can keep important content from being archived. Additionally, files needed in order to successfully reproduce the appearance and behavior of the original website can also be blocked, resulting in a poor-quality archived website.

In the literature about Information Quality that was reviewed in Chapter 2, completeness is often seen as a major dimension of quality. It is present in the work of Bruce and Hillman (2004), Batini and Scannapieco (2016), and Taylor (1986) (though he calls it comprehensiveness). It is therefore tempting to see completeness as its own separate dimension of IQ in web archives, different from correspondence; however this is a fallacy. An archived

Table 4.8

*Examples of Completeness Problems Caused by Robots Exclusions*

<b>Ticket Name</b>	<b>Text of the Ticket</b>
tickets_2012/ticket69	it looks as if perhaps robots.txt is responsible for blocking the capture of these javascript files needed to render the page
tickets_2013/ticket260	For <a href="http://www.stateu.edu/news/newsletter/">http://www.stateu.edu/news/newsletter/</a> it looks like the newsletter itself is on the host news.stateu.edu and that file is blocked by robots.txt
tickets_2013/ticket296	It looks like in this case, the “Older Posts” page was not captured because it was blocked by robots.txt
tickets_2013/ticket395	another seed (DRTV) blocked about 50% of the site with a robots.txt command
tickets_2015/ticket795	It is very clear from the post crawl report that there were many, many image files blocked via robots.txt from media_archive.medialab.stateu.edu

website can have a lack of correspondence with the original website yet still be perfectly complete. For example, it can have all the same components of the original, yet still look or behave differently from it. However, the reverse is not true: an archived website cannot be incomplete, yet still have 100% correspondence with the original. In logic, correspondence is known as a *necessary cause*:

If  $x$  is a necessary cause of  $y$ , then the presence of  $y$  necessarily implies the presence of  $x$  with a probability of 100%. The presence of  $x$ , however, does not imply that  $y$  will occur.

(Ohio State University, 2011, “Introduction of causal reasoning”, para. 4)

The presence of a lack of completeness ( $y$ ) always implies the presence of a lack of correspondence ( $x$ ); however, the presence of correspondence does not imply a lack of

completeness. Therefore, completeness is not a core category in the theory, but rather a sub-category.

#### 4.2.3. The Dimension of Relevance

Relevance is another dimension of quality that appears very often throughout the data (451 mentions across 127 tickets) and is also one of the most complex and difficult to describe. Much of this difficulty is due to the vague ways in which people refer to relevance. AIT clients seem to have a clear mental model of what is “relevant” or “irrelevant” content in their web archives, but they do not always articulate it explicitly. They use these internal concepts of relevance/irrelevance to delimit the boundaries of a web archive or an archived website: what is inside is (or should be) relevant, anything outside is (or should be) irrelevant.

They have a few ways of determining what is irrelevant content, and the most common types are websites or webpages:

- I. containing off-topic content (topic relevance)
- II. in quantity or volume that is unexpected or excessive (size relevance)

##### 4.2.3.1. Topic Relevance

Most AIT clients use their accounts to create topical collections, which cover a single topic or news event, such as human rights or the Arab Spring of 2010. As such, AIT clients tend to have a fairly well-defined scope for what they wish to collect. For example, one client described his scope, “The goal is to crawl only those pages and items dealing directly with Warner State (faculty, clubs, school announcements, etc)”. When a web archive or an archived website contains content about a different topic than is expected or desired, a topic relevance problem occurs. The clients implicitly assumed that a web archive will only include content that is closely related to that of the larger web archive. In reality, due to crawler settings, scoping rules, and the nature of the web, web archives often include content that is not topic-specific.

Table 4.9

*Examples of Topic Relevance Problems*

<b>Ticket Name</b>	<b>Text of the Ticket</b>
tickets_2012/ticket41	The problem is, that a lot of unrelated content is being displayed: sites we are not supposed to have in our collection, social network pages like xing and facebook, porn and dating sites, some of them even with illegal content, and so on
tickets_2012/ticket671	I noticed that we captured a message board that has a lot of unwanted garbage posted on it
ticket_2013/605	the seed <a href="http://www.oakschools.org">http://www.oakschools.org</a> has tons of garbage URLs
tickets_2012/ticket53	Is there any way to disassociate a website from our collection? For instance, in a couple of public demos we've had something outside of our collecting scope and possibly problematic appear in our collection (anti-US propaganda, pornography, etc.). I know this is the nature of web archiving, but thought I would ask in case there's a way we can go in and unhitch those specific domains despite the fact that we were the original crawlers.

This dimension and its homonymous code is mentioned 93 times throughout 54 tickets. Table 4.9 contains examples of tickets where AIT clients have detected topic relevance problems. These off-topic archived websites are described as being of little relevance and superfluous (“unwanted garbage”) and AIT clients were usually eager to remove them from their web archive.

No matter how narrow the collecting scope of a collection, determining what is relevant or not relevant is often not an easy task. In some cases, clients would flag content as irrelevant or unwanted when it was actually necessary to preserve the functionality of archived pages. A website often contains pages, or elements that are not obviously important but help “behind

Table 4.10

*Examples of Seemingly Irrelevant Content that is Actually Relevant*

<b>Ticket Name</b>	<b>Text of the Ticket</b>
tickets_2013/ ticket3350	The host yting.com serves code that effects the layout and client-side functionality of YouTube content
tickets_2012/ ticket125	It looks like there are a fair number of URLs for different sizes of the same image.
tickets_2014/ ticket147	In order to successfully archive Facebook there are a couple other hosts you'll need to ignore robots.txt for: fbcdn.net, akamaihd.net

the scenes” to make other elements or pages render correctly or function properly. This is knowledge that is known by the partner specialist, but usually unknown or invisible to the client. AIT employees often had to explain the true nature of this seemingly irrelevant content.

Table 4.10 shows examples of seemingly irrelevant content that is actually important. In the first and third examples, the AIT employee explains how pages from seemingly irrelevant host domains are actually needed to successfully archive YouTube and Facebook. In the second example, she explains how images of different sizes can each have their own URLs, which is useful when trying to archive image-heavy websites.

#### 4.2.3.2. Size Relevance

Just as a web archive can be perceived to have missing content (a lack of completeness), it can also be seen as having *too much* content. One of the unexpected findings that emerged during the coding phase was that AIT clients were worried as much about the overabundance of content in their web archives as about their completeness. They delved deep into the details of crawl statistics, logs, and reports, and readily wrote if they felt that an archived website or an entire web archive was much larger than expected. During their



examinations, they usually employed the following strategies to detect problems with size relevance.

- I. Looking at the overall size of a crawl
- II. Looking for duplicate content or at the number of times a specific site or URL was captured
- III. Comparing the size of an older web archive to that of a more recent web archive.
- IV. Comparing an older version of an archived website to a more recent version of an archived website
- V. Looking explicitly for crawler traps

Table 4.11

*Examples of General Size Relevance Problems*

<b>Ticket Name</b>	<b>Text of the Ticket</b>
tickets_2012/ticket17	The crawl took 12 hours and returned 103,173 documents and 3.1GB of data. This can not be correct. Crawling the whole law.stateu.edu domain with my constraints yields 20,300 +- docs
tickets_2012/ticket125	There are only 170 photos on this site but I ended up with 15K new URLs
tickets_2014/ticket2679	There were more than 300,000 URLs queued when my time limit ran out! Looking through the queued URLs, it looks like this site is using some jQuery tools (Colorbox, Superfish) that I'm not at all familiar with. Have you seen any sites like this before? Any suggestion for what I might be able to exclude without losing content?
tickets_2015/ticket1062	One seed, www.derap.net, brought in over 40000 URLs, all spam

When AIT clients perceived a web archive or a website to have too much content, they usually assumed the extra content was not needed and asked how to remove it or how

Table 4.12

*Examples of Size Relevance Problems Caused by Crawler Traps*

<b>Ticket Name</b>	<b>Text of the Ticket</b>
tickets_2012/ ticket212	most of the site content consists of urls with a dynamically generated 22-digit date/time stamp. This causes endless loops where the same small sets of documents are continually revisited because every visit generates a url with a time stamp several seconds later, so heritrix doesn't realize it's already seen that page.
tickets_2013/ ticket3349	regarding one particular host "www.epmonthly.com." I've checked the Queued Report and see what appears to be some kind of crawler trap.
tickets_2014/ ticket232	I added that as a host constraint and ran another test crawl. It seems to have dramatically cut down on the number of queued URLs, however there still seems to be a crawler trap of some sort, as there are still more URLs being crawled than necessary.
tickets_2015/ ticket513	the latest test crawl for the Institutional Collection shows 2 hosts with many queued urls: iym.ptsem.edu and www.facebook.com. The first appears to be a crawler trap.

to refine the scope of future crawls in order to avoid capturing it. Table 4.11 shows some examples of size relevance problems.

In the field of web archiving, a web archive with too much content often occurs as a consequence of a "crawler trap", which occurs when a crawler gets "stuck" in an endless loop, capturing the same content again and again. The calendar pages of many websites are notorious for causing crawler traps. These relationship between size relevance and crawler traps is illustrated in the tickets seen in Table 4.12.

A related phenomenon occurs when size relevance problems are caused by other types of problematic content that do not cause *not* a crawler trap. Due to the nature of web

crawlers and the websites they visit, a crawler can capture the same files several times, capture the same content from different URLs, or can even appear to have captured URLs that do not exist. All of these situations can be seen in the examples in Table 4.13. In ticket 233, the AIT client specifically refers to these non-existent URLs as invalid and says they “don’t work”.

It should be noted that size relevance is also tied to topic relevance because AIT clients judge large amounts of content as being irrelevant or unrelated to their collection goals. This might lead some to say that topic relevance and size relevance are the same thing; however, this is a mistake. AIT clients judge archived webpages to have topic relevance problems if their intellectual content falls outside of their collecting scope, whereas they judge webpages to have size relevance problems precisely *because* of their large size and heavy presence in the web archive. They do not normally look at the content of these suspect websites or pages, but judge them in a *prima facie* way.

Whereas topic relevance has been addressed in the literature by (AlNoamany et al., 2015), size relevance has not. There has been no previous work stating that an archived website might be deemed not relevant simply because of its size and not its intellectual content.

#### 4.2.4. The Dimension of Archivability

As was explained in section 2.5.3.2, the notion of archivability has already received some attention from academic researchers. It was defined by Brunelle, Kelly, Weigle, and Nelson (2015) as the ease with which a website can be archived. I redefine archivability as the intrinsic properties of a website that make it easier or more difficult to archive. Archivability is highly dependent on the technology being employed to do web archiving. As technology evolves over time, web components that were previously thought to be unarchivable might become archivable. Archivability proved to be a prominent dimension, as it appears 101 times in 78 tickets. The data showed several factors that greatly affect the archivability of a

Table 4.13

*Examples of Size Relevance Problems Caused by Problematic Content*

<b>Ticket Name</b>	<b>Text of the Ticket</b>
tickets_2013/ ticket3045	I was wondering if you know anything about websites that, when crawled, seem to produce a lot of duplicated files, with the only difference being the addition of a slash after the URL...This particular crawl seemed to produce several hundred such situations, perhaps more. I was wondering if there is an easy way to factor out the dups
tickets_2013/ ticket3128	The css and js files in the report seem to be duplicates that simply exist in different directories
tickets_2015/ ticket233	<p>First, I was surprised that there are a lot of jpg urls that don't work</p> <p><a href="http://www.inhousersearch.org/action/49/2_07_06_01_858_2465.jpg">http://www.inhousersearch.org/action/49/2_07_06_01_858_2465.jpg</a></p> <p><a href="http://www.inhousersearch.org/action/2/2_07_19_01_884_2585.jpg">http://www.inhousersearch.org/action/2/2_07_19_01_884_2585.jpg</a></p> <p>I would have thought these would be valid.</p> <p>TONs and TONs of js and css files that are all invalid, examples:</p> <ul style="list-style-type: none"> <li>• <a href="http://www.inhousersearch.org/action/js/css/jquery-ui-1.8.13.custom.css">http://www.inhousersearch.org/action/js/css/jquery-ui-1.8.13.custom.css</a></li> <li>• <a href="http://www.inhousersearch.org//action/js/lib/jquery-1.7.2.min.js">http://www.inhousersearch.org//action/js/lib/jquery-1.7.2.min.js</a></li> <li>• <a href="http://www.inhousersearch.org//css/action/standard/ie-8.css">http://www.inhousersearch.org//css/action/standard/ie-8.css</a></li> </ul>

website. Archivality problems occur when a website:

- I. has changed the way the content is delivered to the user.

Table 4.14

*Examples of Archivability Problems Caused by Websites Changing How it Delivers Content to Users*

<b>Ticket Name</b>	<b>Text of the Ticket</b>
tickets_2012/ ticket08	<p>Both Facebook and Twitter have made some changes recently to the way they set up their sites, which requires a little bit of work on our end to catch up.</p> <ol style="list-style-type: none"> <li>I. For Facebook, your site was archived, there is just an issue that is keeping the archived page from displaying normally. Our engineers are working on this and it should be fixed this week. I will let you know as soon as I have further information.</li> <li>II. For Twitter, they recently removed the “more” button from twitter feeds and instead users access older tweets by scrolling down the page. The way this feature is set up makes it difficult for our crawlers to access the older content that is not displayed automatically.</li> </ol>
tickets_2012/ ticket129	Facebook made a change to the settings for their stylesheets
tickets_2013/ ticket258	We are still generally able to capture the initial content on a Facebook timeline; however the most recent change from Facebook has made it one again difficult to capture dynamically loading content as a user scrolls down through the page

II. is media-heavy or contains much dynamic content.

III. renders content in a unique, “non-standard” way.

Table 4.14 presents examples of the first situation. Many websites routinely change the way the content is delivered to the user, thus a website can go from being easily archivable to practically unarchivable fairly quickly. As one AIT employee said: “The web, and specifically social networking sites can be a moving target.” When websites change their internal functionality, it can result in the archived website looking different from the original (tickets 08 and 129) and missing content (tickets 08 and tickets 258).

Cases where archivability was negatively impacted by the heavy presence of dynamic content are shown in Table 4.15. Generally, sites that utilize technologies such as JavaScript, Flash, and streaming audio and video are difficult to capture and render like the original. This finding is consistent with the work of Banos et al. (2013) and Brunelle, Kelly, Weigle, and Nelson (2015). A special case of this situation is seen with websites that are database and form or search-driven, such as library catalogs, web forms, or search engines. As the AIT employee explains, these are elements that depend on a myriad of complex, dynamic interactions that cannot be replicated in an archived website.

Sometimes websites will have unique or unusual ways of rendering content, which can negatively affect archivability, as seen in Table 4.16. For example, some content management systems can create endlessly repeating directory structures (such as <http://somesite.com/news>, <http://somesite.com/news/news>, and <http://somesite.com/news/news/news>). The presence of these will cause the crawler to go into infinite loops (crawler traps) in an attempt to capture all levels of the website. This can lead to poor-quality archived websites, stalled or incomplete crawls, and large amounts of unnecessary data.

Archivability is special in that it seems to be a dimension of IQ that is perceived by AIT employees (that is, web archivists), but rarely by AIT clients. In the data shown in this section, all of the people who referred to websites as being difficult to capture were AIT employees. Archivability is a dimension of IQ that is specific to web archivists: they need to know if a website is archivable *before* capturing it in order to ensure a high-quality

Table 4.15

*Examples of Archivability Problems Caused by Websites with Dynamic Content*

<b>Ticket Name</b>	<b>Text of the Ticket</b>
tickets_2012/ ticket113	flash and Javascript can be difficult to capture or display sometimes
tickets_2012/ ticket76	streaming video can be difficult to archive sometimes
tickets_2013/ ticket369	It looks like the site uses a fair bit of javascript to generate those "printer friendly" pages, but I'm not sure how feasible capture is
tickets_2014/ ticket2884	Regarding the tabs on the Press Room URL, I am not sure if we will be able to capture this content due to the dynamic way in which these links are generated
Special case: Websites that are database and form or search-driven	
tickets_2012/ ticket57	if database driven parts of sites have direct links to the content, the crawler will capture those, however the crawler can't enter search terms or interact with forms, so if that is the only way to access the database content, the crawler likely will not automatically be able to access that content
tickets_2013/ ticket3481	Because of their interactive nature, search boxes cannot operate in an archived website in the same way as they would on the live web
tickets_2014/ ticket2893	Search boxes are something that will not behave in an archived site like they do on the live web. We can archive content that would be returned by using the search function (as you noticed with the "Browse All Projects" button) however, the crawler is not able to archive the database or search engine that the live site search runs off of

archived website. Users, on the other hand, only care about archivability *after* something goes wrong. Archivability is thus a *latent* dimension, because it is hidden from most people. Only web archivists, who have a deep knowledge of and experience with the technical process of archiving websites, are able to determine a website's archivability and judge how it will affect the quality of its archived counterpart.

