Correspondence as the Primary Measure of Quality for Web Archives: A Grounded Theory Study

Brenda Reyes Ayala[1]

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

August 27, 2020

24th International Conference on Theory and Practice of Digital Libraries (TPDL 2020)
Lyon, France

Overview I

1. Introduction

2. Contributions to the Research Area

3. Methodology, Data Collection, Analysis

- Methodology
- Data Collection and Analysis

4. Results: A Grounded Theory of IQ for Web Archives

- A Theory of IQ for Web Archives
- Correspondence

5. Conclusion

Terms and Definitions

- Web Archiving: the action of storing websites to preserve them as a historical, informational, legal, or evidential record
- Web Archive: a system that contains such records

Timeline of Web Archiving

1996 Internet Archive is founded with the mission of creating a universally accessible digital library. National Library of Australia inaugurates the first-ever web archiving program by a national library

2000 Library of Congress began its Minerva Project (now the Library of Congress Web Archives)

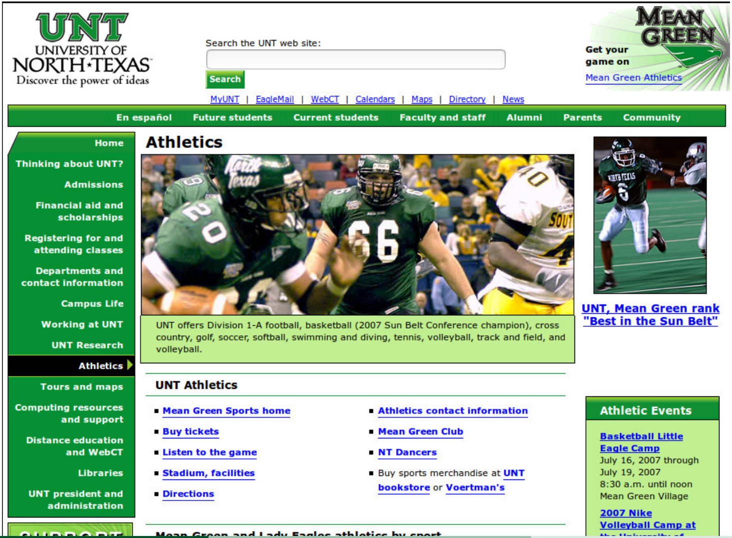
2003 International Internet Preservation Consortium (IIPC) is founded with the mission of "improving the tools, standards and best practices of web archiving while promoting international collaboration, broad access and use of web archives for research and cultural heritage"

2004 British Library launches the UK Web Archive

2006 National Library of France launches the French Web Archive

Good-Quality Archived Website

A good-quality archived version of the UNT Athletics site from 2007


UNT
UNIVERSITY OF NORTH TEXAS
Discover the power of ideas
Search the UNT web site:
Search
MyUNT | EagleMail | WebCT | Calendars | Maps | Directory | News
MEAN GREEN
Get your game on
Mean Green Athletics
En español
Future students
Current students
Faculty and staff
Alumni
Parents
Community
Home
Thinking about UNT?
Admissions
Financial aid and scholarships
Registering for and attending classes
Departments and contact information
Campus Life
Working at UNT
UNT Research
Athletics
Tours and maps
Computing resources and support
Distance education and WebCT
Libraries
UNT president and administration

Athletics

UNT offers Division 1-A football, basketball (2007 Sun Belt Conference champion), cross country, golf, soccer, softball, swimming and diving, tennis, volleyball, track and field, and volleyball.

