A grounded theory of information quality for web archives: Dimensions and applications

Dr. Brenda Reyes Ayala[1]

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

May 24, 2022

IIPC Web Archiving Conference (WAC) 2022
Q&A Session 7: Advancing Quality Assurance for Web Archives: Putting Theory into Practice

Overview I

1. Introduction

2. Methodology, Data Collection, Analysis

- Data Collection and Analysis

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

- A Theory of IQ for Web Archives
- Correspondence
- Relevance
- Archivability

4. Conclusion

References

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?

Grounded Theory Methodology (GT)

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

When should GT be applied?

Some situations where GT is applicable:

1. A field has a need for theoretical explanations and models. (Grbich, 2012)
2. A human-centred approach is desired. As its name implies, GT is heavily “grounded” in rich contextual data gathered from empirical research with people.

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

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 (Reyes Ayala, 2020)
   - Visual
   - Interactional
   - Completeness
2. **Relevance**: pertinence of the contents of an archived website to the original
   - Topical
   - Functional
3. **Archivability**: intrinsic properties of a website that make it more difficult to archive. A latent IQ dimension

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

Visual Correspondence in the Data

Example 1

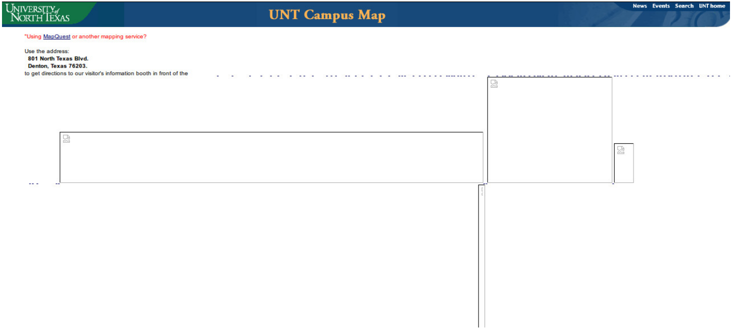
One thing related though, the page is not capturing its look and feel well... Any suggestions? It’s missing the background and objects are not in the right locations

Example 2

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

Visual Correspondence Problems in Real Life

An archived version of the UNT Campus Map from 2004, almost unusable



UNIVERSITY of NORTH TEXAS

UNT Campus Map

News Events Search UNT home

"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

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

Interactional Correspondence in the Data

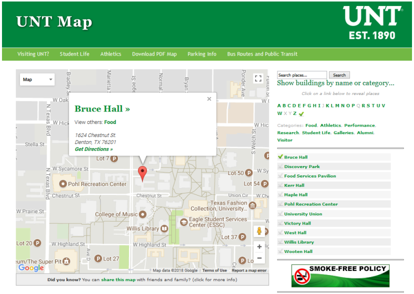
Example 1

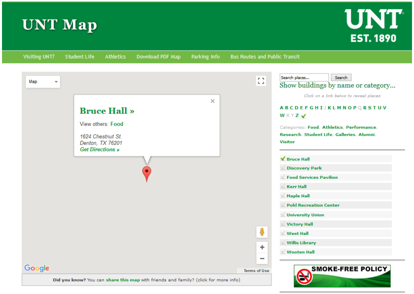
the interactive floorplan isn’t working as it should do - the text should appear over the map when you click on it, rather than in a list underneath

Example 2

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

Interactional Correspondence Problems in Real Life


UNT Map
UNT EST. 1890
Visiting UNT? Student Life Athletics Download PDF Map Parking Info Bus Routes and Public Transit
Map
Bruce Hall »
View others: Food
1624 Chestnut St.
Denton, TX 76201
Get Directions »
Pohl Recreation Center
College of Music
Willis Library
Eagle Student Services Center (ESSC)
Texas Fashion Collection, University...
Search places... Search
Show buildings by name or category...
Click on a link below to reveal places
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
Categories: Food, Athletics, Performance, Research, Student Life, Galleries, Alumni, Visitor
Bruce Hall
Discovery Park
Food Services Pavilion
Kerr Hall
Maple Hall
Pohl Recreation Center
University Union
Victory Hall
West Hall
Willis Library
Wooten Hall
SMOKE-FREE POLICY
Did you know? You can share this map with friends and family? (click for more info)
Google


UNT Map
UNT EST. 1890
Visiting UNT? Student Life Athletics Download PDF Map Parking Info Bus Routes and Public Transit
Map
Bruce Hall »
View others: Food
1624 Chestnut St.
Denton, TX 76201
Get Directions »
Search places... Search
Show buildings by name or category...
Click on a link below to reveal places
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
Categories: Food, Athletics, Performance, Research, Student Life, Galleries, Alumni, Visitor
Bruce Hall
Discovery Park
Food Services Pavilion
Kerr Hall
Maple Hall
Pohl Recreation Center
University Union
Victory Hall
West Hall
Willis Library
Wooten Hall
SMOKE-FREE POLICY
Did you know? You can share this map with friends and family? (click for more info)
Google
Terms of Use

Screenshot of live website　　Screenshot of archived website

Figure: The archived website does not display the map correctly and does not allow users to interact with it

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

Completeness in the Data

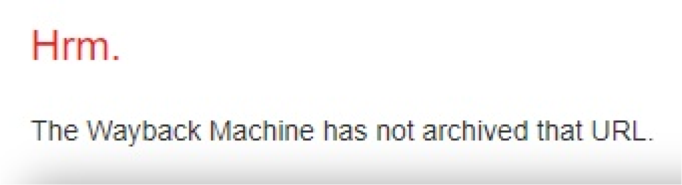
Example 1

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

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

Completeness Problems in Real Life



Hrm.

The Wayback Machine has not archived that URL.

Topical Relevance

The relation between the topic of a file, webpage, or website and the topic of the larger web archive to which it belongs

Topical Relevance in the Data

Example 1

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.

Example 2

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

Relevance Problems in Real Life

- Relevance problems are often not visually apparent
- Instead, it can manifest as *size relevance* problems - too much content
- Some webpages also have *functional relevance*. They are not directly relevant to the topic of the web archive, but are necessary for replaying its "look and feel"
- Examples: Cascading style sheets and JavaScript files
- Functional Relevance: the relation between a file, webpage, or website to accomplishing the goal of successfully replaying an archived website

Archivability in the Data

Example 1

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

Example 2

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

Archivability Problems in Real Life

- Archivability is not a dimension of quality that is directly perceived by many people (latent) and framed in terms of correspondence problems.
- The degree of archivability of a website can be estimated *a priori* by calculating how much of it is composed of dynamic content, such as JavaScript (Brunelle, Kelly, Weigle, & Nelson, 2015). However, its true archivability of a website can only be determined *a posteriori*.
- Archivability can change dramatically over time.

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 correspondence, relevance, and archivability would still be important to the quality in web archives
3. Having clear concepts based on how 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.

References I

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. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Aldine Transaction.

Grbich, C. (2012). *Qualitative data analysis: An introduction* (2nd ed.). London: SAGE Publications Ltd.

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.