Furthermore, archivability is a dimension of quality that *precedes* the other dimensions of relevance or correspondence. Archivability can be measured *before* a website is captured, whereas a dimension such as correspondence can only be measured afterwards. It also singular in that one determines archivability by inspecting the original website, while the other dimensions are measured by inspecting the archived website itself. This makes it even more beholden to time than previous researchers (Brunelle, Kelly, Weigle, & Nelson, 2015) had surmised. If the original, live version of a website has disappeared, then archivability cannot be determined. Only by having access to the original website can archivability be measured and its effect on quality estimated. The uniqueness of the archivability dimension is a new finding that has not been seen before in the literature.



Table 4.16

*Examples of Archivability Problems Caused by Websites Rendering Content in Unique Ways*

<b>Ticket Name</b>	<b>Text of the Ticket</b>
tickets_2014/ ticket464	The way that this site does it's navigation is significantly more complicated than your average site due to the form based dropdowns that you notice to the right of the pagination at the top of the list. The "Sort" and "per page" options are actually forms, so instead of simply clicking on links to subsequent or previous pages (the way that most sites do pagination), the crawler would actually have to select an option from the dropdown and submit a form each time, in order to get content back. These are types of interactive behavior the crawler does not perform by default, so it will require additional development...Because this site is so uniquely complicated in the way it has implemented pagination, any work our engineers put into developing a new crawling feature to capture it would be very specific to this site and likely not transferrable to other examples
tickets_2013/ ticket3423	We do see these types of repetitive URLs from time to time, and they appear to be generated by code in certain implementations of content management systems like Drupal
tickets_2013/ ticket3001	After taking a look at the queued URLs for this host, it appears that the crawler is running into a trap that we see from time to time on some websites (including some Drupal sites) where the site generates links with repeating directories
tickets_2012/ ticket86	The issue with your <a href="http://www.pl.gov/tef/">http://www.pl.gov/tef/</a> site is one that we see from time to time, where something in the way the site is put together creates urls with repeating directories that all point back to the same page

## CHAPTER 5

### FINDINGS: OPERATIONALIZING INFORMATION QUALITY FOR WEB ARCHIVES

#### 5.1. Defining the Universe of Web Archiving

As Lazarsfeld explained, the use of mathematics as an aid in the social sciences does not lead to new findings, but it can help to clarify complex relationships. Set theory, the branch of mathematics dealing with logic, sets, and their relationships, is a useful tool that can be applied quite naturally to a web archiving context. In set theory, a *set* can be defined as a collection of definite objects (Pinter, 2014, p. 213). Similarly, a website, which is a group of elements such as HTML pages, images, scripts, and videos can also be represented as a set.

Definition 5.1.1. We define the website  $O$  as a set of elements.

- I.  $O$  is a finite set, that is, it has a finite number of elements. The size of  $O$  is a natural number,  $n$ .
- II. The elements of  $O$  can be represented as a series of components  $\{c_1, c_2, c_3, \dots, c_n\}$ . These components are elements such as HTML pages, images, and scripts that help to make the website look and act in a certain way.

The first part of this definition makes the important claim that components of a website  $O$  form a *finite*, countable set. In doing this, I am describing the website as a *closed world*; it is assumed that all the elements of a website are represented in the components,  $c_n$ , of the set  $O$ . The definition presented here adopts Batini and Scannapiecco's (2016) closed world assumption (CWA).

Just as with a website, an archived website can also be represented as a set,  $A$ , made up of components. The act of web archiving can then be represented as a function that maps the set  $O$ , the original website to the set  $A$ , the archived website. In set theory, a function such as this is written as  $f : O \rightarrow A$ . During the process of archiving a website,

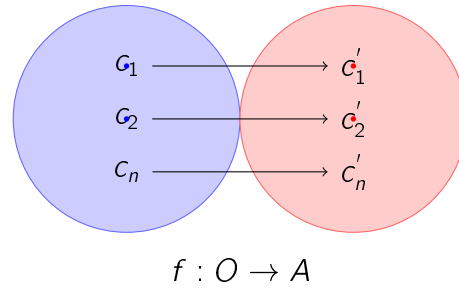


Figure 5.1. The function  $f : O \rightarrow A$  maps the original website,  $O$ , to the archived website  $A$  every component,  $c_i$  of  $O$  is mapped to a component  $c'_i$  of the set  $A$ . Figure 5.1 illustrates the process of web archiving as a mapping between these two sets.

The function  $f : O \rightarrow A$  can then be defined more formally, as is done in set theory (Pinter, 2014, p. 50):

Definition 5.1.2. The function  $f : O \rightarrow A$  has the following properties:

- I.  $\forall c \in O, \exists c' \in A : (c, c') \in f$ . This is another way of saying that every element  $c$  in  $O$  has an image  $c'$  in  $A$ .
- II. If  $(c, c'_1) \in f$  and  $(c, c'_2) \in f$ , then  $c'_1 = c'_2$ . This is another way of saying that if  $c \in O$ , the image of  $c$  is called  $c'$  and is unique.

It is important to remember that Definition 5.1.2 describes an *ideal* state of perfect quality. It makes not only the closed world assumption, but also holds that every component of  $O$  will be perfectly mirrored in  $A$ . As the different dimensions of IQ are described, I will address how other, less ideal states of quality come to be.

For every component,  $c$ , of an original, live website, there exists an identical component  $c'$  in an archived website.

$$(26) \quad \forall c \in O, \exists c' \in A : c = c'$$

According to the rules of set theory, this would give us the following result:  $A = O$ .

## 5.2. Operationalizing Correspondence

In Chapter 4, correspondence was introduced as the most prominent dimension of IQ in a web archive. This section covers how correspondence could be operationalized as a metric that can then be applied to a web archive. These metrics were arrived at by a mixture of several techniques, including panel analysis, inducting reasoning, the literature review, and my own experience working at the Internet Archive.

### 5.2.1. Operationalizing Visual Correspondence

In previous sections, visual correspondence was defined as the similarity in appearance between the original website and the archived website. Section 2.5.3.3 discussed the VQI system developed by the Swiss National Library for their web archives. This system helps Swiss web archivists decide if a website has changed and needs to be archived again. It uses the Euclidean distance to compare a screenshot of the archived website to a screenshot of the live website and determine any differences. If the distance is sufficiently high, this means the website has changed its content and web archivists should prepare to archive it again.

This same process can be applied in a different way to assess the visual correspondence of the archived website and detect any possible IQ problems. A high value of Euclidean distance would indicate greater differences between the original and archived websites, and thus a lower degree of visual correspondence. The formula for Euclidean distance, as applied to the calculation of visual correspondence is shown in Equation 27. Visual correspondence is shown as being inversely proportional to the Euclidean distance between the original and archived webpages.

$$(27) \quad VC(O, A) = \frac{1}{ed(O, A)} = \frac{1}{\sqrt{\sum_{i=1}^N (c_i - c'_i)^2}}$$

In order to calculate visual correspondence by comparing images, it is not necessary to use only the Euclidean distance. In his book *Image Registration: Principles, Tools and Meth-*

ods, Goshtasby (2012) compared the performance of 27 similarity and dissimilarity measures used for image comparison. He concluded that “an absolute conclusion cannot be reached about the superiority of one measure against another”(Goshtasby, 2012, p. 57). Although he did not declare a single measure to be the best one, he did state that the experimental results revealed that the Pearson correlation coefficient, Tanimoto measure, minimum ratio, L1 norm, L2 norm (Euclidean distance) overall performed better than other measures (Goshtasby, 2012, p. 57). Given these results, other similarity measures might also be used to measure visual correspondence.

### 5.2.2. Operationalizing Interactional Correspondence

In Section 4.2.2, interactional correspondence was introduced as a sub-category of the correspondence dimension of IQ. A problem with interactional correspondence occurs when a user’s interaction with the archived website is different from that of the original, unexpected, or deficient. For example, on the live website, a web archivist clicks on a link and is taken to the corresponding target of that link, that is, another webpage. She expects the same thing to happen on the archived version of the original page. If it does not, and she is not taken to a different webpage, the archived website lacks interactional correspondence. Problems with interactional correspondence occur when there is mismatch between a user’s expectation of website behavior and the actual behavior displayed by the archived website.

Interactional correspondence can be explained as a simple logical argument, with a premise and a conclusion. If  $p$  occurs, then  $q$  must occur, or in mathematical notation:  $p \rightarrow q$ . This can be applied to the previous example, where  $p$  = “clicking on a link” and  $q$  = “being taken to a different webpage”. An archived website displays a lack of interactional correspondence when  $p \rightarrow \neg q$ .

In the context of the Web, a single website is a complex aggregation of components such as HTML files, CSS scripts, and interactive elements that work together to give a website its look and feel. The interplay between these components is what determines a

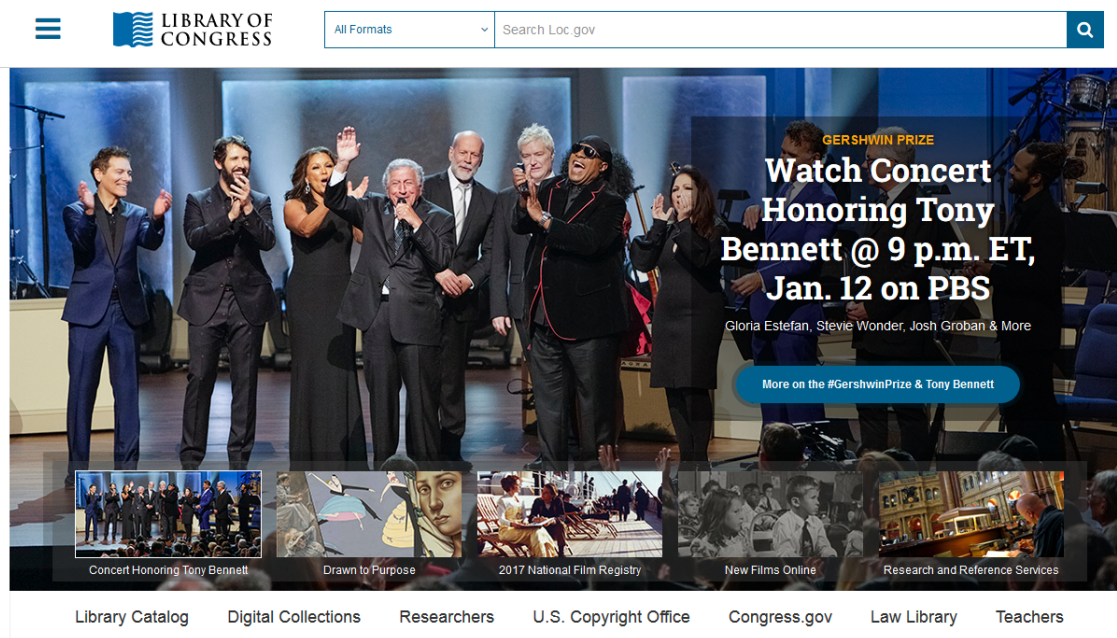


Figure 5.2. Screenshot of the Library of Congress website as seen on January 11, 2018.

website’s behavior. The simple task of loading a website on a browser is actually a process of actions that take place “behind the scenes” to produce a website. Figure 5.2 is an example, a screenshot from the Library of Congress website as seen on January 11, 2018 with the Firefox browser in Windows (“Firefox Quantum (Version 57.0.4)”, 2017).

Most users will be exposed only to this visual portion of the site. However, behind the scenes there is a wealth of activity that is occurring. The Firefox Quantum browser can display this information through its “Developer Tools” features. For example, its Network Monitor tool shows all the network requests Firefox makes as it loads the Library of Congress website. A screenshot from the Network Monitor is shown in Figure 5.3. As can be seen from the image, the Network Monitor shows that no less than 53 network requests were made over 4.78 seconds to be able to load the homepage of the Library of Congress website.

The list shown in Figure 5.3 is very extensive, and covering it in detail is beyond the scope of this dissertation; however, a few key elements can be explained and analyzed in more depth. These are shown in more detail in Table 5.1. The table shows one example of each for

Status	Method	File	Domain	Cause	Type	Transferred	Size	Time
200	GET	/	www.loc.gov	document	html	16.18 KB	82.50 KB	63 ms
304	GET	base.css?63001.62676	www.loc.gov	stylesheet	css	cached	803.66 KB	15 ms
304	GET	jquery-1.8.2.js	www.loc.gov	script	js	cached	260.63 KB	15 ms
304	GET	base.js?63001.62676	www.loc.gov	script	js	cached	915.37 KB	0 ms
304	GET	share.js?63001.62676	www.loc.gov	script	js	cached	413.47 KB	0 ms
304	GET	satelliteLib-6b47f831c184878d7338d4683ecf773a...	assets.adobedtm.com	script	js	cached	71.36 KB	0 ms
304	GET	gershwin_2018.jpg	www.loc.gov	img	jpeg	cached	299.91 KB	16 ms
304	GET	gershwin_2018_thumb.jpg	www.loc.gov	img	jpeg	cached	39.45 KB	16 ms
304	GET	drawn_to_purpose.jpg	www.loc.gov	img	jpeg	cached	288.25 KB	16 ms
304	GET	drawn_to_purpose_thumb_l.jpg	www.loc.gov	img	jpeg	cached	36.46 KB	78 ms
304	GET	titanic_R.jpg	www.loc.gov	img	jpeg	cached	183.44 KB	32 ms
304	GET	titanic_thumb_R.jpg	www.loc.gov	img	jpeg	cached	36.48 KB	16 ms
304	GET	film_duck_R.jpg	www.loc.gov	img	jpeg	cached	82.78 KB	32 ms
304	GET	film_duck_thumb_R.jpg	www.loc.gov	img	jpeg	cached	23.85 KB	32 ms
304	GET	carousel-5-research_reference.jpg	www.loc.gov	img	jpeg	cached	99.69 KB	16 ms
304	GET	carousel-5-research_reference-thumb.jpg	www.loc.gov	img	jpeg	cached	27.33 KB	16 ms
304	GET	t-dec-img3.jpg	www.loc.gov	img	jpeg	cached	18.21 KB	0 ms
304	GET	f-dec-img3.jpg	www.loc.gov	img	jpeg	cached	30.52 KB	16 ms
304	GET	2018_LCM_0102_thumbnail.jpg	www.loc.gov	img	jpeg	cached	53.00 KB	16 ms
304	GET	lc_jeff_exterior.jpg	www.loc.gov	img	jpeg	cached	55.15 KB	16 ms
304	GET	lincoln_voice.jpg	www.loc.gov	img	jpeg	cached	47.87 KB	16 ms
304	GET	JacquelineWoodson.jpg	www.loc.gov	img	jpeg	cached	87.99 KB	16 ms
304	GET	eg-read-01.jpg	www.loc.gov	img	jpeg	cached	44.48 KB	16 ms
304	GET	drawn_to_purpose_blog.jpg	www.loc.gov	img	jpeg	cached	60.86 KB	16 ms
304	GET	01-PlanYourVisit.jpg	www.loc.gov	img	jpeg	cached	22.28 KB	16 ms

53 requests | 5.60 MB / 2.46 MB transferred | Finish: 4.78 s | DOMContentLoaded: 828 ms | load: 2.24 s

Figure 5.3. The Network Monitor tool for Firefox shows the all the network requests the browser makes when it loads the Library of Congress website.

HTML, CSS, and image requests, and two examples of JavaScript requests. Each element is responsible for a specific part of the site's look and feel. Item no. 50, the `jwplayer.js` element, is actually not located on the main `www.loc.gov` site, but is actually from another URL, `cdn.loc.gov`. This illustrates how disparate components, some of them from different locations, are put together to make a website.

Every time a user interacts in any way with the site, by clicking on a link, or even hovering the mouse over an element, an item is added to the list of network requests. If the user goes to a different page, a brand-new list of network requests is generated. An interactive application on the web, such as a map, can easily generate hundreds or even thousands of network requests. A failed network request can negatively affect the user experience. In the case of archived websites, failed network requests can render the website almost unusable.

Maps are a good example of archived websites that rarely, if ever, have high levels of

Table 5.1

*Sample of Network Requests from Library of Congress Website*

#	File	Domain	Cause	Meaning
1	/	www.loc.gov	document	Initial request for browser to load homepage
2	base.css?63001.62676	www.loc.gov	stylesheet	Loads the <i>base.css</i> stylesheet, which controls the look of the page
5	share.js?63001.62676	www.loc.gov	script	Loads the <i>share.js</i> script, which handles the social media features of the page
7	gershwin_2018.jpg	www.loc.gov	img	Loads an image
50	jwplayer.js	cdn.loc.gov	xhr	Loads the <i>jwplayer.js</i> media player, which controls the slideshow at the top of the page

interactional correspondence. Figure 5.4 shows a screenshot of the interactive campus map of the University of North Texas. The map allows users to navigate through an interactive map of UNT, search for specific campus buildings, find appropriate parking spaces, and filter buildings by category (athletics, research, etc.). Because of its interactive nature, this website is almost impossible to archive correctly.