UNT, Mean Green rank "Best in the Sun Belt"

UNT Athletics

- Mean Green Sports home
- Buy tickets
- Listen to the game
- Stadium, facilities
- Directions
- Athletics contact information
- Mean Green Club
- NT Dancers
- Buy sports merchandise at UNT bookstore or Voertman's

Athletic Events

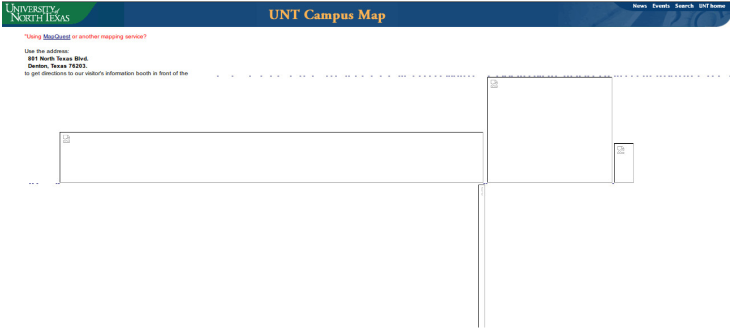
Basketball Little Eagle Camp
July 16, 2007 through July 19, 2007
8:30 a.m. until noon
Mean Green Village

2007 Nike Volleyball Camp at the University of

Mean Green and Lady Eagles athletics by sport

Low-Quality Archived Website

An archived version of the UNT Campus Map from 2004, almost unusable



UNT Campus Map

"Using MapQuest or another mapping service?

Use the address:
801 North Texas Blvd.
Denton, Texas 76203.
to get directions to our visitor's information booth in front of the

Purpose and Research Question

Purpose

To build a theory of IQ for web archives that is grounded in human-centred data

Research Question

What is the human-centred definition of quality for web archives?

Significance

This paper makes the following contributions:

1. The paper presents the first application of grounded theory to discipline of web archiving.
2. It introduces the first theory of quality developed specifically about web archives, and lays the groundwork for future theoretical and practical developments in the field.
3. The theory is human-centred and grounded in how subject-matter experts in the field of web archiving perceive quality.
4. The theory is *comprehensive*, incorporating and unifying the work of previous researchers on web archives.
5. The theory is independent of the technology currently in use to create web archives, making it suitable to a wide variety of platforms, preservation contexts, and situations.

Grounded Theory Methodology (GT)

- Introduced by Barney Glaser and Anselm Strauss in their 1967 book *The Discovery of Grounded Theory*
- The discovery of theory from data, systematically obtained and analyzed

Differences between GT and Logico-formal Theory

Characteristic	Traditional Approach	Grounded Theory
Literature Review	Before data collection	Throughout data collection, analysis
Method	Compare only “comparable" groups	Compare any groups
Sampling	Statistical sampling	Theoretical sampling
Data	Field notes, interviews, observations	Wide variety of materials
Data Collection	After theory is formulated	At any time
Purpose	To verify theory	To generate theory
Goal	To establish fact	To establish structural boundaries of fact
View theory as	A perfected product	An ever-developing entity

Building a 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.

1. Negotiated a researcher agreement with the Internet Archive to obtain a large cache of AIT support tickets
2. Cleaned the data and imported it into Nvivo software package
3. Used open coding and theoretical memos to identify the main concepts and categories present in the data
4. Created a preliminary theory of IQ for web archives
5. Used the constant comparison method to refine and improve the theory
6. Performed literature review

Sample AIT Support Ticket I

AIT client

In the collection, there seems to be a problem with the way the _____.com is being captured–when I try to access it through Wayback, all I get is a blank page (see attached). Does the way the _____.com site is built that makes it uncapturable? I did a test run before adding it to our collection, and don't remember this being a problem. Thanks!

Sample AIT Support Ticket II

AIT employee

Hi _____, It looks like most of the content from the _____.com site has been captured, but display is a bit tricky due to the extensive use of javascript and flash on this site. For now, you can view the archived content for this site by disabling javascript in your browser while you're browsing through the site in wayback.

It may not have the exact look and feel of this site on the live web, but you should be able to see most of the archived content. I will also send this along to our engineers to look into a more long-term solution to allow users to more easily view this content, and we'll update you when we have more information.

Please let me know if you have any further questions!

