A Grounded Theory of Information Quality in Web Archives

[Extended Abstract] *

Brenda Reyes Ayala
University of North Texas
Department of Information Science
Denton, Texas USA
Brenda.Reyes@unt.edu

Keywords

web archiving; information quality; grounded theory, panel analysis

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 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.

In the field of web archiving, there has been only one comprehensive definition of Information Quality (IQ) in a web archive, put forward by Masanés [13]. 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

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 model 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. Almost 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.

The goal of this research is to build a model 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 definition of information quality (IQ) for web archives?

RQ 2 Which aspects of IQ can be most successfully measured?

The study presented in this dissertation will research the topic of information quality in web archives using the Grounded Theory (GT) approach.

2. LITERATURE REVIEW

2.1 Frameworks, Theories, and Models of Information Quality

The field of Information Science has produced several models of IQ. Taylor [19] incorporated IQ as part of his value-added model. The value-added model described 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*. They are called this because they enhance or add value to the information being presented by the interface. 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. As Taylor defines it, quality is "a user criterion which has to do with excellence or in some cases truthfulness in labeling". Quality has the values of accuracy (error-free), comprehensiveness (completeness of coverage), currency (recency of the data), validity (information can be judged to be sound), and reliability (trust the user has in the system's reliability).

In 2002, Soo Young Rieh [17] 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.

In their paper, Zhu and Gauch [20] 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 of currency, availability, information-to-noise ration, authority, popularity and cohesiveness.

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

$$G_i = \overline{W}_i * (a''_s * \overline{T}_i + b''_s * \overline{A}_i + c''_s * \overline{I}_i + d''_s * \overline{R}_i + e''_s * \overline{P}_i + f''_s * \overline{C}_i) \quad (1)$$

where $\overline{W}_i, \overline{T}_i, \overline{A}_i, \overline{I}_i, \overline{R}_i, \overline{P}_i$ are the means of information quantity, currency, availability, information-to-noise ratio, authority, and popularity of site i across topics relevant to the query, $\overline{C}_i$, is the cohesiveness of site i, and a''_s, b''_s, c''_s, d''_s, e''_s, and f''_s are the weights representing the importance of each quality metric.

2.2 Quality Assurance in Web Archiving

The process of archiving a website usually occurs in the following order, as noted in Reyes Ayala (2015) [15]:

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.

In their survey of web archiving practices, Reyes Ayala, Phillips, and Ko (2014) [16] 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.
- Quality problems are noted, either in a spreadsheet or in another system such as a database.

The authors noted that the process of QA for web archives is an onerous one because it involves manually inspecting hundred if not thousands of archived websites. This necessitates a significant time commitment from web archivists, and a specialized knowledge and skills.

2.3 Research on Quality in Web Archives

In the field of web archiving, researchers have recently begun to address the topic of quality for web archives. Though they do not offer a comprehensive definition of quality, as Masanés [13] does, some *do* examine specific facets of quality such as coherence and completeness. Some of the researchers have also attempted to operationalize individual aspects of quality and to create metrics to effectively measure it.

2.3.1 The Notion of Coherence in a Web Archive

In their paper, "'Catch me if you can': Visual analysis of coherence defects in web archiving" Spaniol, Mazeika, Denev, and Weikum (2009) [18] 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. Coherence defects 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 [10] introduced the 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.

The work of Ainsworth, Nelson, and Van de Sompel [1] 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.

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" [1]. 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:

1. Prima Facie Coherent: The embedded memento existed in its archived state at the time the root memento was captured.

2. Prima Facie Violative: The embedded memento did not exist in its archived state at the time the root memento was captured.

3. Possibly Coherent: The embedded memento could have existed in its archive state at the time the root memento was captured.

4. Probably Violative: The embedded memento probably did not exist in its archived state at the time the root memento was captured.

5. Coherence Undefined: There is not enough information to determine coherence state.

The authors 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-Datetime), but also the content of the two mementos in order to determine coherence.

Ainsworth and Nelson [2] were also concerned with defining quality as meeting measurable characteristics. Their work elaborates on the notion of coherence put forward by [10]. 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 [10] 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. 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 resemblance to the original. [2] 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."[2].

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, 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. [7] 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.

The authors proposed a metric to assess this damage that is based on three factors: the MIME type, size, and location of the embedded resource [7]. 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. The authors focus on three types of embedded resources: images, multimedia elements, and stylesheets, and assign calculate their importance in terms of the possible damage caused to the archived website if these were missing. 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, or $D_m = \frac{D_{m_{actual}}}{D_{m_{potential}}}$. 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."[7].