Figure 5.5 shows a screenshot of the archived version of the UNT Campus Map, as captured by the Internet Archive in 2017. As can be seen from the image, the map portion of the website is completely blank. Users can still click on the green links on the right-hand side of the page and a small location window or bubble will still show up; however, the background will always be blank. It is evident that the archived version lacks interactional correspondence, therefore its usefulness as a map is greatly diminished.

As can be expected, the list of network requests generated by the the original, live website looks very different from that of its archived counterpart. To reach the exact site as it appears in the screenshots, one must navigate to [maps.unt.edu](https://maps.unt.edu), click on the “Food”



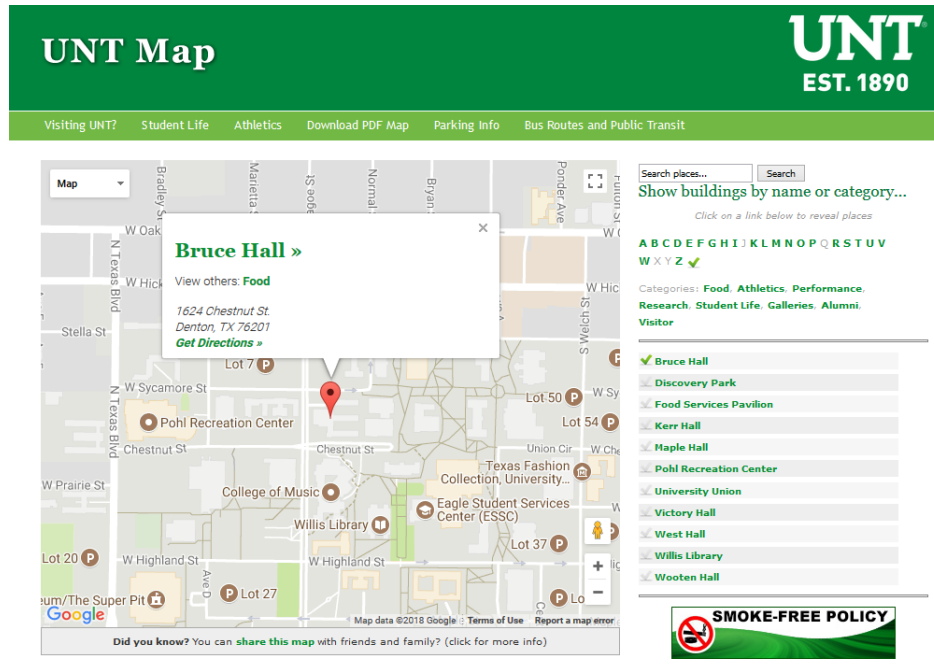


Figure 5.4. Screenshot of the campus map of the University of North Texas as seen on January 12, 2018.

category and select “Bruce Hall”. This causes the map to center itself on the location of Bruce Hall and display an informative window. These three interactions with the map, generated 108 network requests, as shown by the Network Tools windows in Firefox. In contrast, the same set of interactions generated 164 for its archived version.

Table 5.2 lists some of the most relevant network errors for the UNT Campus Map. As seen on the table, network errors typically have a status of “404” on the Network Monitor. The first error occurs because the archived website cannot load the “StaticMapService.GetMapImage” component from the Google server. This image is responsible for creating the visual portion of the map, and if it goes missing, it renders the map as a blank section, as was seen in Figure 5.5. The second error is due to the inability to execute the “ViewportInfoService.GetViewPortInfo” script on the archived website. GetViewPortInfo is a piece of JavaScript code that is responsible for centering the map on a particular geo-location, as specified by its latitude and longitude. On the original website, the script is executed when

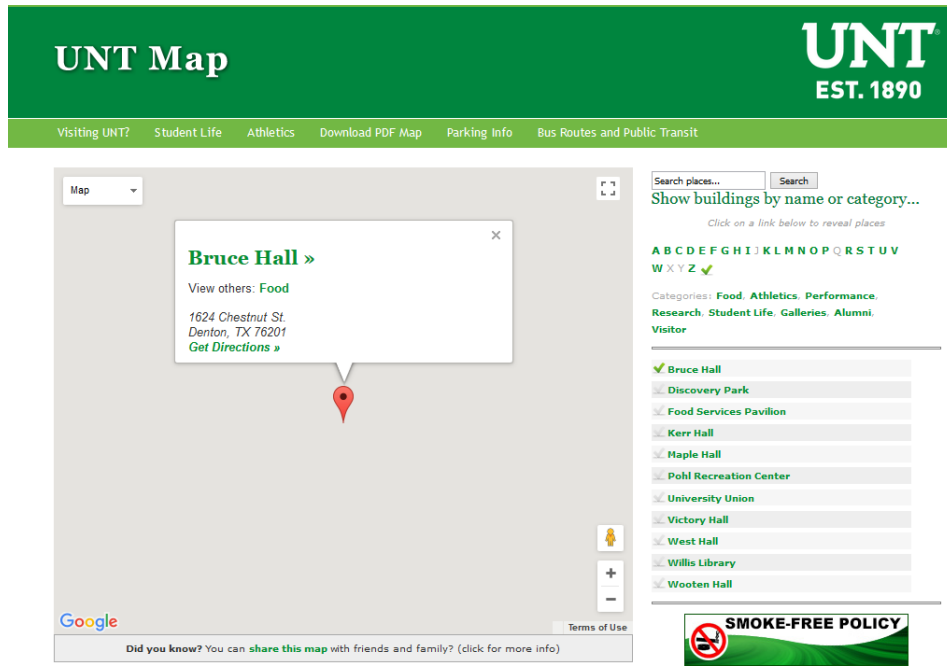


Figure 5.5. Screenshot of an archived version of the UNT Campus Map. The archived website does not display the map correctly and does not allow users to interact with it.

Retrieved from

<https://web.archive.org/web/20170910180007/http://maps.unt.edu/>

the user clicks on a location, such as “Bruce Hall”, on the map. Google would extract the coordinates of the building and center the map on its location so the user could see it. This script does not execute in the archived version.

The third and fourth network errors are also due to JavaScript code that fails to execute on the archived version of the map. The third error, “AuthenticationService.Authenticate” is a script that provides Google with an API key needed to access its map services. If the authentication process fails, as it did in the archived website, Google will not respond to any map requests issued by the browser, leading to a loss of map functionality. The fourth error occurs because “ga.js”, a Google script responsible for tracking page views and access statistics also failed to execute.

From this example, it is evident that network requests offer valuable information about

Table 5.2

*Sample of Network Requests that Generated Errors for the Archived Version of the UNT Campus Map.*

File Name	Cause	Domain		Status	
		Live Site	Archived Site	Live Site	Archived Site
StaticMapService. GetMapImage	img	maps.googleapis.com	web.archive.org	200	404
ViewportInfoService. GetViewportInfo	script	maps.googleapis.com	web.archive.org	200	404
AuthenticationService. Authenticate	script	maps.googleapis.com	web.archive.org	200	404
ga.js	script	google-analytics.com	web.archive.org	200	404

possible problems with interactional correspondence. Comparing the list of network requests of the original website to the network requests of the archived website allows a web archivist to pinpoint which components are causing errors. Therefore, network requests provide a way forward in trying to measure the interactional correspondence of an archived website. The degree of interactional correspondence, or *IC*, can be calculated by looking at the set of network requests of the archived website,  $N_A$ , as compared to the set of network requests of the original website,  $N_O$ .

Because *IC* calculates the degree of difference between the original website and its archived version, only the network requests common to both  $N_O$  and  $N_A$  are of interest. In other words, if  $x$  is a network request, then  $x : x \in (N_O \cap N_A)$ . Additionally, the network request must not have produced an error, that is, have a status of “404” on the list of network requests. If  $N_E$  is defined as the set of network requests in both the original and the archived

websites that caused errors in the archived website, then its complement  $N'_E$  can be defined as the set of network requests that *did not* cause errors in the archived website. These definitions are summarized in 5.2.1.

Definition 5.2.1. The set  $N_E$  has the following properties:

- I.  $N_E = \{x : x \in (N_O \cap N_A) \text{ and } x \text{ has caused an error in } N_A\}$
- II.  $N'_E = \{x : x \in (N_O \cap N_A) \text{ and } x \text{ has not caused an error in } N_A\}$
- III.  $N_E \cup N'_E = N_O \cap N_A$

Equation 28 presents the formula for measuring interactional correspondence. IC is thus defined as the cardinality of the set of successful (non-404) network requests in the archived website that were also in the original website, divided by the cardinality of the set of network requests of the original website. It calculates the portion of an archived website that offers the same user interaction as the original.

$$(28) \quad IC = \frac{|N'_E|}{|N_O \cap N_A|}$$

Equation 28 can be easily applied to measure interactional correspondence. Suppose there exists a live website where  $|N_O| = 100$ ,  $|N_A| = 120$ , and  $|N_O \cap N_A| = 90$ . Of the 90 network requests that are present both in the original and the archived site, there were 20 errors. Therefore  $|N_E| = 20$  and  $|N'_E| = 70$ . Using the formula the degree of IC is equal to  $70/90$  or  $0.\bar{7}$ . Therefore the archived website has 77.78 % of the interactional correspondence of the original. This formula can be applied to measure the interactional correspondence of a single webpage, or an entire website.

Furthermore, a more nuanced approach can also be employed that takes into account the importance of each element and its effect on the quality of the archived website. Recall that in Table 5.2, not all network requests resulted in the loss of map functionality

for the archived website. The errors “StaticMapService.GetMapImage”, “ViewportInfoService.GetViewPortInfo”, and “AuthenticationService.Authenticate” did have a deleterious effect on the final rendering of the archived map, however, the “ga.js” did not. This particular script is used by Google to track website traffic and compile access statistics, and has no effect on the user’s interaction with a site. This situation of unequal effects can be operationalized by giving each component of the website a different weight that corresponds to its effect on the interactional correspondence of the archived website.

This approach is informed by the work of Banos and Manolopoulos (2015), which was covered in Section 2.5.3.2. Banos and Manolopoulos (2015) defined archivability as a series of facets  $F$ , each having a weight  $w$ , as presented in Equation 29.

$$(29) \quad WA = \sum_{\lambda \in \{A,S,C,M\}} w_{\lambda} F_{\lambda}$$

In their work, the authors gave each facet a different weight. For facets of high importance,  $w_{\lambda}$  was set to 4,  $w_{\lambda} = 2$  for facets of medium importance, and  $w_{\lambda} = 1$  for facets with low importance.

Taking this approach as the starting point, then each network request  $x_i$  of an archived website can be assigned a weight  $w_i$ . Errors that contribute significantly to the interactional correspondence of the site, such as the “StaticMapService.GetMapImage” JavaScript error, are given  $w_i = 4$ . Errors with medium significance can have  $w_i = 2$ , and errors of low significance, such as the “ga.js” error can have  $w_i = 1$ . As in the original formulation,  $N_E$  is still defined as the set of network requests in both the original and the archived websites that caused errors in the archived website, and its complement  $N'_E$  is still the set of network requests that *did not* cause errors in the archived website. However, this time there are a few additional details about  $N_E$ , shown in Definition 5.2.2.

Definition 5.2.2. The set  $N_E$  can be defined as  $N_E = \{x : x \in (N_{E_H} \cup N_{E_M} \cup N_{E_L})\}$ ,

where

- I.  $N_{E_H}$  is the set of network requests in both the original and the archived websites that caused errors of *high importance* in the archived website
- II.  $N_{E_M}$  is the set of network requests in both the original and the archived websites that caused errors of *medium importance* in the archived website
- III.  $N_{E_L}$  is the set of network requests in both the original and the archived websites that caused errors of *low importance* in the archived website
- IV.  $N_{E_H} \cap N_{E_M} \cap N_{E_L} \equiv 0$ , that is, the three sets are disjoint because a network request can only have only one type of importance: medium, high, or low importance

$N_E$  can be decomposed into a vector of network requests, each with its own weight according to its importance. Then the formula for  $N_E$  can be rewritten as:

$$N_E = w_i * |N_{E_H}| + w_i * |N_{E_M}| + w_i * |N_{E_L}|$$

where  $w_i = 4$  for network requests of high importance,  $w_i = 2$  for network request of medium importance and  $w_i = 1$  for network requests of low importance

Similarly,  $N'_E$ ,  $N_O$  and  $N_A$  can also be rewritten in a similar manner:

$$N'_E = w_i * |N'_{E_H}| + w_i * |N'_{E_M}| + w_i * |N'_{E_L}|$$

$$N_O = w_i * |N_{O_H}| + w_i * |N_{O_M}| + w_i * |N_{O_L}|$$

$$N_A = w_i * |N_{A_H}| + w_i * |N_{A_M}| + w_i * |N_{A_L}|$$

$$N_O \cap N_A = w_i * |N_{O_{A_H}}| + w_i * |N_{O_{A_M}}| + w_i * |N_{O_{A_L}}|$$

For  $N_O \cap N_A$ ,  $N_{OA}$  refers to the set of components that are present in both  $N_O$  and  $N_A$ . As with  $N_E$ , the weighted components of  $N'_E$ ,  $N_O$ , and  $N_A$  are disjoint. The original formulation for IC then becomes the weighted version shown in Equation 30:

$$(30) \quad IC = \frac{N'_E}{N_O \cap N_A} = \frac{w_i * |N'_{E_H}| + w_i * |N'_{E_M}| + w_i * |N'_{E_L}|}{w_i * |N_{OA_H}| + w_i * |N_{OA_M}| + w_i * |N_{OA_L}|}$$

Equation 30 can be applied to measure the weighted interactional correspondence. Suppose there exists a live website where  $|N_{OA}| = 150$ ,  $|N_E| = 50$  and  $|N'_E| = 100$ . Of the 150 network request present in both the original and archived website, there are 55 high-imporantance ones, 55 of medium importance, and 40 low-importance requests, therefore  $|N_{OA_H}| = 55$ ,  $|N_{OA_M}| = 55$ , and  $|N_{OA_L}| = 40$ . Of the 50 network requests that returned errors, the errors according to their importance were:  $|N_{E_H}| = 35$ ,  $|N_{E_M}| = 12$ , and  $|N_{E_L}| = 3$ . Therefore  $|N'_{E_H}| = 20$ ,  $|N'_{E_M}| = 43$ , and  $|N'_{E_L}| = 37$ . The calculation for the IC, according to the Equation in 30 would then be:

$$\frac{35(4) + 12(2) + 3(1)}{55(4) + 55(2) + 40(1)} = \frac{167}{370} = 0.45$$

Therefore the archived website has 45 % of the interactional correspondence of the original. This formula can be applied to measure the interactional correspondence of a single webpage, or an entire website.

### 5.2.3. Operationalizing Completeness

As was explained in the Section 5.1, web archiving can be characterized as a function  $f : O \rightarrow A$ . In an ideal state of quality, every component of  $O$  would be perfectly mirrored in  $A$ . In other words, when the web archiving process has resulted in an archived website of perfect quality,  $f : O \rightarrow A$  can be seen as a *bijective* function. In a bijective function, every

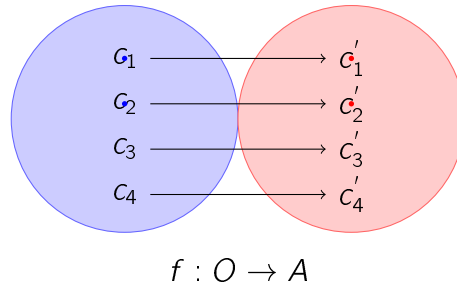


Figure 5.6. The function  $f : O \rightarrow A$  is bijective, so every element in  $O$  has *exactly one* corresponding element in  $A$ .

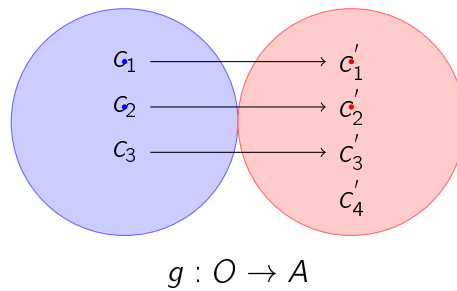


Figure 5.7. The function  $f : O \rightarrow A$  is injective, so there are elements in  $A$  that are not present in  $O$ .

element of  $O$  has *exactly one* image in  $A$ , and every element of  $A$  has a corresponding image (called the pre-image) in  $O$  (Pinter, 2014, p. 54).

This characterization is illustrated in Figure 5.6.

Set theory can also provide ways of describing more common, less-than-ideal states of quality in an archived website. In Section 4.2.3.2, I introduced the concept of size relevance. Problems with size relevance occurred when web archivists perceived a web archived to have too much content, and thus deemed that content irrelevant and fit to be removed. This situation could be described in terms of an *injective* function  $g : O \rightarrow A$ . In an injective function, every element of  $A$  has *no more than one* pre-image in  $O$ . Therefore, if  $(c_i, c'_i) \in g$ , and  $(c_j, c'_j) \in g$ , then  $c_i = c_j$  (Pinter, 2014, p. 52).

