WebArchiving@UNT

Current Quality Assurance Practices in Web Archiving

Prepared By

Brenda Reyes Ayala
Brenda.Reyes@unt.edu

Mark E. Phillips
Mark.Phillips@unt.edu

Lauren Ko
Lauren.Ko@unt.edu

August 19, 2014

UNT Libraries
THE POWER OF IDEAS STARTS HERE

Document Information	
Title	Current Quality Assurance Practices in Web Archiving
Author(s)	Reyes Ayala, Brenda. Phillips, Mark E., Ko, Lauren
Original Creation Date	2014-08-19
Version	1.0
Date of Current Version	2014-08-19
Revised By	Reis, Nancy
Description	This paper presents the results of a survey of quality assurance practices within the field of web archiving. To understand current QA practices, the authors surveyed 54 institutions engaged in web archiving, which included national libraries, colleges and universities, and museums and art libraries.
Rights Information	Copyright 2014 Web Archiving Project
Licensing Information	This work is licensed under the Creative Commons Attribution-NonCommercial 3.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc/3.0/>. This license allows you to copy, distribute, transmit, or adapt the content of this work to derivative work. The license requires that you credit the original creators of the work (attribution). This work may not be used for commercial purposes.

Revision History	
Previous Version	None
Changes	None

Table of Contents

Current Quality Assurance Practices in Web Archiving

> “And what is good, Phaedrus,
> And what is not good—
> Need we ask anyone to tell us these things?”
> — Robert M. Pirsig, *Zen and the Art Of Motorcycle Maintenance: An Inquiry into Values*

Abstract

This paper presents the results of a survey of quality assurance practices within the field of web archiving. It was undertaken to address a considerable knowledge gap: practitioners do not know if and how their peers are conducting a quality assurance QA process and generally do not share this information. Consequently, there are no agreed-upon quality standards or processes. To understand current QA practices, the authors surveyed 54 institutions engaged in web archiving, which included national libraries, colleges and universities, and museums and art libraries. The results identified quality as an important issue in web archiving and quality assurance as a process that almost all web archiving institutions undertake, usually after the capture process. Quality assurance is currently undertaken manually at most places, necessitating a significant time commitment from web archivists as well as specialized training, knowledge, and skills. The authors identify a need for the development of tools to automate the QA process.

Introduction

Web archiving is the action of storing Internet resources to preserve them as a historical, informational, legal, or evidential record. The process involves three stages: selecting relevant resources for preservation, gathering and storing them, and providing for their access. A web archive is a system which contains such records.

In recent years, web archiving has become an increasingly common practice in libraries around the world, as national libraries, such as the Library of Congress and the National Library of Australia, seek to preserve their national digital heritage. Many universities have also begun archiving the web, often to create subject-specific collections of web sites that supplement their existing print and digital collections. The International Internet Preservation Consortium (IIPC), founded in 2003, has as its goal the improvement of “the tools, standards and best practices of web archiving while promoting international collaboration and the broad access and use of web archives for research and cultural heritage” (International Internet Preservation Consortium, 2012).

As a member of IIPC, the University of North Texas (UNT) has long been involved with web archiving. Its CyberCemetery project, which began in 1997, is an archive of government web sites that have ceased operation (usually web sites of defunct government agencies and commissions that have issued a final report). UNT Libraries has collaborated with the IIPC in a variety of projects, including the development of the URL Nomination Tool to allow communities of subject specialists to collaboratively recommend websites for upcoming web harvests. During the course of this research, we contacted many IIPC members to solicit their views about the notion of quality in a web archive.