The potential damage $D_{m_{potential}}$ is the sum of the importance of each embedded resource, as shown in Equation 2. 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.

$$D_{m_{potential}} = \frac{\sum_{i=1}^{n_{[I,MM]}} D_{[I|MM]}(i)}{n_{[I|MM]}} + \frac{\sum_{i=1}^{n_C} D_c(i)}{n_C} \quad \forall\{I = Images, MM = Multimedia, C = CSS\} \quad n \in R \tag{2}$$

[3] also addressed quality problems that could affect the coherence of a web archive, such as off-topic web pages. 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. According to their results, the cosine similarity method proved the best at detecting off-topic web pages. 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.

2.3.2 The Notion of Archivability

In their iPres paper "CLEAR: A Credible Method to Evaluate Website Archivability", Banos, Kim, Ross, and Manolopou-

los (2013) [5] introduced the concept of website archivability. Archivability was defined as the "sum of the attributes that make a website amenable to being archived" [5]. 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 usage. 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 four facets: accessibility, standards compliance, cohesion, and metadata usage. 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.

[6] 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. Once the value for each facet has been calculated, the total archivability score for the website can also be calculated using $WA = \sum_{\lambda \in \{A,S,C,M\}} w_\lambda F_\lambda$. In this formula, 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, while w represents each facet's significance, or weight.

Other researchers have also focused on the notion of archivability and attempted to operationalize it. In their paper "The impact of JavaScript on archivability", [8] defined archivability as the ease with which a website can be archived, which is similar to the concept put forward by [6]. 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[8].

However, obtaining a perfect copy of the original is an onerous process, made more difficult by the widespread use of the JavaScript programming language. 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 [8]. To study the impact of JavaScript on archivability, the researchers studied the quality of the archived URLs and their use of the JavaScript language, and presented several metrics to measure their archivability.

[8] also presented the *content complexity* metric for a URL, which is measured as the number of `<script>` tags present inside its HMTL markup, or $CC = \sum script\ tags \in HTML$. They found that over half of the URLs in their collection used JavaScript to load embedded resources. Similarly, JavaScript was responsible for 52.7% of all missing embedded resources during the same time period. 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.

3. METHODOLOGY

Previous research on the notion of quality in web archives has been mostly system-centered, that is, focused on the process and technologies of web archiving. Instead, my research is user-centered; it seeks to analyze the behavior of users and creators of web archives, and how they envision the concept of information quality.

3.1 Phase 1: Building a Substantive Theory of Quality in a Web Archive

The Internet Archive's Archive-It (AIT) 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 400 clients (called "partners") consisting of universities, state libraries and archives, museums, and national libraries in several countries [4]. AIT clients comprise a wide range of job roles, education, and experience. They range from dedicated web archivists who work full-time on preserving the web for large organizations, to employees of small-corporations who do web archiving part-time, from those well-versed in the complexities of Internet technologies to those just beginning to grasp the process of web archiving. The sheer size and diversity of the AIT user base makes it an optimal subject of study.

During this initial phase, I have collected and analyzed support tickets that have been submitted to the Internet Archive's Archive-It service. The tickets collected were Level 1 support tickets that had been submitted by AIT clients over the course of several years, and include the initial question submitted by the client, the response given by the AIT partner specialist, and any subsequent communication between the two. To this end, I negotiated a researcher agreement with the Internet Archive. Among other conditions, the research agreement stipulated that the researcher anonymize any personal or institutional information present in the tickets, as well as any other potentially identifying information.

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. They were in the form of a large file in XML format. 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 written by myself After the tickets were cleaned, they were imported into the NVivo software package, a popular software for qualitative data analysis [14]. My research focuses on tickets in which the client discusses a perceived flaw in an individual archived website or an entire web archive. 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."
- "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 contraints yields 20,300 +- documents."

The AIT tickets were analyzed using the Grounded Theory (GT) methodology. GT, created by Barney Glaser and Anselm Strauss. GT is defined as "the discovery of theory from data - systematically obtained and analyzed in social research" [11]. GT is an inductive methodology created explicitly for generating theory in the Social Sciences: working closely from the data, the researcher begins the work of generating a theory. 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" [9]. Coding allows the researcher to discover what is happening in the data and to grapple with what it means. GT also uses 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". A category is a conceptual element of the theory, while a property, is a conceptual aspect or element of a category. The categories with the most explanatory power are called core categories.