Figure 5.7 illustrates what an injective function looks like in the context of web archiving. Every element in  $O$  has a corresponding element in  $A$ ; however, there is a single element



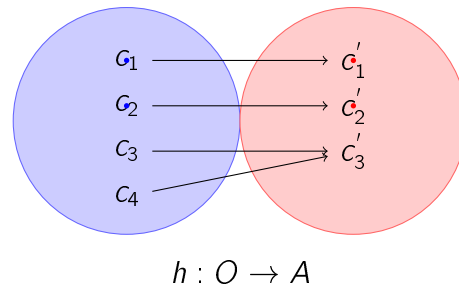


Figure 5.8. The function  $h : O \rightarrow A$  is surjective, so there are elements in  $O$  that are not present in  $A$ .

in  $A$ ,  $c'_4$ , which is not an image of any element in  $O$ . Therefore  $A$  could be perceived as having too much content because it has more elements than  $O$ . In a web archiving context, this problem is multiplied, since a web archive can contain thousands or even millions of objects that are not in the original.

Just as the problem of too much content can be explained in terms of set theory, so can the problem of too little content, explained in Section 4.2.2.1 as a lack of completeness. As was explained, a completeness problem occurs when the original website's content has not been captured or is not present in the archive. Lack of completeness is caused by the absence of needed content.

This case can be described in terms of a *surjective* function in set theory. In a surjective function  $h : O \rightarrow A$ , every element of  $A$  must have *at least one* pre-image in  $O$  (Pinter, 2014, p. 52). A surjective function may map more than one element of  $O$  to the same element in  $O$ . Figure 5.8 illustrates what a surjective function looks like in the context of web archiving. As can be seen in the image, elements  $c_1$ ,  $c_2$ , and  $c_3$  from set  $O$  are mirrored in set  $A$ . However, element  $c_4$  does not have a corresponding mirror image,  $c'_4$ , instead it points to  $c'_3$ ,  $c_3$ 's image. From a web archiving perspective,  $A$  could be seen as being incomplete because it does not adequately mirror  $O$  since it has too little content.

As has been seen, the type of function responsible for creating the web archive can have an impact on its final quality, especially in terms of size. Injective functions lead to too

much content, while surjective functions lead to too little content, and thus a completeness problem. In terms of set theory,  $|O|$  can be defined as the cardinality (size) of  $O$ , or the number of its components, and  $|A|$  can be similarly defined as the cardinality (size) of  $A$ . Table 5.3 summarizes the different types of web archiving functions discussed in this section, their dimensions, and their impact on the quality of the final archive.

Table 5.3

*Summary of the Types of Web Archiving Functions and their Quality*

Function Name	Type	Dimensions	Description
$f : O \rightarrow A$	bijjective	$ A  =  O $	Perfect quality. Ideal state.
$g : O \rightarrow A$	injective	$ A  >  O $	Lower quality. Too much content.
$h : O \rightarrow A$	surjective	$ A  <  O $	Worst quality. Completeness problem due to missing content.

quality

Now that injective mappings between the original website and its archived version have been identified as the cause of completeness problems, completeness can be fully operationalized. In web archiving, it is not enough to know if an archived website contains more or less content than the original, *how much more or how much less content* is just as important. As was discussed in Section 2.5.3.3, the traditional Information Retrieval notions of similarity can be useful when calculating the difference between a website and its archived version.

For example, completeness can be operationalized as the cosine similarity between  $O$ , the original website and  $A$ , the archived website, as seen in Equation 31.

$$(31) \quad cosine(O, A) = \frac{O \cdot A}{|O||A|}$$

Cosine similarity was chosen as a measure of completeness because of its prior use in (AlNoamany et al., 2015) for detecting off-topic webpages. However, other similarity

measures such as Jaccard similarity and Euclidean distance might also be used. This flexible approach to measurement is consistent with Lazarsfeld's notion of the *interchangeability of indices*. As Lazarsfeld noted, "the findings of empirical social research are to a considerable extent invariant when reasonable substitutions from one index to another are made" (Lazarsfeld, 1959, p. 64). Simply put, when formulating the relationships between variables, the researcher will find that many measures are similar and lead to similar empirical results. Thus, substituting one measure for another, or adding additional measures to the formula is unlikely to change the direction of the general relationship.

If the original website and its archived version are expressed as bit vectors  $O$  and  $A$  that contain all the components,  $c$ , of a website, such as text, images, video, etc, then the cosine similarity becomes:

$$(32) \quad \text{cosine}(O, A) = \frac{O \cdot A}{\|O\| \|A\|} = \frac{\sum_{i=1}^n c_i * c'_i}{\sqrt{\sum_{i=1}^n c_i^2} * \sqrt{\sum_{i=1}^n c_i'^2}}$$

$$O = \langle c_1, c_2, c_3, c_4, c_5, \dots, c_n \rangle$$

$$A = \langle c'_1, c'_2, c'_3, c'_4, c'_5, \dots, c'_n \rangle$$

- In cosine similarity, the values calculated range between 0, for vectors that do not share any components, to 1, for vectors that are identical, to -1, for vectors that point in opposite directions. The values of a vector can be binary, that is, 0 or 1. Let us assume that the value of each component,  $c$ , is also binary. So  $c_n = 0$  if the component is absent, and  $c_n = 1$  if the component is present.
- Let us assume that the original website,  $O$ , always has all of its components, so  $O = \langle 1, 1, 1, 1, 1, \dots, 1 \rangle$

- The archived website,  $A = \langle c'_1, c'_2, c'_3, c'_4, c'_5, \dots, c'_n \rangle$ , since we do not yet know the values of  $A$ .

Then the magnitudes of the original site and the archived site can be calculated:

$$|O| = \sqrt{\sum_{i=1}^n c_i^2} = \sqrt{\sum_{i=1}^n 1^2} = \sqrt{n}$$

$$|A| = \sqrt{\sum_{i=1}^n c_i'^2} = \sqrt{c_1'^2 + c_2'^2 \dots + c_n'^2}$$

As well as their dot product:

$$O \cdot A = \langle 1, 1, \dots, 1 \rangle \cdot \langle c'_1, c'_2, \dots, c'_n \rangle = \sum_{i=1}^n (1)c'_i = \sum_{i=1}^n c'_i$$

Substituting these values into the equation, we get the following, generalized version of completeness:

$$(33) \quad \text{cosine}(O, A) = \frac{\sum_{i=1}^n c'_i}{\sqrt{n * \sum_{i=1}^n c_i'^2}}$$

To be even more consistent with Lazarsfeld's notion of interchangeability of indices, Appendix E presents an exploration of completeness that is similar to the one presented in this section. The only difference is that, instead of using the cosine similarity coefficient, I instead use the Jaccard similarity.

### 5.3. Operationalizing Relevance

In Section 4.2.3, relevance was explained as a dimension of quality that was notably difficult to describe. Much of this was due to the vague, imprecise ways, in which AIT clients expressed whether an archived webpage was relevant or irrelevant. However, relevance could be defined as being of two types:

- I. Topic relevance: how close the topic of an archived webpage or website is to the topic that is expected or desired by the creator.

- II. Size relevance: how close the the quantity or volume of the archived website or the entire archive is to that which is expected or desired by the creator.

Clearly, relevance is a dimension of quality which depends entirely on the users' perception, which makes it difficult, though not impossible, to operationalize. For example, in traditional Information Retrieval, the relevance of a document to a specific query is determined through the following process:

- I. Research subjects are presented with a query. This is often a question or topic someone would like to know more about, such as *What were the causes of World War II?* or *I'd like to know more about the health benefits of a vegetarian diet.*
- II. Subjects are presented with a number of documents that might be related to the query. They judge each document as being "relevant" or "not relevant" to the topic. In Information Retrieval, this human judgment is regarded as the "ground truth".
- III. The researchers then design an IR system that best approximates human judgments of relevance.

This process is hardly applicable to the field of web archiving for a multitude of reasons. First, the AIT interface, which clients use to create and manage their own web archives, does not have a query interface that can be used to execute topic-based queries. AIT clients do not query their own web archives in the traditional sense; rather, they judge whether an archived webpage or website is relevant or not by looking at crawl reports generated by the AIT system. Crawl reports contain detailed information about the crawl that was run in order to capture the desired websites, information such as the size of the crawl, the number of files captured, and their MIME type (HTML, image, video, etc.).

Second and most important, as was seen in Section 4.10, AIT clients often made mistakes when judging whether or not archived content was relevant, regularly flagging web content as irrelevant, when it was actually necessary to properly display an archived website. AIT employees, who had a deep knowledge of the web archiving process, were usually the

ones who were able to correctly distinguish between relevant and non-relevant elements in a web archive. In the context of web archiving, human judgments of relevance, traditionally taken to be the ground truth in Information Retrieval, are unreliable.

This section is an attempt to address these problems by proposing relevance measures that would not necessarily need ground truth judgments from human subjects. Not all the measures proposed in this section need be applied. However, when they are taken together, they form a good approximation of how humans perceive IQ in a web archive.

### 5.3.1. Topic Relevance

As AlNoamany et al. (2015) noted, a method can be devised to detect off-topic web pages in a web crawl. According to their experiments, cosine similarity was successful at detecting which pages had moved away from the original scope of a web archive. Their research was covered in Section 2.5.3.1. Their approach, which required comparing the text of web pages, is noteworthy; however it was some considerable weaknesses. Many web pages today are media-heavy and dynamic, containing more videos, JavaScript, and interactive elements than they do text. The text included in these dynamic elements is often in the form of code that controls how the user will interact with page, but which is often not relevant to the topic of the page itself. In other words, the media is itself the relevant content, not the actual HTML text on the page. A text-comparison approach might not be enough to correctly ascertain whether a web page is off-topic in a web archive.

In this section, I propose a different approach to measuring topic relevance, one which is based, not on comparing the text of web pages, but instead on certain patterns which are seen again and again in web archives. Two operationalizations of topic relevance are proposed: one that takes into account the distances between vertices in a web graph and a second one that takes into account the differences between domain names.

In visual depictions of the Web, it is typically represented as a *graph*. A graph is a structure that “connects points called *vertices* using lines called *edges*” (Erciyes, 2014, p. 11).

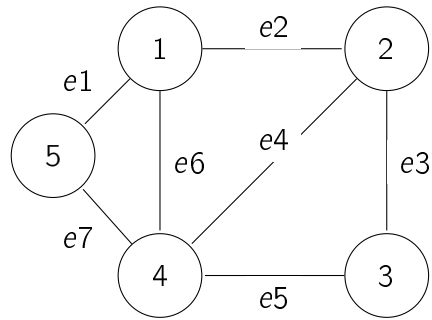


Figure 5.9. A sample graph with five vertices and seven edges

Figure 5.9 shows a sample graph.

In a graph  $G$ , the distance between two nodes  $u$  to  $v$  is written as  $d(u, v)$  and is defined as the number of edges in the shortest path from  $u$  to  $v$  (Bonato, 2005, p. 18). Equation 34 shows the formula for average distance of  $G$ , where  $S$  is the set of pairs of distinct nodes  $u, v$  of  $G$  with the property that  $d(u, v)$  is finite.

$$(34) \quad L(G) = \sum_{\{u,v\} \in S} \frac{d(u, v)}{|S|}$$

According to this definition, the distance between nodes 3 and 5 in Figure 5.9 can be defined as  $d(3, 5) = 2$ . This is because the shortest path between 3 and 5 has two edges:  $e5$  and  $e7$ .

A website can be represented as one such graph, with the pages being vertices, and links being edges. Figure 5.10 shows part of the link structure of the Library of Congress website as it was in March 2018. This graph was produced with data from the Screaming Frog SEO Spider software (“Screaming Frog SEO Spider (Version 9.2)”, 2018). It shows the homepage [www.loc.gov](http://www.loc.gov) and the web pages it links to. In this graph, the distance between *teachers* and *lessons* is  $d(\textit{teachers}, \textit{lessons}) = 2$  and  $d(\textit{share.js}, \textit{classroommaterials}) = 3$ .

I propose a way to measure topic relevance by using the distance between two components in the web graph. When crawling a website, a web crawler typically begins from a *seed*,

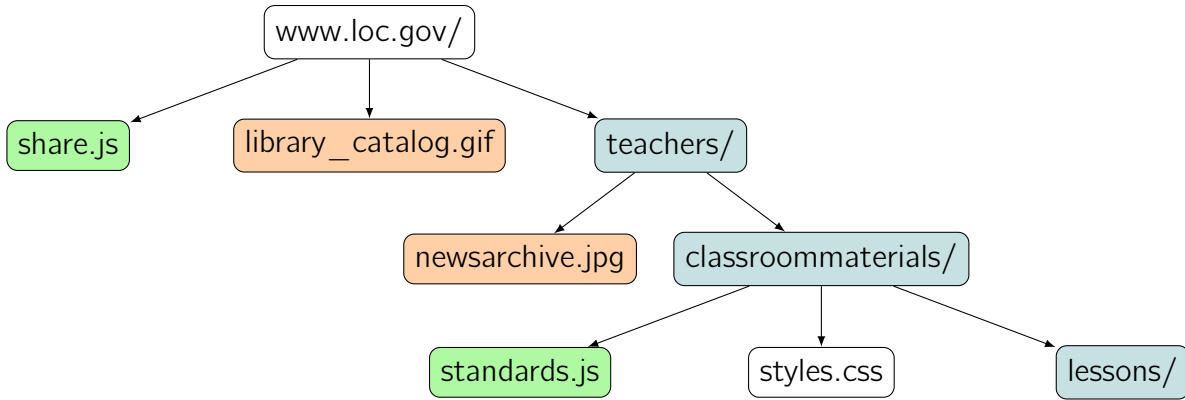


Figure 5.10. A partial view of the hierarchy for the Library of Congress website, as of March 30, 2018. HTML pages are shown in blue, images are shown in orange, and JavaScript items are shown in green.

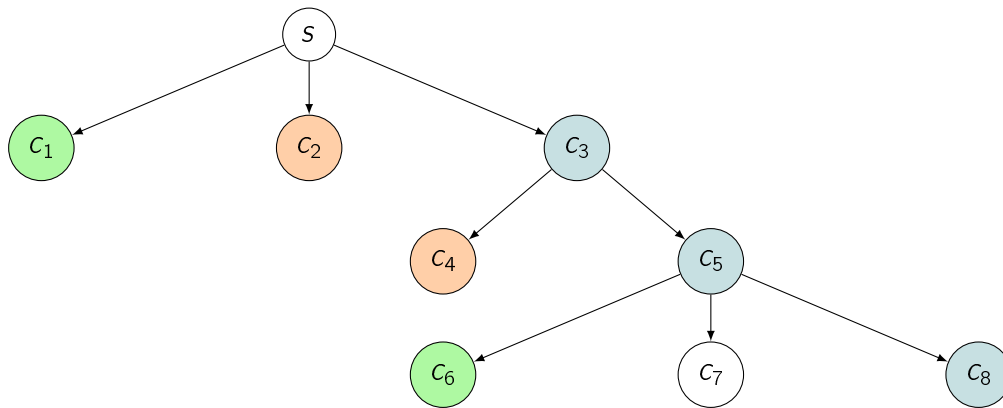


Figure 5.11. Link graph of a website to be archived showing the seed  $s$  and the pages  $c$  it links to.

which is a URL that acts as the starting point for the web archive. The web crawler will then proceed to follow every link on the seed and archive those subsequent web pages. This seed can be represented as  $s$ , while the rest of the web pages on the graph can be represented as  $c_i$ , where  $i$  is a number between 1 and  $N$ , which is the size of the graph  $G$ . The web graph presented in Figure 5.10 can then be re-drawn as Figure 5.11

$$(35) \quad TR(c_i) = \frac{1}{d(s, c_i)}$$



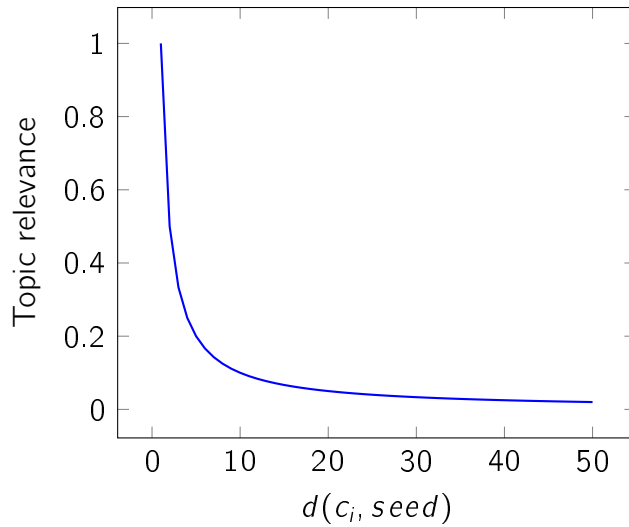


Figure 5.12. The graph of topic relevance as a function of distance from the seed URL.