Definition of Quality in a Web Archive

A particular concern to practitioners is the issue of the quality of their web archives. The concept of quality in a web archive was first discussed in depth by Masanès (2006) in his book *Web archiving*. According to him, quality can be defined using the following two aspects:

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

(Masanès 2006)

According to Masanès, the first aspect, the completeness of a web archive, can be measured either horizontally or vertically. A horizontally complete web archive uses a breadth-first approach, crawling many sites at the surface level. Pages deep in the hierarchy of the site will not be captured. In contrast, a vertically complete web archive does a much deeper crawl of fewer sites using a depth-first approach. Ideally a web archive should be both horizontally and vertically complete. (Masanès, p. 39)

We, the authors, feel that the definition put forward by Masanès, though robust, is too centered on the technological tools needed to archive websites. Phrases such as “target perimeter” and “horizontally complete” refer to crawler specifications and so Masanès’ current definition excludes web archives gathered through alternative methods, such as legal deposit, or a simple transfer of files from one institution to another. Though crawling is currently the most oft-used technology in web archiving, technology is constantly evolving, and it may not remain so forever. We will return to this definition of quality later in the Discussion section.

The Quality Assurance Process for Web Archives

The records in a web archive are often put through a *quality assurance* process, which measures the quality of a resource by comparing it to a standard that must be met. The terms quality control, quality analysis, quality review, and quality assessment have also been used to represent this process. Throughout this paper we will use the term quality assurance, or QA for short.

It appears that there is a considerable knowledge gap: practitioners do not know if and how their peers are conducting a QA process and generally do not share this information. Consequently, there are no agreed-upon quality standards or processes. If they exist, QA procedures are often not publicly available and not thoroughly documented, if at all. This led us to formulate the following research question:

What are the current QA practices in the web archiving community?

This research question can be subdivided into the following questions:

- Do web archiving institutions view quality assurance primarily as an automatic or a manual process?
- How do institutions assure the quality of an individual site?
- What tools do institutions use to perform QA when web archiving? Do they rely mostly on existing systems and tools (such as Archive-It, WAS, Heritrix crawl reports) or do they implement their own systems?
- What kind of information do institutions collect about an individual site during the QA process?
- Is QA implemented for every single site or for only a subset of sites?
- How do institutions deal with crawl problems that might negatively affect the quality of their sites?
- What are the most serious quality problems that institutions encounter when archiving sites?

Taken together we felt that answering these smaller questions would help us answer our main research question. To this end, after conducting some preliminary research on how different institutions conduct their QA processes, we designed a survey to help us accurately describe the status of QA within the web archiving community. Our survey includes data from institutions both inside and outside the IIPC.

Previous Work on Quality in Web Archives

The topic of quality in web archives is not new. It has been addressed in various forums, such as the IIPC General Assemblies, and the annual Archive-It Partners meeting. At the 2010 General Assembly for the International Internet Preservation Consortium, Voorburg (2010, May) described the approach taken by the National Library of the Netherlands (KB) regarding quality assurance. According to Voorburg, the KB would capture websites of interest, which would then be checked for quality by web archiving analysts. The analysts would then decide to archive a site if its quality was sufficient, or choose to reject it if serious quality problems existed. Quality problems were recorded by the analysts using annotations within the Web Curator Tool, an archiving tool originally developed by the British Library and the National Library of New Zealand.

In a second presentation in Vienna, Voorburg (2010, September) explained some of the most important issues pertaining to quality assurance in the field of web archiving. Some of these were:

- The large amounts of data involved in web archiving pointed to a need for automation.
- The quality assurance process was complex, requiring special tools, procedures, and experts with specialized knowledge. It did not scale well.
- The issue of quality itself was confusing; it was difficult to define a “good enough” level of quality for a captured resource.

In order to gain a broader perspective on QA, the KB surveyed six institutions as to their QA practices, most of them national libraries. Based on the results, Voorburg presented a list of top 10 quality issues most often seen in web archiving:

1. URIs could not be discovered by crawler (~ 20 %)
2. Access restrictions (~ 15 %)
3. Multi-media missing in harvest (~ 12 %)
4. Display problematic; rewrites fail (~ 10 %)
5. Speculative links cause crawler traps (~ 6 %)
6. Sites have a near endless URI-space (~ 6 %)
7. Seeds missed during selection (~ 4 %)
8. Selection was too broad (~ 2 %)
9. Redirects broaden crawl scope too much (~ 2 %)
10. The ‘agnostos daemon’ : the (yet) unknown problem…

The results indicated that there was room for better and more efficient QA of harvested websites. In order to achieve this, tools and processes for automating QA were needed as an alternative to the expensive and time-consuming process of manual QA. To achieve this goal, Voorburgh recommended starting an IIPC workgroup on quality assurance, so that members would be able to share knowledge and create a common classification and technology.