A Look at the Data Analyzed

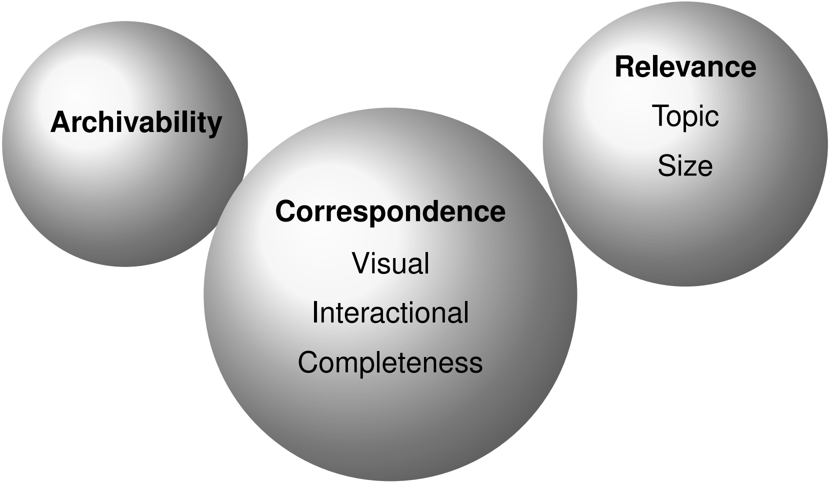
Number of Tickets and Interactions About Information Quality Analyzed Per Year

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

Core Facets of IQ for Web Archives

1. **Correspondence**: similarity between the original and archived websites. Good correspondence requires equivalence, or at least a close resemblance, between the two
   - Visual
   - Interactional
   - Completeness
2. **Relevance**: pertinence of the contents of an archived website to the original. Archived websites must not contain off-topic content (topic relevance) or content in quantity or volume that is unexpected or excessive (size relevance)
   - Topic
   - Size
3. **Archivability**: intrinsic properties of a website that make it more difficult to archive. A latent IQ dimension

The Theory of IQ for Web archives, Visualized



Archivability

Correspondence

Visual

Interactional

Completeness

Relevance

Topic

Size

Visual Correspondence

Similarity in appearance between the original website and the archived website

- When describing a quality problem, AIT clients will often compare the archived website to the original website
- AIT clients have a strong idea of what the archived website should look or behave like and are quick to report any discrepancies

Examples of Visual Correspondence

Example 1

“On the new http://www.stateu.edu/academics page we are not capturing the background images. I cannot figure out why since we are capturing other images from the same directory”

Example 2

“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”

Example 3

“We’re having some trouble with our Facebook site captures not displaying properly (or at all, really)”

Interactional Correspondence

Degree to which a user's interaction with the archived website is similar to that of the original

- 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
- Video content in web archives is notoriously difficult to replay
- Often (but not always), mismatched appearance and behaviours is caused by missing important files that provide needed visual elements or functionality. Completeness and interactional correspondence are separate, but often linked

Examples of Interactional Correspondence

Example 1

“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”

Example 2

“When i click on it, it briefly flashes to the homepage and then it displays a URL with the nationalscience URL in it twice”

Example 3

“Clicking View all comments under an update does not reveal the comments.”

Completeness

Degree to which the archived website contains all of the components of the original

- Occurs when the original website's content has not been captured or is not present in the archive
- Examples include missing search boxes, menus, comments, and entire web pages
- An archived website can have a lack of correspondence with the original website yet still be perfectly complete. However, the reverse is not true: an archived website cannot be incomplete, yet still have 100% correspondence with the original

Examples of Completeness

Example 1

“on all most every blog that we have captured from blogspot the Wayback Machine does not include the subsequent pages beyond the first”

Example 2

“We’re still having some trouble capturing the JavaScript menu at the top of the main page. I know that JS can be wonky.”

Example 3

“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”

Conclusions and Future Work

1. Theory presented here represents the majority of quality problems seen in topic-centred or event-driven web archives
2. If in the future, new technologies were developed to capture websites more successfully, the notions of visual correspondence, interactional correspondence, and completeness would still be relevant to quality in web archives
3. Having clear concepts based on experts perceive the issue of quality can lead to the successful creation of metrics, methods, and tools that will enable web archivists to measure the quality of their web archives

Thanks

Thank you for your time and support.