I used these techniques to identify the main concepts and categories present in the data. I am using the following questions to guide me in my 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?

In GT conventions, the literature review takes place after the main categories have emerged. Accordingly, while coding the data I am also reviewing literature covering IQ and web archiving.

3.2 Phase 2: Identifying the Operationalizable Dimensions of Web Archive Quality

Phase 2 of this study involves operationalizing the dimensions of quality present in the model developed during Phase 1. That is, after I have generated a multidimensional model of IQ, the different dimensions can then be operationalized into mathematical definitions. These definitions can then be used to quantitatively measure the IQ of a web archive. To be in line with the goals of this research, the definitions should be as technology-independent as possible and suitable for use in a wide variety of contexts and platforms.

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. 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 will also be using GT, but will employ Paul Lazarsfeld's panel analysis method to explore the relationships between aspects of IQ and express them mathematically [12]. Though I have yet to begin Phase 2 officially, I have done some exploratory analysis that attempts to operationalize completeness, as can be seen in Section 4.2.

4. PRELIMINARY RESULTS

4.1 Core Facets in the IQ of Web Archives

At the time of writing, I have classified 128 AIT tickets. The following categories have emerged as facets of quality:

- Archivability: the intrinsic properties of a website that make it easier or more difficult to archive. There are a number of factors that could affect the archivability of a website, but two have been noted more in the data: (1) the complexity of the site, which can be operationalized as the total number of its components and (2) dynamism, or the number of dynamic components, such as JavaScript and videos contained in the website.
- Completeness: 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 website's content has not been captured or is not present in the archive.
- Relevance: Creators and curators of archives seem to have a mental model of what is "relevant" or "irrelevant" content in their web archives. They use these concepts to delimit the boundaries of a web archive: what is inside is relevant, anything outside is irrelevant. They have a few ways of determining what is irrelevant content, and the most common types are: sites that were not explicitly on the seed list of the collection (boundary relevance), sites with inappropriate content (topic relevance), and sites in quantity or volume that is unexpected or excessive (size relevance).
- Correspondence: The user expects the archived website to provide the same interaction and user experience as the original, an explicit or implicit comparison

is drawn between the archived website and the original. A problem occurs when the user's interaction with the site is different from that of the original, unexpected, or deficient. Users have a strong idea of what the archived website should look or behave like.

These are only very rough, emergent categories. More time and effort is needed to analyze more tickets and further refine the categories.

4.2 A Mathematical Exploration of Completeness in a Web Archive

Let us define two variables x, the original website and x', the archived website. We can operationalize completeness as the cosine similarity between x, the original website and x', the archived website, as seen in Equation 3.

$$cosine(x, x') = \frac{x \cdot x'}{\parallel x \parallel \parallel x' \parallel} \tag{3}$$

Cosine similarity was chosen as a measure of completeness because of its prior use in [3] 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* [12]. 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". 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 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 cosine similarity becomes:

$$cosine(X, X') = \frac{X \cdot X'}{\parallel X \parallel \parallel X' \parallel} = \frac{\sum_{i=1}^{n} c_i * c'_i}{\sqrt{\sum_{i=1}^{n} {c_i}^2} * \sqrt{\sum_{i=1}^{n} {c'_i}^2}} \tag{4}$$

$$X =< c_1, c_2, c_3, c_4, c_5, , ...c_n >$$
$$X' =< c'_1, c'_2, c'_3, c'_4, c'_5, , ...c'_n >$$

- 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, X, always has all of its components, so $X =< 1, 1, 1, 1, 1, ...1 >$

- The archived website, $X' =< c'_1, c'_2, c'_3, c'_4, c'_5, , ...c'_n >$, since we do not yet know the values of X'.

Then we can proceed to calculate the magnitudes of the original site and the archived site:

$$\parallel X \parallel = \sqrt{\sum_{i=1}^{n} {c_i}^2} = \sqrt{\sum_{i=1}^{n} 1^2} = \sqrt{n}$$

$$\parallel X' \parallel = \sqrt{\sum_{i=1}^{n} {c'_i}^2} = \sqrt{{c'_1}^2 + {c'_2}^2 ... + {c'_n}^2}$$

As well as their dot product:

$$X \cdot X' =< 1, 1, ..1 > \cdot < c'_1, c'_2, ..c'_n = \sum_{i=1}^{n} (1) c'_i = \sum_{i=1}^{n} c'_i$$

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

$$cosine(X, X') = \frac{\sum_{i=1}^{n} c'_i}{\sqrt{n * \sum_{i=1}^{n} {c'_i}^2}} \tag{5}$$

5. EXPECTED CONTRIBUTIONS TO THE RESEARCH AREA

By clarifying the notion of quality for web archives, the resulting model will begin the work of establishing a much-needed theoretical groundwork for the field, which will help its development and growth. My model will help to clarify the concept of information quality as it applies to web archives, describe its facets, and explain it with the fewest possible concepts, and with the greatest possible scope.