The topic relevance of a single node,  $c_i$  can then be calculated as the inverse of the distance between  $c_i$  and  $s$ , as seen in Equation 35. According to this formulation, the greater the distance of a component from the seed URL, the lower its relevance to the rest of the web archive. If this formula is applied to the link graph of the Library of Congress website seen in Figure 5.10, then  $TR(newsarchive.jpg) = \frac{1}{2} = 0.50$   $TR(standards.js) = \frac{1}{3} = 0.33$ ,  $TR(library\_catalog.gif) = 1$  and so on. A component which is linked to directly from the seed page will be deemed to be more relevant than one which is farther down in the web graph. This relationship is depicted graphically in Figure 5.12. From this formula, it can be discerned that the domain of  $TR(c_i)$  is equal to  $(1, N)$ , where  $N$  is the size of the graph. Its range is equal to  $(\infty, 0)$  and thus  $\lim_{i \rightarrow \infty} TR(c_i) = 0$ .

One condition of this operationalized  $TR$  is determining the limits of the index variable  $i$ , that is, the size  $N$  of the web graph  $G$ . Since the Web is such a large and interconnected structure, it is difficult to pinpoint the exact size of a website. In this case, web archivists can choose a limit  $m$  such that  $1 \leq m \leq N$ . The final value of  $m$  will determine the amount of web content web archivists will choose to collect in a web archive. As was seen in Chapter 4.2.3.1, most archivists have a well-defined scope of what they wish to collect. The strength

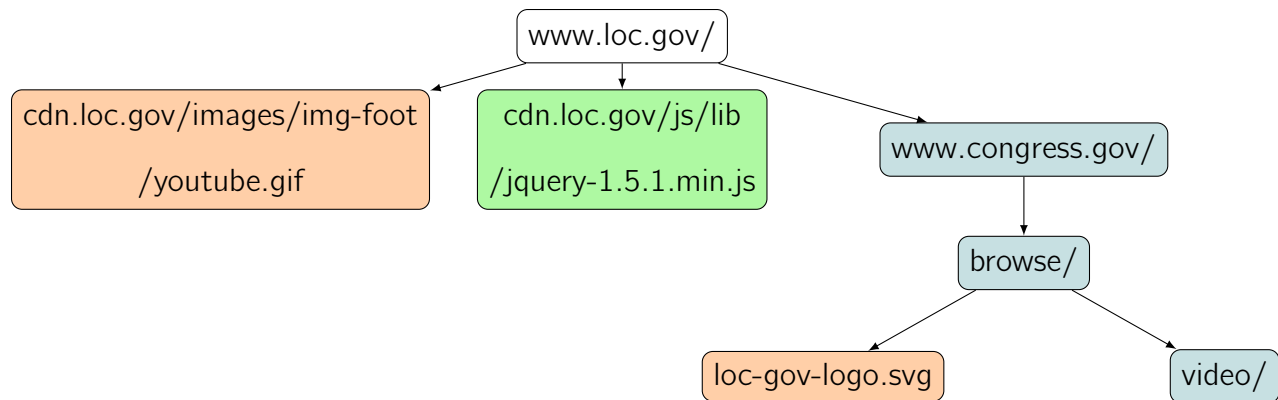


Figure 5.13. External web components linked to from the Library of Congress website, as of April 4, 2018. HTML pages are shown in blue, images are shown in orange, and JavaScript items are shown in green.

of this approach is that it is not necessary to differentiate between different types of content such as images and videos, as Brunelle, Kelly, SalahEldeen, et al. (2015) does. All that matters is whether the component is linked to from the seed URL and its sub-pages.

Another way of measuring the relevance of a website component to the rest of the archive is to examine the domain it is found in. Most websites today contain links to both *internal* and *external* components. Internal components, such as HTML pages, videos, and images, are found in the same domain as the seed URL. However, external components are found on other websites, usually with different domains. Figure 5.13 shows some external web components linked to from the main Library of Congress website. As can be seen from the graph, the seed website [loc.gov](http://loc.gov) links to JavaScript code and an image found on a different site [cdn.loc.gov](http://cdn.loc.gov). It also links to a separate site, the main website of the U.S Congress, [www.congress.gov](http://www.congress.gov).

As was seen in Chapter 4, web components on other websites can have a significant effect on the quality of the archived seed site. One way to assess whether these components are actually relevant to the seed URL and should be collected is by comparing the domain they are found in to the domain of the seed URL. For example, for the Library of Congress

graph, this would entail comparing [www.loc.gov](http://www.loc.gov) to [cdn.loc.gov](http://cdn.loc.gov) and [www.congress.gov](http://www.congress.gov). A number of similarity measures could be used to calculate the differences between the seed domain and these other domains. After the final results are obtained, domains with higher similarity scores would be classified as relevant and their contents would be archived, while domains with lower similarity scores would be seen as less relevant and not as necessary for creating a high-quality web archive. Equation 36 shows this similarity calculation using the cosine similarity.

$$(36) \quad \text{cosine}(S, D) = \frac{S \cdot D}{\|S\| \|D\|}$$

For example, the domain [www.loc.gov](http://www.loc.gov) could be called  $S$ , and then compared to domains  $D_1$ , [cdn.loc.gov](http://cdn.loc.gov), and  $D_2$ , [www.congress.gov](http://www.congress.gov). If cosine similarity is used for this comparison, the results are that  $\text{cosine}(S, D_1) = 0.30$  and  $\text{cosine}(S, D_2) = 0.16$ .  $D_1$  has greater similarity to the original seed domain, and thus can be seen as more relevant to the web archive. The full details of the similarity calculations are shown in Appendix F.

This application of cosine similarity to determine topic relevance has the same condition as the web graph distance in that it needs a limit or threshold value to be effective. Since not every website can be effectively archived, the web archivist should define a limit for the cosine similarity. Websites with cosine similarity scores above this threshold will be archived, while those with lower scores will not. This threshold will help to keep the web archive to a manageable size.

Both of these topic relevance measures are advantageous in that they can be applied *before* the web archive is actually created. Rather than accidentally collecting a large amount of irrelevant pages in a web archive, and then having to correct the problem, the web archivist can use these measures to prevent the problem from happening in the first place. Web archivists can begin by establishing specific limits and thresholds as to what will be collected, such as web components with up to a distance of 10 from the seed URL, or components in

domains with no less than a 0.2 similarity score when compared to the seed domain. This preventative approach can be important for the web archiving institution when trying to save time and money.

### 5.3.2. Size Relevance

A perfectly-archived website  $A$  can be called relevant if  $|A| = |O|$ , that is, the archived website is the same size as the original website. The size can be measured either in terms of number of documents or as the space it occupies on a disk. However, as has been explained in previous sections, this condition is rarely met, as it is more common for an archived website to be smaller or larger than its original. Therefore, the measure of size relevance needs to be adapted to fit a more realistic situation. An archived website  $A$  is relevant, if any of the following conditions apply:

- I. both  $|A| > |O|$  and  $|A| \not\gg |O|$  are true: the archived website is larger than the original website, but not *much* larger. The difference in cardinality between  $|A|$  and  $|O|$  must not exceed a certain user-defined limit, defined as  $k$ . Each web archivist determines how much larger she thinks the archived website can be when compared to the original. In this way:

$$(37) \quad |A| - |O| \leq k$$

- II. the archived website does not contain the same component  $c'$  repeated more than once. Every component is unique and appears only once in the archive.

The first condition is illustrated in Figure 5.14, which shows the original website  $O$ , an archived version of acceptable or good quality (depending on the archivist's judgment), and an archived version of poor quality.  $O$  has three components, which are mirrored in the first archive. However, this same archive also contains many more components, up to  $c'_k$ . As a result, it might still be considered an acceptable or even good-quality web archive because it

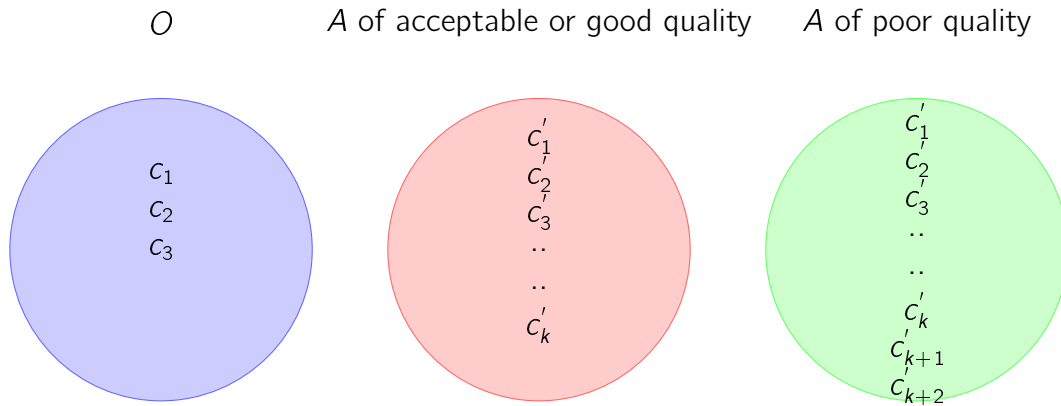


Figure 5.14. The first condition for size relevance: an archived website can be larger than the original, but not much larger. Its threshold must not exceed  $k$ .

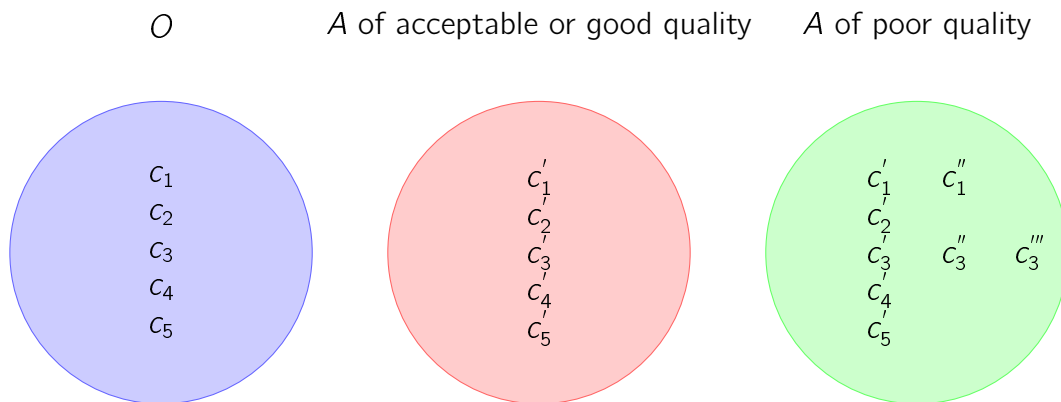


Figure 5.15. The second condition for size relevance: an archived website does not have the same component,  $c'$ , repeated more than once.

does not exceed the  $k$  limit. The second archive contains many more components than the  $k$  limit, and thus it is classified as being of poor quality.

The second condition is illustrated in Figure 5.15, which again shows the original website, a good-quality archived version, and a poor-quality archived version. In this case,  $O$  has five components, which are mirrored in the good-quality archive. However, the second archive contains two copies of the  $c_1$  component, named  $c'_1$  and  $c''_1$ . It also contains three copies of  $c_3$ , named  $c'_3$ ,  $c''_3$ , and  $c'''_3$ . Because the components in the second archive are not unique, it fails the second condition, and it is thus classed as a poor-quality archive.

#### 5.4. Operationalizing Archivability

There are a number of factors that could affect the archivability of a website, as noted in the data and the research of Banos and Manolopoulos (2015). The support tickets made it clear that dynamic components such as JavaScript and videos made a website much more difficult to archive. For example, JavaScript code could create endlessly repeating directories that lead to crawler traps, and render an archived website unusable. Given these findings, the archivability of a website is clearly a measure of *dynamism*, or the number of dynamic components, such as JavaScript and videos, contained in the website. The higher the dynamism of a website, the lower its archivability.

If the set of dynamic components for the original website  $O$  is represented as  $D_O$ , then dynamism can then be represented as the size of  $D_O$ ,  $|D_O|$ . This formulation is shown in Equation 38. For example, for the Library of Congress web graph shown in Figure 5.10, the cardinality of  $O$  is  $|O| = 9$ , since the graph has nine components. Two of these components, the JavaScript files *share.js* and *standards.js* are dynamic, therefore  $|D_O| = 2$ . The dynamism of this website then becomes 2.

$$(38) \quad \text{Dynamism} = |D_O|$$

Since archivability is inversely proportional to dynamism, it can be represented as in Equation 39. The archivability for the Library of Congress website then becomes  $1/2$  or  $0.50$ . If more dynamic components were added in the future, dynamism would increase and correspondingly, archivability would decrease.

$$(39) \quad \text{Archivability} = \frac{1}{\text{Dynamism}} = \frac{1}{|D_O|}$$

Because archivability is an a-priori condition of a website, it can be measured before a web archive is even created, like the dimension of topic relevance. Unlike other measures of

information quality, such as correspondence, it is not dependent on a comparison between the original website and its archived counterpart. Once archivability is known, the web archivist can design crawling strategies and other capture methods to increase the quality of the final archived version.

## CHAPTER 6

### DISCUSSION AND CONCLUSION

#### 6.1. Research Questions

##### 6.1.1. RQ1: What is the Human-Centered Definition of Information Quality (IQ) for Web Archives?

The data analysis yielded three dimensions, or core categories for IQ in web archives: correspondence, relevance, and archivability. Correspondence is the degree of similarity, or resemblance, between the original website and the archived website. Relevance is the pertinence of the contents of an archived website to the original website, while archivability is the degree to which the intrinsic properties of a website make it easier or more difficult to archive. The dimensions of correspondence and relevance each have several sub-dimensions, or sub-categories, as seen below:

#### I. Correspondence

- A. Visual correspondence: similarity in appearance between the original website and the archived website
- B. Interactional correspondence: the degree to which a user's interaction with the archived website is similar to that of the original
- C. Completeness: the degree to which the archived website contains all of the components of the original

#### II. Relevance

- A. Topic relevance: degree to which an archived website (or a web archive) includes only content that is closely related to that of the original website or the topic of the larger web archive
- B. Size relevance: the similarity in size of the archived website to the original website



### III. Archivability

Taken together, these three dimensions meet the requirements specified by Barney Glaser in *Theoretical Sensitivity* and discussed in Section 4.2. As core categories, they account for most of the behavior of web archivists towards the quality of web archives that was seen in the data.

#### 6.1.2. RQ2: How Can IQ in a Web Archive be Measured?

This dissertation proposed measures for every dimension and sub-dimension of IQ. A summary of these is given in Table 6.1. Of the dimensions of IQ discovered, relevance was the most difficult to operationalize for several reasons. First, the data showed that AIT clients expressed the notion of relevance or irrelevance in vague and imprecise ways, which required greater effort to interpret. Second, relevance has traditionally been seen as a dimension of IQ which depends entirely on the users' perception and no human *ground truth* judgments of quality were present due to the difficulties explained in Section 5.3. An added difficulty was the fact that web archivists' judgments of relevance were often incorrect.

The measures of relevance proposed in this dissertation attempt to approximate relevance as *judged by AIT web archivists and not AIT clients*. It seemed that only the AIT employees had enough knowledge and experience of web archiving to accurately judge which content was relevant to a web archive. In order to be successfully applied, the metrics for topic and size relevance both require a limit determined by the web archivist. For example, topic relevance requires a limit  $m$  for the index variable  $i$ , such that  $1 \leq m \leq N$ . Similarly, the metric for size relevance is also dependent on a threshold  $k$  such that  $|A| - |O| \leq k$ . Both of these variables need to be properly set by a knowledgeable and experienced web archivist.

Nevertheless, the metrics proposed for relevance and archivability have definite advantages. For one, unlike correspondence, they are not dependent on a comparison between the archived website and the original. Topic relevance, size relevance, and archivability can be calculated for a website *before* it is even archived. This is useful because it allows web

Table 6.1

*Dimensions of Information Quality in a Web Archive and their Corresponding Measures*

<b>Dimension</b>	<b>Measure</b>
<b>Correspondence</b>	
Visual Correspondence	$VC(O, A) = \frac{1}{ed(O, A)}$
Interactional Correspondence (Original Version)	$IC = \frac{ N'_E }{ N_O \cap N_A }$
Interactional Correspondence (Weighted Version)	$IC = \frac{w_i *  N'_{E_H}  + w_i *  N'_{E_M}  + w_i *  N'_{E_L} }{w_i *  N_{O_{A_H}}  + w_i *  N_{O_{A_M}}  + w_i *  N_{O_{A_L}} }$
Completeness (Cosine Similarity Version)	$\frac{\sum_{i=1}^n c'_i}{\sqrt{n * \sum_{i=1}^n c_i'^2}}$
<b>Relevance</b>	
Size relevance	$ A  -  O  \leq k$
Topic relevance	$TR(c_i) = \frac{1}{d(s, c_i)}$
	$cosine(S, D) = \frac{S \cdot D}{\ S\  \ D\ }$
<b>Archivability</b>	$Archivability = \frac{1}{Dynamism} = \frac{1}{ D_O }$

archivists to anticipate the size and nature of the web archive before it is created and helps them to prevent future problems or issues.