Our survey expands on Voorburgh’s work. Our sample was diverse, it included respondents who worked at large national libraries, universities, and various other institutions. Many of them did not have a custom, in-house system to carry out their web archiving activities, but relied on pre-built “off the shelf” web archiving tools. Examples of these tools include the Internet Archive’s Archive-It service and the California Digital Library’s Web Archiving Service (WAS). We believe by surveying not just IIPC members, but also reaching out to institutions outside of the IIPC, we have captured an accurate picture of the web archiving community and its practices.

Methods

Three people were involved in this project: the research lead Brenda Reyes Ayala, and the co-authors, Mark E. Phillips and Lauren Ko, who are later referred to as the coders. All three are part of the Web Archiving Team at the University of North Texas Libraries.

Document Analysis and Email Communications

In order to answer our research question, we searched the web for publicly available documents on QA from different institutions with web archiving programs. We found that very few institutions made documentation on their QA processes publicly available. To address this, we decided to expand our efforts to find more information. We found a 2009 discussion on the QA process that had taken place in the IIPC listserv (Kobus, 2009); in this discussion, some of the participants described the QA processes undertaken by their respective institutions. The research lead contacted the participants to ask if their QA processes were still the same or had changed. We also posted a message asking for information on QA on the IIPC listserv and contacted respondents for additional information. All of these sources, online documents, listserv postings, and emails, were analyzed for content. Over time, we began to form a rough sketch of QA in the web archiving community.

Interviews and Meetings

In addition to compiling documentation, we conducted interviews with web archiving staff at the University of North Texas and met (remotely) with several web archivists from the Library of Congress. During the meeting, they explained their QA process and also demonstrated their custom interface for assuring the quality of their captured web resources.

Survey

Despite the helpful information we gained from the collected documents and the interviews, we felt we still lacked sufficient information to accurately describe the status of QA within the web archiving community. We felt that a survey instrument would be the best way to gain a comprehensive view of the ways web archivists handle the issue of quality.

To this end, we designed an online survey for staff at institutions with web archiving programs. It was composed of 24 multiple-choice, fill-in-the-blank, short answer, and long-answer questions. We built the survey using the research suite Qualtrics and distributed it in several ways. First, we sent a message with a link to the survey to the IIPC listservs, and also compiled a separate email distribution list of people we knew to be working in the field. We also disseminated information about the survey during our presentations at the IIPC 2013 General Assembly in Ljubljana, Slovenia and the Texas Conference on Digital Libraries (TCDL) in Austin, Texas.

The survey was anonymous. We did not collect any personally identifying information about the respondents, but only asked for the names of the institutions where they worked. At the end of the survey, participants were asked if they would like to be contacted in the future about QA issues. If they answered “Yes,” they had the choice to provide their contact information. The survey remained open and accessible for several months while responses were gathered.

Analysis of Survey Results

Once we closed the survey, we analyzed the results using SPSS, a popular statistical package for the social sciences. This was a relatively straightforward process when dealing with closed-ended questions; however, the survey also included several in-depth, open-ended questions that needed to be analyzed and interpreted. The research team decided to analyze the following three types of answers separately:

1. Answers to long-answer questions that required the participant to elaborate on a topic.
2. Answers to multiple-choice questions where the participant chose the “other” category and was asked to briefly state an answer.
3. Answers that were vague or confusing.