Practitioners in the field of web archiving will also benefit from this model because they will be able to apply it to measure the quality of their own web archives. Because the model will be technology-independent, it will be applicable across a wide variety of different web archiving technologies, platforms, and systems. Knowing which aspects of web archive quality can be measured and how to measure them will allow web archiving professionals to improve the Quality Assurance processes for their organizations.

6. ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Jiangping Chen and the members of my committee, Dr. Oksana Zavalina, Dr. Cornelia Caragea, Dr. Shawne Miksa, and Dr. Kathryn Masten-Cain. I would also like to thank Lori Donovan and Jefferson Bailey of the Internet Archive.

7. REFERENCES

[1] S. Ainsworth, M. L. Nelson, and H. Van de Sompel. A framework for evaluation of composite memento temporal coherence. *Computing Research Respository (CoRR)*, abs/1402.0928, 2014.

[2] S. G. Ainsworth and M. L. Nelson. 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, 2015.

[3] Y. AlNoamany, M. C. Weigle, and M. L. Nelson. Detecting off-topic pages in web archives. In S. Kapidakis, C. Mazurek, and M. Werla, editors, *Research and Advanced Technology for Digital Libraries: Lecture Notes in Computer Science*, volume 9316, pages 225–237, Cham, Switzerland, 2015. Springer International Publishing.

[4] Archive-It. Learn more. https://archive-it.org/learn-more, 2014.

[5] V. Banos, Y. Kim, S. Ross, and Y. Manolopoulos. CLEAR: A credible method to evaluate website archivability. Presented at the 10th International Conference on Preservation of Digital Objects (iPRES 2013), Sept. 2013.

[6] V. Banos and Y. Manolopoulos. A quantitative approach to evaluate website archivability using the CLEAR+ method. *International Journal on Digital Libraries*, pages 1–23, 2015.

[7] J. F. Brunelle, M. Kelly, H. SalahEldeen, M. C. Weigle, and M. L. Nelson. Not all mementos are created equal: measuring the impact of missing resources. *International Journal on Digital Libraries*, pages 1–19, 2015.

[8] J. F. Brunelle, M. Kelly, M. C. Weigle, and M. L. Nelson. The impact of Javascript on archivability. *International Journal on Digital Libraries*, pages 1–23, 2015.

[9] K. Charmaz. *Constructing grounded theory: A practical guide through qualitative analysis*. SAGE Publications Ltd, London,England, 1 edition, 1 2006.

[10] D. Denev, A. Mazeika, M. Spaniol, and G. Weikum. The SHARC framework for data quality in web archiving. *The VLDB Journal*, 20(2):183–207, Mar. 2011.

[11] B. Glaser and A. Strauss. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine Transaction, 2009.

[12] P. F. Lazarsfeld. Problems in methodology. In R. Merton, L. Broom, and L. Cottrell, editors, *Sociology Today*, pages 39–78. Basic Books, New York, 1959.

[13] J. Masanès. *Web archiving*. Springer, Berlin; New York, 2006.

[14] QSR International. Nvivo product range. http://www.qsrinternational.com/nvivo-product, 2016.

[15] B. Reyes Ayala. Web archiving bibliography 2013. Research report, University of North Texas, 2013.

[16] B. Reyes Ayala, M. E. Phillips, and L. Ko. Current quality assurance practices in web archiving. Research report, University of North Texas, 2014.

[17] S. Y. Rieh. Judgment of information quality and cognitive authority in the web. *Journal of the American Society for Information Science and Technology*, 53(2):145–161, 2002.

[18] M. Spaniol, A. Mazeika, D. Denev, and G. Weikum. "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*, pages 27 – 37, 2009.

[19] R. S. Taylor. *Value-added Processes in Information Systems*. Ablex Publishing Corporation, Norwood, NJ, 1986.

[20] X. Zhu and S. Gauch. 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*, SIGIR '00, pages 288–295, New York, NY, USA, 2000. ACM.