### 6.1.3. Unexpected or Surprising Findings

During the course of this research, many findings came to light that were surprising and unexpected. The first unexpected finding was that in the final theory, completeness is a sub-dimension of IQ (part of the correspondence dimension) rather than having its own dimension. This differs from previous theories and models of IQ in Information Science, Computer Science, and Philosophy. Those theories usually describe completeness as a major dimension (or category) of IQ. The status of completeness illustrates the differences between

web archives and other digital objects. In web archiving, the degree of similarity between the live website and its archived version are more important than the completeness of the latter.

As was discussed in Section 5.3, in traditional Information Retrieval research, users' judgments of relevance are taken to be the "ground truth". However, the data showed that AIT clients often made mistakes when judging whether or not archived content was relevant, regularly flagging web content as irrelevant, when it was actually necessary to properly display an archived website. AIT employees, who had a deep knowledge of the web archiving process, were usually the ones who were able to correctly distinguish between relevant and non-relevant elements in a web archive. This leads me to conclude that for the dimension of relevance in web archives, it is the perspective of an experienced web archivist (and not that of the end user) that should be considered as the ground truth.

This finding also had another consequence. At the beginning of this dissertation research, I intended to create a theory that stemmed purely from the perspective of the AIT clients; however, the results showed that the viewpoint of the AIT employees proved very significant. The final theory described in this dissertation includes the viewpoints of both AIT clients, who come from widely different technical backgrounds and experience, and AIT employees, who are seasoned web archivists. As a result, the final theory is broader in scope than originally intended. However, as Glaser and Strauss stated, the more a theory includes information about different groups of people, the greater its generalizability and predictability.

A second surprising finding about relevance was the importance of size relevance for AIT clients. As was noted in Section 4.2.3.2, AIT clients were worried as much about the overabundance of content in their web archives as about their completeness. This directly contradicted my initial assumption that the size of a web archive would not be important for them due to the amorphous nature of the Web and the abundance of storage space. Size relevance emerged as a major sub-category and was therefore operationalized. Whereas topic relevance has been addressed in the literature by other researchers, size relevance has

not. There has been no previous work stating that an archived website might be deemed not relevant simply because of its size and not its intellectual content.

The special nature of archivability as a dimension of IQ was also one of the more surprising findings. In Section 4.2.4, archivability was described as a latent dimension, because was not obvious to most AIT clients. Only AIT employees, who have a deep knowledge of an experience with the technical process of archiving websites, were able to determine a website's archivability and judge how it will affect the quality of its archived counterpart. This is a new finding that has not been seen in the previous work on archivability.

## 6.2. Limitations of the Study

There is no research study that is carried out without any issues, challenges, or unexpected twists and turns. The following section describes the issues that arose and how they were addressed, as well as other unexpected challenges I encountered. Furthermore, the scopes and limitations of both the study and the theory are explained.

### 6.2.1. Methodological Issues

During grounded theory research, several methodological issues can arise. The most serious one is that the theory created using the GT approach might not properly describe the actual data. For this dissertation, this issue was averted through the use of purposeful peer review. Committee members were periodically invited to audit the entire research project, including the codebook, preliminary findings, and core categories. In addition to the committee members, employees of the Internet Archive were also invited to see the findings. Furthermore, in the summer of 2017, I presented the theory and my preliminary findings at the doctoral consortium of the Joint Conference of Digital Libraries, where I received feedback from my peers and other Information Science researchers.

Transparent documentation and rich descriptions were also provided in the dissertation. This involved including the original research agreement, anonymization process, examples of tickets, and the entire codebook in the dissertation. In Chapter 4, for each core category or

sub-category that is introduced, I present many examples of it that come from the raw data. All of these efforts were made to enhance the credibility, dependability, confirmability, and transferability of the IQ theory.

Another issue that arose during Phase 2 was that my original intent to use Lazarsfeld's panel analysis method to operationalize the dimensions of quality was not entirely successful. There were several reasons for this. First, the panel analysis approach, as described in Table 3.2 makes heavy use of time as an important variable to take into account. While time is indeed an important variable that should be considered for web archives, it did not arise as a category (indeed it rarely surfaced) in the data. According to the rules of classical grounded theory, a category cannot be created if it is not present in the data, and thus it was not included in the final theory. Additionally, I felt that research into web archives was not yet sufficiently mature so as to provide an in-depth understanding of how time affects web archives.

Despite these issues, Lazarsfeld's general principles of qualitative mathematics proved invaluable while operationalizing the dimensions of quality. His process for translating data into an empirical index, described in Section 3.2 was followed during Phase 2. Lazarsfeld notion of the interchangeability of indices also proved invaluable when crafting the similarity measures, as it states that substituting one index for another, or adding additional indices to the formula is unlikely to change the direction of the general relationship. As a result, I was able to pick a similarity measure (say, cosine similarity) to explain an IQ dimension, but was also able to say that other measures such as Jaccard similarity or Euclidean distance would also prove useful. Overall, Lazarsfeld methods, while not followed to the letter, provided me with flexible principles that enabled creative and adaptable ways of operationalizing quality.

#### 6.2.2. Scope and Limitations

The theory presented in this dissertation is a *substantive* theory, that is, it is specific to the context of web archiving and not meant to describe the construct of IQ in a more general

form. The theory is delimited because it is specific to small or medium-size web archives that are focused on covering a single topic or an event. It is not meant to describe larger web archives such as the .gov or .fr, which preserve an entire country's national domain. The theory also makes other important assumptions that are reiterated here:

- Countable and finite: the original and archived websites form a finite, countable set
- Closed World Assumption (CWA): all the elements of a website are represented in the components,  $c_n$ , of the set  $O$  or  $c'_n$  of the set  $A$
- Web page vs. website vs. web archive: the theory focuses mostly on IQ at the webpage level; however, some dimensions such as topic and size relevance are more appropriately measured at the website and web archive level, as seen in Table 6.2

Table 6.2

*Dimensions of IQ and the Levels to Which They are Best Applied*

<b>Dimension</b>	<b>Best Applied To</b>
<b>Correspondence</b>	
Visual Correspondence	Webpage
Interactional Correspondence	Webpage
Completeness	Webpage, Website, Web Archive
<b>Relevance</b>	
Size relevance	Web Archive
Topic relevance	Webpage, Website, Web Archive
<b>Archivability</b>	Webpage, Website

It is also important to note that the correspondence of an archived webpage might not always be easily measurable. For example, if the original site has been lost, there is no way to compare it to the archived version, so a measure of correspondence cannot be calculated. The operationalized definitions of correspondence presented in this dissertation assume that

there exists a live version of a website to which the archived version can be compared.

Another complicated issue to consider is that it is becoming increasingly difficult to determine what is the “ground truth” version of a website. As pointed out by Ainsworth et al. (2014) and Ainsworth and Nelson (2015), webpages are composite objects, containing a multitude of components such as images, CSS style sheets, videos, and JavaScript files. As the Internet becomes more dynamic, webpages and websites are increasingly personalized to a specific user’s needs: a user’s profile page on a social media website will be markedly different from another user’s. Also, many websites have versions tailored to the user’s platform (mobile websites differ from “desktop” websites), their geographical location, and other variables. Given that many million of users see only their own version of a website, which version is the “true” website that should be preserved? No entity or technology can possibly archive all versions of a website.

Though this is a significant concern for web archivists and institutions involved in preserving cultural heritage, it is simply not borne out by the data analyzed for this project. In the AIT tickets, the clients did not express concern for preserving multiple versions of a website. This may be because highly customized websites might not yet be common enough for them to be a significant concern for AIT clients. Generally, AIT users expressed very clear ideas as to what a high-quality archived website should look and behave like; there was a single, clear version of a website that they wished to preserve, not multitudinous personalized versions.

### 6.3. Contributions of the Study

This theory of IQ for web archives meets the standards set forward by Glaser and Strauss in *The discovery of grounded theory* and discussed in section 3.1.2. Glaser and Strauss (2009) require that a good grounded theory closely fit the data and also be clear, understandable, applicable to a multitude of diverse daily situations, and flexible. The theory advanced here is substantive, heavily grounded in the data, and meant for measuring the IQ

of web archives. It contains no dimension or sub-dimension of IQ that is not reflected in the data.

Furthermore, the theory is explained in a thorough manner and illustrated through the use of examples. It is also applicable to many situations found in web archiving, and accounts for many common quality problems with web archives, such as lack of completeness, and lack of correspondence between the original website and its archived counterpart. The operationalizations of the IQ dimensions given in Chapter 5 are highly flexible and open to being modified by other researchers or by web archivists themselves. Throughout the section, I present metrics that correspond to each quality dimension, but emphasizes that other metrics could also be utilized.

The theory, as presented in Chapter 4, is also independent of the technology that is currently in use for creating web archives. If, in the future, there was a shift away from using crawlers and the Wayback Machine to create and view archived websites, the definitions of correspondence, relevance, and archivability would remain the same. The operationalized definitions of these dimensions, as presented in Chapter 5, might need to be altered somewhat, but they are still general enough to still be applicable and useful.

The theory presented in this dissertation also has theoretical completeness, defined by Glaser (1978) as the ability to explain as much variation as possible with the fewest possible concepts and the greatest possible scope. The final theory has three major dimensions and five sub-dimensions. Taken together they represent the great majority of IQ problems seen in topic-centered or event-driven web archives today. As the first theory developed specifically about web archives, it lays the groundwork for future theoretical developments in the field.

## 6.4. Future Directions

### 6.4.1. Applying the Operationalized Definitions of IQ

The purpose of operationalizing the dimensions of IQ was to develop metrics that would enable web archivists to measure the quality of their web archives. Several of the



metrics presented can be applied

In order to measure the visual correspondence of a web archive, a program could be developed that is similar to the VQI system implemented by the Swiss National Library (described in Section 2.5.3.3). This program would work by navigating to both the live website and its archived counterpart, taking screenshots of both, and then calculating the Euclidean distance between them. There are many current tools that have the ability to take website screenshots, such as headless browser software PhantomJS (“PhantomJS (Version 2.1.1)”, 2016), and the software-testing suite Selenium (“Selenium (Version 3.81)”, 2017). Image comparison software would then be used to calculate the distance between the screenshot images.

In Section 5.2.2, errors in interactional correspondence were detected by using the Network Monitor tool present in the Firefox browser (“Firefox Quantum (Version 57.0.4)”, 2017). This tool shows all the network requests (and their errors) that are behind the functionality of a site. The interactional correspondence between a live website and an archived website could be measured by employing these tools to detect dissimilar interactions, and then calculating the final value using the formula provided.

Completeness can be measured in several ways. The first and easiest one for existing Archive-It users would be to examine the crawl logs of their web archives, and then extract a list of files contained in the archived website. Then they could use software such as the Screaming Frog SEO Spider (“Screaming Frog SEO Spider (Version 9.2)”, 2018) to obtain a list of the files present in a live website. A program could be written that would then compare the two file lists, detect any missing files in the archived website, and calculate the degree of completeness. If the crawl logs are not available, then software that generates a list of the files present in the web archives could also be used.

Measuring size relevance would require similar software to that used for completeness. The crawl logs could be used to generate an estimate of the size of the web archive and then

compare it to the size of the live website. Link analysis software such as Screaming Frog SEO Spider (“Screaming Frog SEO Spider (Version 9.2)”, 2018) can also generate estimates as to the size of a website. In order to measure topic relevance, an idea of the link structure of the website would be needed. Software such as Screaming Frog SEO Spider (“Screaming Frog SEO Spider (Version 9.2)”, 2018) or the Archives Unleashed Toolkit (“Archives Unleashed Toolkit”, 2018) are adept at creating link graphs that represent the network structure of a website. One of the most challenging aspects of this process might be determining the threshold value,  $k$  or  $m$ , that will help to delimit the size of the web graph.

Like the other dimensions, archivability would not be too difficult to measure. It would simply require a list of the dynamic elements of a website. The Archive-It platform has the ability to generate lists of the files present in an archived website, and sort them by type (image, JavaScript, HTML). Link analysis software and even the Network Monitor tool could be employed to do the same for the live version of the site. These tools would help the web archivist measure the dynamism of a website.

Once the software to measure IQ has been built, experiments could be carried out to determine which metrics perform best. For example, in this dissertation, I presented cosine similarity as a measure for calculating completeness; however, other similarity measures such as Euclidean distance or Jaccard similarity might also be appropriate. The same situation would apply to measuring visual correspondence or topic relevance. Lazarsfeld principle of the interchangeability of indices might likely come into play, where some measures might be only marginally better than others at measuring quality, but the general direction of the relationship remains unchanged.

In some instances, the correct formula for an IQ dimension might be more complex than originally stated. For example, visual correspondence is defined as  $VC(O, A) = \frac{1}{ed(O, A)}$ . In reality, the formula might be something closer to  $VC(O, A) = \frac{\alpha}{ed(O, A)}$ , where  $\alpha$  is some constant with a limited range of values. Details such as these would need

time and effort to be adequately worked out.

The tasks described here are very time-consuming, but are not conceptually difficult to carry out. The most difficult challenge lies, not in applying the metrics, but in securing access to a dataset that would allow these metrics to be used and refined. In order to do this, a researcher agreement with an institution that possesses web archives would be necessary. Such an agreement would provide me with the ability to access the web archives as well as grant me permission to run the IQ measuring software on their web archives. This might present additional legal issues and complications that would necessitate some time to sort out.

#### 6.4.2. Other Research Directions

As I coded the support tickets submitted by AIT clients, I realized that the expectations they had for web archives were often in conflict with the practical realities of web archiving. For example, the clients often assumed that a website had a specific size which would also be reflected in the archived site. Since the original website had  $X$  number of documents, it would also follow that the archived website also had  $X$  number of documents. However, the tickets analyzed showed that the reality did not reflect their expectations. This mismatch is summarized by the code “expectations vs. reality”, described in the NVivo codebook in Table D.3. This findings points to some promising research topics regarding the mismatch between how humans perceive web archives and how web archives are actually constructed. I have already published some preliminary work on this topic in the JCDL workshop paper “Web archives: A preliminary exploration of user expectations vs. reality”, and might develop it further in the future (Reyes Ayala, 2017).

Another possible research direction involves exploring the notion of time and its effect on the IQ of a web archive. Other researchers such as Ainsworth et al. (2014) and Brunelle, Kelly, Weigle, and Nelson (2015) have previously explored the effects of time on coherence and archivability, respectively. Future research could focus on how time can affect the overall

IQ of a website, not just a single category or dimension.

APPENDIX A

RESEARCHER AGREEMENT WITH ARCHIVE-IT



# Web Data Research Agreement



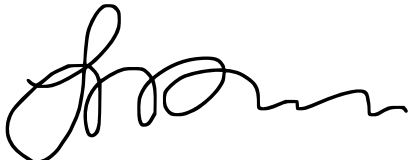

**THIS AGREEMENT** is made on the 1st day of July, 2016, by and between Internet Archive (IA) based in San Francisco, CA and Brenda Reyes Ayala, Phd Candidate in Information Science, University of North Texas.

**I.** IA agrees to provide Archive-It partner support ticket data for tickets submitted over a period of one year, from January 1, 2012 - December 31, 2012. This data set will include initial partner query and any IA staff responses in xml format.

**II.** Signatory below agrees that he shall:

- a. NOT provide partner support ticket information to third parties without express permission of Internet Archive.
- b. Anonymize any personal or institutional information from the tickets for publication purposes. This includes: Names (personal or institutional), institutional web domains, other potentially identifying information.
- c. Provide no third-party replay of archived URLs beyond linking to their existing availability through Archive-It or Internet Archive.
- d. Cite Internet Archive or Archive-It in publications, presentations, blog posts, or similar outputs related to use and study of this collection.
- e. Provide an opportunity for IA staff to review findings and suggest factual corrections at least 4 weeks prior to submission for any formal or informal publication.
- f. Uphold Internet Archive's Terms of Use posted at <http://archive.org/about/terms.php>.

This agreement shall be binding upon the parties, their successors, assigns and personal representatives.

