"Im not sure how feasible capture is": archivability as a dimension of website quality

Dr. Brenda Reyes Ayala[1],

[1] Associate Professor
School of Library and Information Studies, University of Alberta
Edmonton, Alberta, Canada
brenda dot reyes at ualberta dot ca

February 21, 2025

21st Conference on Information and Research Science Connecting to Digital and Library Science (IRCDL 2025), Udine, Italy

Overview

1. **Introduction**
2. **Related Work**
3. **Methods**
4. **Results and Discussion**
5. **Discussion**
6. **Conclusion**
7. **References**

The context of web archives

- Web archiving is the practice of preserving web content.
- Carried out by institutions such as libraries, governments, and universities to preserve digital cultural heritage.
- Day-to-day tasks for archiving the web:
  1. Appraisal and Selection: institutions decide specifically which websites they want to collect.
  2. Scoping: institutions may opt to archive portions of a website, whole sites, or even entire web domains.
  3. Data Capture: institutions fine-tune how they want to capture their data through decisions about crawl (capture) frequency and types of files to archive or not archive.
  4. Storage and Organization: This step includes a temporary or long-term storage plan for the archived data.
  5. **Quality Assurance and Analysis**: institutions review what they have archived and how well the resulting collection satisfies the goals they set at the beginning of the life cycle.

The context of web archives, cont.

- Most popular web archiving service is the Internet Archive's Archive-It (AIT), which helps organizations build and manage their own web archives
- It currently has over 800 clients *partners* consisting of universities, archives, museums, and libraries in over 24 countries (Archive-It, 2021)
- In 2022, the National Digital Stewardship Alliance (NDSA) conducted another survey of web archiving practices worldwide (Abrams et al., 2023). The authors found that, of the over 190 institutions that had web archiving programs in place, 71% capture content with AIT
- Web archiving is a field with few conceptual tools or theoretical definitions

A theory of IQ for web archives I

In a previous work I presented a grounded theory of Information Quality (IQ) for web archives, derived from an analysis of tickets submitted to the Internet Archive's AIT service by web archivists (Reyes Ayala, 2020)

1. Correspondence: degree of similarity, or resemblance, between the original website and the archived website
   - Visual correspondence: similarity in appearance between the original website and the archived website
   - Interactional correspondence: the degree to which a user's interaction with the archived website is similar to that of the original
   - Completeness: the degree to which the archived website contains all of the components of the original
2. Relevance: pertinence of the contents of an archived website to the original website

A theory of IQ for web archives II

  - 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
  - Size relevance: the similarity in size of the archived website to the original website

3. **Archivability**: degree to which the intrinsic properties of a website make it easier or more difficult to archive

Research Questions

1. How do people perceive the notion of archivability in web archives?
2. How does website archivability affect web archives and thus the future historical record?

Evaluating website archivability

- In “CLEAR: A Credible Method to Evaluate Website Archivability”, (Banos, Kim, Ross, & 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).
- Introduced facets designed to determine the archivability of a website, the Credible Live Evaluation of Archive Readiness, or CLEAR, method.
- These facets were: accesibility, standards compliance, cohesion, and metadata usage. (Banos & Manolopoulos, 2015):

Another definition of archivability

- In their paper “The impact of JavaScript on archivability”, (Brunelle, Kelly, Weigle, & Nelson, 2015) defined archivability as the ease with which a website can be archived
- 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 et al., 2015)
- 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 et al., 2015)
- (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”

Grounded theory

- Barney Glaser and Anselm Strauss created the methodology of Grounded Theory (GT), which they defined as “the discovery of theory from data - systematically obtained and analysed in social research’ (Glaser & Strauss, 1967/2009)
- Theory was not a perfected product, but a process, an ever-developing entity
- GT is an inductive methodology; working closely from the data, the researcher begins the work of generating a theory

Data Gathering and Processing

- Accounts of AIt clients are managed by a team of partner specialists
- When a client encounters a problem, she first opens a support ticket using Zendesk
- The ticket is received by a partner specialist, who is then responsible for addressing the issue
- AIT support tickets are a rich source of information regarding quality problems in web archives
- They contain rich descriptions of how quality problems are detected, analysed, and addressed, and are thus an ideal dataset for studying quality in all its dimensions

Process I

1. Negotiated a researcher agreement with the organisation to obtain support tickets from the years 2012 through 2016
2. Tickets collected were Level 1 support tickets that had been submitted by AIT client. They included the initial question submitted by the client, the response given by the AIT partner specialist, and any subsequent communication between the two
3. Final dataset was then imported into the NVivo software package, a popular program for performing qualitative data analysis (QSR International, 2016)
4. 305 tickets and 2544 interactions were analyzed using the GT techniques of open coding and theoretical memos to identify the main concepts and categories present in the data

Process II

5. In order to increase the quality and rigour of the study, I engaged in purposeful peer review. University professors were periodically invited to audit the entire research project, including the codebook, preliminary findings, and core categories
6. In addition to peers, employees of the Internet Archive were also invited to see the findings and comment on them

Archivability as a dimension

- Archivability proved to be a prominent dimension, as it appears 101 times in 78 tickets, behind correspondence and relevance
- Archivality problems occur because a website:
  1. has changed the way the content is delivered to the user.
  2. is media-heavy or contains much dynamic content.
  3. renders content in a unique, non-standard way.