INTERNET ARCHIVE	RESEARCHER OR USER
By: 	By: 
Name: Jefferson Bailey	Name: Brenda Reyes Ayala
Title: Director, Web Archiving Programs	Title: PhD candidate
	Institution: University of North Texas
Date:	Date: 07 / 05 / 2016

## APPENDIX B

A SAMPLE ARCHIVE-IT SUPPORT TICKET WITH XML TAGS

```
<ticket>
  <assigned-at type="datetime">2013-03-27T17:35:39-08:00</
    assigned-at>
  <assignee-id type="integer">54229397</assignee-id>
  <base-score type="integer">80</base-score>
  <created-at type="datetime">2013-03-20T09:50:48-08:00</
    created-at>
  <current-collaborators nil="true"/>
  <current-tags nil="true"/>
  <description>Yesterday, I crawled http://museum.wordpress.com/. It seemed like it was getting way more documents than I anticipated, so I stopped the crawl after a few hours. I realize that would make it miss some pages, but the display (http://wayback.archive-it.org/3711/20130314205229/http://museum.wordpress.com/) is pretty weird-looking. Was this caused by stopping the crawl early & missing essential content that effects the display? Probably should have done a test crawl first: -)</description>
  <due-date type="datetime" nil="true"/>
  <entry-id type="integer" nil="true"/>
  <external-id nil="true"/>
  <group-id type="integer">20004587</group-id>
  <initially-assigned-at type="datetime">2013-03-24T17:59:48-08:00</initially-assigned-at>
```



```
<latest-agent-comment-added-at type="datetime" nil="true"/>
<latest-public-comment-added-at type="datetime" nil="true"/>
<nice-id type="integer">403</nice-id>
<number-of-incidents type="integer">0</number-of-incidents>
<organization-id type="integer" nil="true"/>
<original-recipient-address nil="true"/>
<priority-id type="integer">0</priority-id>
<recipient nil="true"/>
<requester-id type="integer">60437793</requester-id>
<resolution-time type="integer" nil="true"/>
<solved-at type="datetime" nil="true"/>
<status-id type="integer">2</status-id>
<status-updated-at type="datetime">2013-03-30T10:14:03-08:00
  </status-updated-at>
<subject>Archived copy isn't displaying the way I
  anticipated</subject>
<submitter-id type="integer">60437793</submitter-id>
<ticket-type-id type="integer">2</ticket-type-id>
<updated-at type="datetime">2013-03-30T10:14:03-08:00</
  updated-at>
<updated-by-type-id type="integer">0</updated-by-type-id>
<via-id type="integer">0</via-id>
<via-reference-id type="integer" nil="true"/>
<score type="integer">80</score>
<problem-id nil="true"/>
```

```
<has-incidents type="boolean">false</has-incidents>
<comments type="array">
  <comment>
    <author-id type="integer">60437793</author-id>
    <created-at type="datetime">2013-03-25T09:50:48-08:00</
      created-at>
    <is-public type="boolean">true</is-public>
    <type>Comment</type>
    <value>Yesterday, I crawled http://museum.wordpress.com
      /. It seemed like it was getting way more documents
      than I anticipated, so I stopped the crawl after a
      few hours. I realize that would make it miss some
      pages, but the display (http://wayback.archive-it.org
      /3711/20130314205229/http://museum.wordpress.com/) is
      pretty weird-looking. Was this caused by stopping
      the crawl early & missing essential content that
      effects the display? Probably should have done a
      test crawl first: -)</value>
    <via-id type="integer">0</via-id>
    <attachments type="array"/>
  </comment>
  <comment>
    <author-id type="integer">54229397</author-id>
    <created-at type="datetime">2013-03-26T17:59:48-08:00</
      created-at>
```

```
<is-public type="boolean">true</is-public>
<type>Comment</type>
<value>Hi Y,
```

It actually looks like this crawl ran to completion (<https://partner.archive-it.org/archiveit/partner/crawl/report/seed.html?accountId=619&crawlJobId=54746&cid=45734&conversationPropagation=join>), however it looks like some of the stylesheets are blocked by robots.txt. I'm running a quick test to confirm and I'll get back to you as soon as I have more information.

Best,

I,

Partner Specialist, Internet Archive

```
</value>
```

```
<via-id type="integer">0</via-id>
```

```
<attachments type="array"/>
```

```
</comment>
```

```
<comment>
```

```
<author-id type="integer">60437793</author-id>
```

```
<created-at type="datetime">2013-03-27T13:53:29-08:00</
```

```
created-at>
```

```
<is-public type="boolean">true</is-public>
```

```
<type>Comment</type>
```

```
<value>Thanks! Iâ€™ve archived Wordpress blogs before  
and not have problems, but they were ones we run here  
at the library, so they may be a bit different.
```

-Y

```
</value>
```

```
<via-id type="integer">4</via-id>
```

```
<attachments type="array"/>
```

```
</comment>
```

```
<comment>
```

```
<author-id type="integer">54229397</author-id>
```

```
<created-at type="datetime">2013-03-38T12:33:24 -08:00</  
created-at>
```

```
<is-public type="boolean">true</is-public>
```

```
<type>Comment</type>
```

```
<value>Hi Jay,
```

Sorry for the delay! I just wanted to let you know that the test that I ran did not clear up the issue, so I'm having one of our engineers look into this further. I'll let you know once I have additional information.

Best,

I

Partner Specialist , Internet Archive

</value>

<via-id type="integer">0</via-id>

<attachments type="array"/>

</comment>

<comment>

<author-id type="integer">60437793</author-id>

<created-at type="datetime">2013-03-38T14:19:33 -08:00</

created-at>

<is-public type="boolean">true</is-public>

<type>Comment</type>

<value>Hi I,

Sorry it's being such a weird problem, but glad to know it  
evidently isn't just me!:-)

-Y

</value>

<via-id type="integer">4</via-id>

<attachments type="array"/>

</comment>

<comment>

```
<author-id type="integer">54229397</author-id>
<created-at type="datetime">2013-03-38T17:35:40 -08:00</
  created-at>
<is-public type="boolean">true</is-public>
<type>Comment</type>
<value>Hi Y,
```

Upon further investigation, it looks like the stylesheets are embedded in such a way that they are not being captured by default. If you expand scope to include URLs that contain the following text: <http://s0.wp.com/wp-content/themes/> you should be able to crawl the site again and capture the stylesheets and this site should display normally.

Do let me know if this doesn't solve the problem or if you have any further questions!

Best,

I

Partner Specialist, Internet Archive

```
</value>
```

```
<via-id type="integer">0</via-id>
```

```
<attachments type="array"/>
```

```
</comment>
```

```
<comment>
```

```
<author-id type="integer">60437793</author-id>
<created-at type="datetime">2013-03-39T08:07:25 -08:00</
  created-at>
<is-public type="boolean">true</is-public>
<type>Comment</type>
<value>Great, thanks! Iâ€™ll give it a try.
```

-Y

```
</value>
  <via-id type="integer">4</via-id>
  <attachments type="array"/>
</comment>
</comments>
<ticket-field-entries type="array"/>
<linkings type="array"/>
<channel nil="true"/>
<permissions>
</permissions>
</ticket>
```

## APPENDIX C

### ANONYMIZING THE SUPPORT TICKETS



This appendix presents the process I used to anonymize the support tickets received from the Internet Archive's AIT Team. Unless otherwise stated, none of the URLs presented in this dissertation actually exist.

The following methods were employed to remove potentially identifying information from the data. All of the data presented has been anonymized using *at least two* of these methods; a significant amount of data has had all three anonymization methods applied to it.

- I. Obscuring the domain names. Some URLs have had the domain name obscured in their anonymous form:

Original: <http://www.somewebpage.org/remembering/life-work>

Anonymized: [http://www.\\_\\_.org/remembering/life-work](http://www.__.org/remembering/life-work)

Original: <http://public-museum.uk/roman-scrolls>

Anonymized: [http://\\_\\_.uk/roman-scrolls](http://__.uk/roman-scrolls)

- II. Scrambling or replacing the words in a URL with others. Tickets with mentions of institutions had their domain names scrambled or replaced with non-existent names:

Original: <http://www.governmentagency.gov/ourteam>

Anonymized: <http://www.pl.gov/tef>

Original: I captured one video in my last crawl with collection StateCollege Social Media

Anonymized: I captured one video in my last crawl with collection MSU Social Media

- III. Replacing named entities with others. URLs that contained names of places, people, and things were replaced with those of others:

Original: <http://www.famousenglishwriter.uk.gov>

Anonymized: <http://mishima.jp>

Original: [www.NameofAcademySports.org](http://www.NameofAcademySports.org)

Anonymzed: [www.oursports.edu](http://www.oursports.edu)

APPENDIX D  
NVIVO CODEBOOK

This appendix contains the entire codebook used for coding the data in Nvivo. Not all the categories present in the codebook were used for the final analysis and theory.

Table D.1

*Nvivo Codebook for Correspondence and Its Subcategories*

<b>Hierarchical Category Name</b>	<b>Description</b>	<b># Tickets Containing Code</b>	<b># Coding References</b>
<b>correspondence</b>	Degree of similarity between the archived website and the original website.	226	852
appearance of archived website	The archived website has a different appearance from the original, or there is something else wrong with it.	91	160
completeness	Refers to the completeness of an archived website. The desired content has not been captured or is not present in the archive.	157	478
completeness/ add more content	User wishes to add more content to web archive.	69	122
completeness/ add more content/ expanding scope to include more content	The user or partner specialist will expand the scope of the crawl using the AIT interface in order to capture more content.	29	35

Continued on next page

**Table D.1 – continued from previous page**

<b>Hierarchical Category Name</b>	<b>Description</b>	<b># Tickets Containing Code</b>	<b># Coding References</b>
completeness/ add more content/ General add content	Add content, not a specific method to do it.	16	24
completeness/ add more content/ ignore robots.txt	The user is requesting to ignore robots.txt. Robots.txt is a file created by a website that determines if and what a crawler can crawl. Sometimes, robots.txt can keep a crawler from capturing desired content.	34	41
completeness/ aparent display is sue is actually a capture issue	The component of the website does not display appropriately; however the underlying problem is that the content has not been captured.	23	23
completeness/ blocked by robots.txt	A site's robots.txt file does not allow it to be crawled. By default the AIT crawler follows the robots.txt settings; however, sometimes these rules need to be ignored in order to fully capture a site.	26	27

Continued on next page

**Table D.1 – continued from previous page**

<b>Hierarchical Category Name</b>	<b>Description</b>	<b># Tickets Containing Code</b>	<b># Coding References</b>
completeness/ general completeness	General completeness issue, unspecified.	111	163
explicit comparison between the archived website and the original	The partner draws an explicit comparison between the archived website and the original website.	47	64
implicit comparison between the archived website and the original	The partner draws an implicit comparison between the archived website and the original website. The original website is not mentioned; however, the user has a strong idea of what the archived website should look or behave like.	52	72
user interaction is different	The user's interaction with the site is different from that of the original, unexpected, or deficient.	49	72

Table D.2

*Nvivo Codebook for Relevance and Its Subcategories.*

<b>Hierarchical Category Name</b>	<b>Description</b>	<b># Tickets Containing Code</b>	<b># Coding References</b>
<b>relevance</b>	The archived website contains content that matches the collecting scope of the web archive, in both content and quantity.	127	451
size_relevance	The web archive contains websites in quantity or volume that is unexpected or excessive.	107	351
size_relevance/ other_size_relevance	Other, miscellaneous types of size relevance.	1	1
size_relevance/ removing or reducing content	Removing some content from an already-existing web archive or a crawl. Different than blocking access to it. Can also encompass reducing the amount of content.	73	126
size_relevance/ removing or reducing content/ delimiting	An effort to restrict the capture of content. Perhaps to keep undesirable or irrelevant content from being captured.	63	103

Continued on next page

**Table D.2 – continued from previous page**

<b>Hierarchical Category Name</b>	<b>Description</b>	<b># Tickets Containing Code</b>	<b># Coding References</b>
size_relevance/ removing or re- ducing content/ general removing content	Removing content, method unspecified.	7	7
size_relevance/ removing or re- ducing content/ out of scope	This is an in-vivo code. The user may be referring to: 1) urls that were not captured and are shown as “out of scope” in the AIT interface or 2) content that is outside the collecting scope of the web archive or institution. Users may or may not want to capture out of scope content.	10	11
size_relevance/ too much content	User believes there is too much content being captured.	65	167
size_relevance/ too much con- tent/ crawler trap	In-vivo code. A crawler trap occurs when the crawler “gets stuck” crawling a site. This can cause a number of problems, such as: a crawl is stalled, it never finished, or it results in quality issues.	24	35

Continued on next page



**Table D.2 – continued from previous page**

<b>Hierarchical Category Name</b>	<b>Description</b>	<b># Tickets Containing Code</b>	<b># Coding References</b>
size_relevance/ too much content/ duplicate content	The web archive has duplicates of content, which is thought of as unnecessary.	10	17
size_relevance/ too much content/ general too much content	Too much content, reason not specified.	48	62
topic relevance	The web archive or archived website contains content about a different content than is expected or desired.	54	93
topic relevance/ irrelevant content	Content that is deemed not appropriate or not relevant to the collection or crawl.	46	81
topic relevance/ irrelevant content/ blocking access to irrelevant, low-quality or undesirable content	The user seeks to restrict Wayback access to irrelevant or undesirable content. Also used to block access to low-quality content.	19	30

Continued on next page

**Table D.2 – continued from previous page**

<b>Hierarchical Category Name</b>	<b>Description</b>	<b># Tickets Containing Code</b>	<b># Coding References</b>
topic relevance/ relevant content	A reference to content that is desired. The users usually want to capture this content because of various reasons: a) it is in-line with the institution's collecting guidelines and so is of interest b) it is part of a website that the user wishes to capture.	12	12
unknown relevance	Content whose relevance to the collection or crawl is unknown.	7	7

Table D.3

*Nvivo Codebook for Other Categories.*

<b>Hierarchical Category Name</b>	<b>Description</b>	<b># Tickets Containing Code</b>	<b># Coding References</b>
archivability	The site has intrinsic qualities that make it impossible or difficult to archive or replay.	78	101
between-crawl comparison	The user compares a current crawl to a previous crawl, usually the size or number of urls.	21	30
compromise	The client or partner specialist talks about a trade-off. It is unlikely that all aspects of quality can/will be achieved.	4	4
confusion	User experiences confusion.	7	8
expectations vs. reality	There is a mismatch between how a user thinks a web archive or single archived website works, and how it actually works.	18	24

Continued on next page

**Table D.3 – continued from previous page**

<b>Hierarchical Category Name</b>	<b>Description</b>	<b># Tickets Containing Code</b>	<b># Coding References</b>
expectations vs.reality/ all URLs are the same	Users did not draw fine distinctions (or any distinction at all) between the concepts of domain and sub-domain in URLs, or how types of URLs are treated differently by a crawler, such as unt.edu and www.unt.edu	14	18
goodness	Encompasses the ways in which people define what is normal, correct, or proper for an archived website.	84	125
inconsistent quality	Usually occurs when a user compares two web archives, or different captures of the same content. One capture exhibits different quality from the other.	6	6
notion of good enough	There is an acceptable, “good enough” level of quality.	12	15
pre-emptive question	The client asks for help prior to taking any action in order to prevent or minimize problems.	14	14
problem cannot be resolved	There is no solution to the quality problem at the moment.	14	15

Continued on next page

**Table D.3 – continued from previous page**

<b>Hierarchical Category Name</b>	<b>Description</b>	<b># Tickets Containing Code</b>	<b># Coding References</b>
space constraints	There are limitations as to the amount of space or number of documents a crawl can have. Sometimes these concerns are articulated very clearly (my crawl is 7GB when it should be 2GB) or sometimes very vaguely.	9	10
temporal incoherence	The site has changed and the archive has not changed to match it. The user may or may not be aware of this change and perceives it as a problem.	10	10
workaround	A workaround solution is a partial or temporary solution to the quality problem.	15	18

## APPENDIX E

### COMPLETENESS IN AN ARCHIVED WEBSITE USING THE JACCARD SIMILARITY

This appendix contains a mathematical exploration of completeness using the Jaccard similarity.

$$(40) \quad J(x \cap x') = \frac{|x \cap x'|}{|x \cup x'|} = \frac{|x \cap x'|}{|x| + |x'| - |x \cap x'|}$$

Let us define two variables  $x$ , the original website and  $x'$ , the archived website. We can operationalize completeness as the Jaccard similarity between  $x$ , the original website and  $x'$ , the archived website, as seen in Equation 40

If we express  $x$  and  $x'$  as bit vectors  $X$  and  $X'$  that contain all the components,  $c$ , of a website, such as text, images, video, etc, then the Jaccard similarity becomes:

$$(41) \quad J(X \cap X') = \frac{X \cdot X'}{|X|^2 + |X'|^2 - X \cdot X'}$$

$$X = \langle c_1, c_2, c_3, c_4, c_5, \dots, c_n \rangle$$

$$X' = \langle c'_1, c'_2, c'_3, c'_4, c'_5, \dots, c'_n \rangle$$

- In Jaccard similarity, the values of a vector can be binary, that is, 0 or 1. Let us assume that the value of each component,  $c$ , is also binary. So  $c_n = 0$  if the component is absent, and  $c_n = 1$  if the component is present.
- Let us assume that the original website,  $X$ , always has all of its components, so  $X = \langle 1, 1, 1, 1, 1, \dots, 1 \rangle$
- The archived website,  $X' = \langle c'_1, c'_2, c'_3, c'_4, c'_5, \dots, c'_n \rangle$ , since we do not yet know the values of  $X'$ .

Substituting these values into the equation, we get the following:

$$J(X \cap X') = \frac{\langle 1, 1, \dots, 1 \rangle \cdot \langle c'_1, c'_2, \dots, c'_n \rangle}{|\langle 1, 1, \dots, 1 \rangle|^2 + |\langle c'_1, c'_2, \dots, c'_n \rangle|^2 - (\langle 1, 1, \dots, 1 \rangle \cdot \langle c'_1, c'_2, \dots, c'_n \rangle)}$$

then we calculate the square of the dimensions of the original site and the archived site:

$$|X|^2 = \sum_{i=1}^n 1 = n$$

$$|X'|^2 = \sum_{i=1}^n c_i'^2 = c_1'^2 + c_2'^2 \dots + c_n'^2$$

As well as their dot product:

$$X \cdot X' = \langle 1, 1, \dots, 1 \rangle \cdot \langle c_1', c_2', \dots, c_n' \rangle = \sum_{i=1}^n (1)c_i' = \sum_{i=1}^n c_i'$$

Substituting these values into the equation, we get the following, generalized version of completeness:

$$(42) \quad J(X \cap X') = \frac{\sum_{i=1}^n c_i'}{n - \sum_{i=1}^n c_i'^2 - \sum_{i=1}^n c_i'}$$



## APPENDIX F

### MEASURING TOPIC RELEVANCE USING COSINE SIMILARITY

For the Library of Congress graph in Figure 5.13, a measure of topic relevance could be calculated by comparing the seed domain [www.loc.gov](http://www.loc.gov) to the domains [cdn.loc.gov](http://cdn.loc.gov) and [www.congress.gov](http://www.congress.gov). Cosine similarity is a measure of the cosine of the angle between two vectors. In this case, the domain names would be represented as vectors, with the seed domain [www.loc.gov](http://www.loc.gov) being vector  $S$ , and the domains [cdn.loc.gov](http://cdn.loc.gov) and [www.congress.gov](http://www.congress.gov) being vectors  $D_1$  and  $D_2$ . The cosine similarity between the seed domain  $S$  and any domain  $D$  is then shown in Equation 43.

$$(43) \quad \text{cosine}(S, D) = \frac{S \cdot D}{\|S\| \|D\|}$$

Table F.1

*Vector Representation of the Domain Names  $S$ ,  $D_1$ , and  $D_2$*

<b>Letter</b>	<b>Frequency in <math>S</math></b>	<b>Frequency in <math>D_1</math></b>	<b>Frequency in <math>D_2</math></b>
c	1	1	1
d	0	1	0
e	0	0	1
g	1	1	1
l	1	1	0
n	0	1	1
o	2	2	2
r	0	0	1
s	0	0	2
v	1	1	1
w	3	0	3
.	2	2	2

The domains  $S$ ,  $D_1$ , and  $D_2$  now need to be properly represented as vectors. A common approach to representing textual information in the field of Information Retrieval is to create a list of the letters present in all the strings of text. Then counting the number of times each letter appears in the text string (the frequency). The frequency values will form the vector representation of the text string. Table F.1 shows this process as applied to the three domains. According to the table, the vector  $S$  now has the value  $S = \langle 1, 0, 0, 1, 1, 0, 2, 0, 1, 3, 2 \rangle$ , while  $D_1 = \langle 1, 1, 0, 1, 1, 1, 2, 0, 0, 1, 0, 2 \rangle$  and  $D_2 = \langle 1, 0, 1, 1, 0, 1, 2, 1, 2, 1, 3, 2 \rangle$ . The cosine similarity between these vectors can now be calculated. The results are:

$$\text{cosine}(S, D_1) = 0.30$$

$$\text{cosine}(S, D_2) = 0.16$$

From the results, it can be seen that  $D_1$ , the domain [cdn.loc.gov](http://cdn.loc.gov) is more similar to the seed vector  $S$  than  $D_2$ . Therefore, it can be concluded that the contents of [cdn.loc.gov](http://cdn.loc.gov) are more relevant to the seed URL and should be archived.

## REFERENCES

- Ainsworth, S. G., & Nelson, M. L. (2015). Evaluating sliding and sticky target policies by measuring temporal drift in acyclic walks through a web archive. *International Journal on Digital Libraries*, 16(2), 129–144. doi: 10.1007/s00799-014-0120-4
- Ainsworth, S. G., Nelson, M. L., & Van de Sompel, H. (2014). A framework for evaluation of composite memento temporal coherence. *Computing Research Respository (CoRR)*, abs/1402.0928. Retrieved from <http://arxiv.org/abs/1402.0928>
- AlNoamany, Y., Weigle, M. C., & Nelson, M. L. (2015). Detecting off-topic pages in web archives. In S. Kapidakis, C. Mazurek, & M. Werla (Eds.), *Research and Advanced Technology for Digital Libraries: Lecture Notes in Computer Science* (Vol. 9316, pp. 225–237). Cham, Switzerland: Springer International Publishing. Retrieved from [http://dx.doi.org/10.1007/978-3-319-24592-8\\_17](http://dx.doi.org/10.1007/978-3-319-24592-8_17) doi: 10.1007/978-3-319-24592-8\_17
- Archive-It. (2014). *Learn more*. Retrieved from <https://archive-it.org/learn-more>
- Archives Unleashed Toolkit [Computer software]. (2018). The Archives Unleashed Project.
- Bailey, J., Grotke, A., McCain, E., Moffatt, C., & Taylor, N. (2016). *Web archiving in the united states: A 2016 survey* (Research Report). Retrieved from <http://ndsa.org/publications/>
- Banos, V. (2012). *Archiveready.com: Website archivability evaluation tool*. Retrieved from <http://archiveready.com/>
- Banos, V., Kim, Y., Ross, S., & Manolopoulos, Y. (2013, September). *CLEAR: A credible method to evaluate website archivability*. Presented at the 10th International Conference on Preservation of Digital Objects (iPRES 2013), Lisbon, Portugal. Retrieved from [http://www.academia.edu/10967309/CLEAR\\_a\\_credible\\_method\\_to\\_evaluate\\_website\\_archivability](http://www.academia.edu/10967309/CLEAR_a_credible_method_to_evaluate_website_archivability)
- Banos, V., & Manolopoulos, Y. (2015). A quantitative approach to evaluate website archiv-

- ability using the CLEAR+ method. *International Journal on Digital Libraries*, 1-23. doi: 10.1007/s00799-015-0144-4
- Batini, C., Palmonari, M., & Viscusi, G. (2012, July). The many faces of information and their impact on information quality. In P. Illari & L. Floridi (Eds.), (pp. 5–23). Birmingham, UK. Symposium conducted at the AISB/IACAP World Congress 2012.
- Batini, C., & Scannapieco, M. (2016). *Data and information quality: Dimensions, principles and techniques (data-centric systems and applications)* [Kindle book]. Cham, Switzerland: Springer International Publishing.
- Bonato, A. (2005). A survey of models of the web graph. In *Proceedings of the First International Conference on Combinatorial and Algorithmic Aspects of Networking* (pp. 159–172). Berlin, Heidelberg: Springer-Verlag. doi: 10.1007/11527954\_16
- Bruce, T. R., & Hillman, D. I. (2004). The continuum of metadata quality: Defining, expressing, exploiting. In D. I. Hillmann & E. L. Westbrook (Eds.), *Metadata in practice* (pp. 238–256). Chicago, IL: American Library Association.
- Brügger, N. (2009, February). Website history and the website as an object of study. *New Media & Society*, 11(1-2), 115–132. doi: 10.1177/1461444808099574
- Brunelle, J., Kelly, M., SalahEldeen, H., Weigle, M. C., & Nelson, M. L. (2015). Not all mementos are created equal: measuring the impact of missing resources. *International Journal on Digital Libraries*, 1-19. doi: 10.1007/s00799-015-0150-6
- Brunelle, J., Kelly, M., Weigle, M., & Nelson, M. L. (2015). The impact of JavaScript on archivability. *International Journal on Digital Libraries*, 1-23. doi: 10.1007/s00799-015-0140-8
- Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis* (1st ed.). London, UK: SAGE Publications Ltd.
- Cook, M. (1999). *Management of information from archives* (2nd ed.). Aldershot, UK: Gower Publishing Company.

- Denev, D., Mazeika, A., Spaniol, M., & Weikum, G. (2011, March). The SHARC framework for data quality in web archiving. *The VLDB Journal*, 20(2), 183–207. doi: 10.1007/s00778-011-0219-9
- End user. (2013). In D. Downing (Ed.), *Barron's business guides: Dictionary of computer and internet terms (11th ed.)*. Hauppauge, NY: Barron's Educational Series.
- Erciyas, K. (2014). *Complex networks : An algorithmic perspective*. Boca Raton, FL: CRC Press.
- Evans, G. L. (2013). A novice researcher's first walk through the maze of grounded theory: Rationalization for classical grounded theory. *Grounded Theory Review*(3). Retrieved from <http://groundedtheoryreview.com/2013/06/22/a-novice-researchers-first-walk-through-the-maze-of-grounded-theory-rationalization-for-classical-grounded-theory/>
- Firefox Quantum (Version 57.0.4) [Computer software]. (2017). Mountain View, CA: Mozilla.
- Floridi, L. (2013). Information quality. *Philosophy & Technology*, 26(1), 1-6. doi: 10.1007/s13347-013-0101-3
- Foot, K., & Schneider, S. (2010). Object-oriented web historiography. In N. Brügger (Ed.), *Web history* (pp. 61–79). New York: Peter Lang.
- Glaser, B. (1978). *Theoretical sensitivity: Advances in the methodology of grounded theory*. Mill Valley, CA: The Sociology Press.
- Glaser, B., & Strauss, A. (2009). *The discovery of grounded theory: Strategies for qualitative research* [Kindle book]. Aldine Transaction. Retrieved from <http://amazon.com/o/ASIN/0202302601/> (Original work published 1967)
- Goshtasby, A. A. (2012). *Image registration: Principles, tools and methods*. Berlin, Germany: Springer Science and Business Media.
- Grbich, C. (2012). *Qualitative data analysis: An introduction (2nd ed.)*. London: SAGE

Publications Ltd.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2002). *Applied statistics for the behavioral sciences* (5th ed.). Boston, MA: Cengage Learning.

International Internet Preservation Consortium. (n.d.). *About IIPC*. Retrieved from <http://netpreserve.org/about-us>

International Internet Preservation Consortium. (2016). *Member archives*. Retrieved from <http://netpreserve.org/resources/member-archives>

Internet Archive. (2012). *About IA*. Retrieved from <http://archive.org/about/>

ISO Technical Committee ISO/TC 46. (2009). *Information and documentation - WARC file format* (Tech. Rep. No. ISO 28500:2009). Geneva, Switzerland: International Organization for Standardization. Retrieved from [http://www.iso.org/iso/iso\\_catalogue/catalogue\\_tc/catalogue\\_detail.htm?csnumber=44717](http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=44717)

Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Krathwohl, D. R. (2009). *Methods of educational and social science research: The logic of methods* (3rd ed.). Long Grove, IL: Waveland Press, Inc.

Kuny, T. (1997, September). *A digital dark ages? Challenges in the preservation of electronic information*. Presented at the 63rd International Federation of Library Associations and Institutions (IFLA) Council and General Conference, Copenhagen, Denmark. Retrieved from <http://archive.ifla.org/IV/ifla63/63kuny1.pdf>

Lazarsfeld, P. F. (1959). Problems in methodology. In R. Merton, L. Broom, & L. Cottrell (Eds.), *Sociology Today* (pp. 39–78). New York: Basic Books.

Masanès, J. (2006). *Web archiving*. Berlin, Germany: Springer.

Mohr, G., Kimpton, M., Stack, M., & Ranitovic, I. (2004). Introduction to heritrix, an archival quality web crawler. In *In proceedings of the 4th International Web Archiving Workshop (IWAW'04)*. Bath, UK.

- PhantomJS (Version 2.1.1) [Computer software]. (2016). Hidayat, Ariya.
- Pinter, C. C. (2014). *A book of set theory*. Mineola, NY: Dover Publications.
- QSR International. (2016). *Nvivo product range*. Retrieved from <http://www.qsrinternational.com/nvivo-product>
- Reyes Ayala, B. (2013). *Web archiving bibliography 2013* (Research Report). Retrieved from <http://digital.library.unt.edu/ark:/67531/metadc172362/>
- Reyes Ayala, B. (2017, June). Web archives: A preliminary exploration of user expectations vs. reality. In E. Fox (Chair), *Web archiving and digital libraries*. Workshop conducted at the meeting of the Joint Conference of Digital Libraries (JCDDL), Toronto, Canada.
- Reyes Ayala, B., Phillips, M. E., & Ko, L. (2014). *Current quality assurance practices in web archiving* (Research Report). Retrieved from <http://digital.library.unt.edu/ark:/67531/metadc333026/>
- Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the web. *Journal of the American Society for Information Science and Technology*, 53(2), 145–161. doi: 10.1002/asi.10017
- Ruge, G. (1992, January). Experiment on linguistically-based term associations. *Information Processing & Management*, 28(3), 317–332. doi: 10.1016/0306-4573(92)90078-E
- Salton, G., Wong, A., & Yang, C. S. (1975, November). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613–620. doi: 10.1145/361219.361220
- Santos, R. (2016, July 16). *How do I compare two images to see if they are equal?* San Francisco, CA: Internet Archive. Retrieved from <http://web.archive.org/web/20160716062247/http://www.lac.inpe.br/JIPCookbook/6050-howto-compareimages.jsp>
- Sarkar, D. (2016). *Text analytics with python: A practical real-world approach to gaining actionable insights from your data*. Berkeley, CA: Apress.



- Schellenberg, T. R. (1975). *Modern archives principles and techniques (midway reprints)* (1st ed.). Chicago, IL: The University of Chicago Press.
- Screaming Frog SEO Spider (Version 9.2) [Computer software]. (2018). Portsmouth, UK: Screaming Frog Limited.
- Selenium (Version 3.81) [Computer software]. (2017). Selenium Project.
- Spaniol, M., Mazeika, A., Denev, D., & Weikum, G. (2009). "Catch me if you can": Visual analysis of coherence defects in web archiving. In *Proceedings of the 9th International Web Archiving Workshop (IWAW), Corfu, Greece, September 30 - October 1, 2009* (pp. 27 – 37).
- Stvilia, B. (2006). *Measuring information quality* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database: (Order No. 3223727).
- Swiss National Library. (2015). *Visual quality indicator (VQI)* (Tech. Rep.). Bern, Switzerland.
- Taguchi, G., Elsayed, E. A., & Hsiang, T. C. (1988). *Quality engineering in production systems (mcgraw hill series in industrial engineering and management science)*. New York, NY: McGraw-Hill College.
- Taylor, R. S. (1986). *Value-added processes in information systems*. Norwood, NJ: Ablex Publishing Corporation.
- United National Educational Scientific and Cultural Organization. (2003). *Charter on the preservation of the digital heritage*. Retrieved from [http://portal.unesco.org/en/ev.php-URL\\_ID=17721&URL\\_DO=DO\\_TOPIC&URL\\_SECTION=201.html](http://portal.unesco.org/en/ev.php-URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION=201.html)
- University of North Texas. (2004, July 22). *Campus map: University of North Texas*. San Francisco, CA: Internet Archive. Retrieved from <http://web.archive.org/web/20040722064240/http://www.unt.edu/pais/map/campusmap.htm>
- University of North Texas. (2007a, July 16). *Admissions: University of North Texas*. San Francisco, CA: Internet Archive. Retrieved from <http://web.archive.org/web/>

[20070716164959/https://www.unt.edu/admissions.htm](https://www.unt.edu/admissions.htm)

University of North Texas. (2007b, July 16). *Athletics: University of North Texas*. San Francisco, CA: Internet Archive. Retrieved from <http://web.archive.org/web/20070716164831/https://www.unt.edu/athletics.htm>

User. (2017). In *Oxford english dictionary online*. New York, NY: Oxford University Press. Retrieved from <http://www.oed.com/view/Entry/220650?rskey=514SYH&result=1&isAdvanced=false>

Voorburg, R. (2010, September). *Improving quality assurance for selective harvesting*. Presented at the International Internet Preservation Consortium Working Group Meeting, Vienna, Austria.

Wikipedia. (n.d.). *List of web archiving initiatives*. Retrieved from [http://en.wikipedia.org/wiki/List\\_of\\_Web\\_archiving\\_initiatives](http://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives)

Zhu, X., & Gauch, S. (2000). Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 288–295). New York, NY: ACM. doi: 10.1145/345508.345602