Examples of archivability problems caused by websites changing how it delivers content to users

Ticket 129

C: I am getting an error on the following Facebook crawl.
AIT: Facebook made a change to the settings for their stylesheets

Ticket 258

C: in the "Township of ___" collection I am trying to capture this facebook site: http://www.facebook.com/pages/township
AIT: 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

Examples of archivability problems caused by websites with dynamic content I

Ticket 369

C: The athletics department has their game day programs online. I see to be able to view the sections but can't see a way to capture printer-friendly formats from their link. Is this possible?

AIT: 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

Examples of archivability problems caused by websites with dynamic content II

Ticket 2884

C: under the About Us tab, under Press Room, the tabs other than News Releases (___ in the news, Annual report, Media Kit, and Social Media) do not work:

AIT: 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 I

Ticket 30

C: Much of this content is located in databases, so, in general I'm curious about how Archive-It will handle these databases.
AIT: 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

Special case: Websites that are database and form or search-driven II

Ticket 3458

C: I would like to know if there is any way I can capture the search feature of the website

AIT: 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

Examples of archivability problems caused by websites rendering content in unique ways I

Ticket 3243

C: I've done a test crawl on all ".stateu.edu" while it has captured thousands of pages it also seems to determine many "stateu.edu" pages to be "out of scope". These pages are not blocked by a robots.txt. Why would that be happening?

AIT: 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

Examples of archivability problems caused by websites rendering content in unique ways II

Ticket 86

C: We ran another test crawl on this site and now we seemed to have opened a can of worms for the main site we wanted to crawl. It looks like we have the Flickr URL under control and may just put a limit on the number captured. The main URL that we want, captured 83, 600 with over 1 million in the que

AIT: The issue with your http://www.pl.gov/tef/ 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

How do people conceptualise the notion of archivability?

- Archivability is not a dimension of quality that is directly perceived by most AIT clients.
- AIT employees were much more likely to perceive a quality problem as an archivability problem.
- Instead, AIT clients framed archivability issues as correspondence or relevance problems that negatively affected the quality of an archived website
- Website archivability can thus be seen as a *latent* dimension, because it is hidden from most people, and framed in terms of other quality problems
- The archivability of a website can only be perceived *after* its archived counterpart exhibits a quality problem.

How does website archivability affect web archives and thus the future historical record?

- A website with low archivability can negatively impact the quality of its archived version by causing correspondence or relevance problems
- Low archivability in a website leads to low-quality web archives, and low-quality web archives lead to low-quality historical records
- Any substantial change in web technologies, standards, or platforms necessitates a change in web archiving practice in order to catch up and create high-quality web archives that result in a high-quality historical record
- The web is no longer archivable, and will become increasingly so as time passes.
- Current web archiving technologies cannot adequately capture the web as it is now, yielding, at worst, highly-degraded versions of the original

Conclusion

Web archives and history

“web archives do not provide a perfect representation of the past...but neither do traditional archives, which have had to be very selective with what they select, appraise, and preserve”(Milligan, 2018)

The future...

Perhaps it is not web archivists, but historians, who will be the most able to deal with the errors and omissions of the past web

References I

Abrams, S., Collier, Z., Colón-Marrero, E., keondra bills freemyn, Krabbenhoeft, N., Wertheimer, M. E., & Wickner, A. (2023, oct). *2022 web archiving survey results* [Research Report]. Retrieved from http://ndsa.org/publications/

Archive-It. (2021). *Learn more.* https://archive-it.org/learn-more.

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

References II

Banos, V., & Manolopoulos, Y. (2015). A quantitative approach to evaluate website archivability using the CLEAR+ method. *International Journal on Digital Libraries*, 1-23. doi: 10.1007/s00799-015-0144-4

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

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)

Milligan, I. (2018). Historiography and the web. In N. Brügger & I. Milligan (Eds.), *The sage handbook of web history* (pp. 16–29). Los Angeles, CA, USA: Sage Publications Ltd.

References III

QSR International. (2016). *Nvivo product range.* Retrieved from http://www.qsrinternational.com/nvivo-product

Reyes Ayala, B. (2020). Correspondence as the primary measure of quality for web archives: A grounded theory study. In M. Hall, T. Merčun, T. Risse, & F. Duchateau (Eds.), *Digital libraries for open knowledge* (pp. 73–86). Cham: Springer International Publishing.