We decided to analyze these using thematic analysis, which is a qualitative research method for “identifying, analysing and reporting patterns [also called themes] within data” (Braun & Clarke, 2006, p.79). The process of thematic analysis involves familiarizing oneself with the data collected, highlighting interesting aspects of the data, grouping them into units called “codes,” and defining and naming themes. To ensure an accurate interpretation of the data, we also performed a coding consistency check. Our entire coding process is described below, and a sample of the codebook appears in Table 1.

1. The research lead carried out an initial analysis of the data and created a first draft of the codebook.
2. The co-authors repeatedly looked over the codes, made corrections, and worked to further refine them.
3. The lead researcher conducted a training session about thematic analysis and how to code for the co-authors. This was also a chance for further refinement of the codebook.
4. The two co-authors, acting as coders, independently viewed and coded the open-ended responses according to the codebook provided.
5. The research lead performed a side-by-side analysis to determine agreements and disagreements between each pair of classifications. She also resolved classification differences between the two coders and merged them into one set of coded data.

Table 1: *Sample of the Codebook Used to Code the Open-ended Responses in the Survey*[1]

code_abbreviation	code_name	description
browse_site	Browse the site to look for problematic content	Describes any activity that involves accessing a site and clicking around to look for problems. Takes place before the crawl begins
robots_exclusions	Check for robots.txt exclusions	Look at robots.txt to see what parts of a site are excluded from being crawled. Takes place before the crawl begins
test_crawl	Perform a test crawl of the site	Process that takes place before capture in order to insure quality. Refers to a preliminary crawl of the site that is conducted before the actual crawl in order to detect possible problems
link_discovery	Run a program to discover links that might be missed by the crawler	Process that takes place before capture in order to insure quality. Involves deploying program to extract links that might be missed by the crawler, and then adding them

[1] It is important to note that the codes in the codebook were not mutually exclusive, and an answer could be coded as corresponding to several codes. For example, if a respondent indicated that she browsed a site to look for problematic content and also performed a test crawl beforehand, her answer might be coded as both “browse_site” and “test_crawl.” This double coding might affect the frequencies below.

id_crawler_traps	Identify potential crawler traps	Process where problematic content that could trap the crawler is identified. Takes place before the crawl begins

Note. The full codebook is included in the appendix at the end of this paper.

Results

We received 54 completed responses to the survey and classified them according to the type of institution, with 88.9 % of respondents coming from either colleges and universities or national institutions such as national libraries.

Table 2: *Survey Respondents Grouped by Type of Institution*

	Frequency	Percent	Cumulative Percent
Colleges & Universities	23	42.6	42.6
National Institutions	25	46.3	88.9
Museums & Art Libraries	2	3.7	92.6
Other	4	7.4	100.0
Total	54	100.0	

When the participants were asked if they conducted a QA process for their archived websites, over 90% replied “Yes.” Of those that replied “No,” the majority answered that it was due to a lack of staffing or lack of funds in their organization.

Table 3: *Do you conduct a QA process for your archived sites?*

	Frequency	Percent
Yes	49	90.7
No	5	9.3
Total	54	100.0

The participants were asked when they conducted the QA process: before, during, and/or after the capture process. Over half of them (55.6%) answered with “before,” while the next largest group responded with “before and after.”

Table 4: *When do you conduct your QA process?*

Answer	Frequency	Percent	Cumulative Percent
never	5	9.3	9.3
after	30	55.6	64.8

during	1	1.9	66.7
during and after	4	7.4	74.1
before and after	8	14.8	88.9
before, during, and after	6	11.1	100.00
Total	54	100.00	

When asked to describe the process undertaken to assure quality *before* a crawl, participants named a variety of strategies. The most common ones were adjusting a crawler's scope rules to deal with problematic content before conducting a crawl, followed by manual browsing of a site to look for problematic content, and performing a test crawl beforehand. A few respondents deployed link discovery software, such as Xenu Link Sleuth, to identify links that might be missed by the crawler.

Table 5: *Please describe the process you use to assure quality before a crawl begins.*

Answer	Frequency	Rank
Adjust scope rules to deal with problematic content	8	1
Browse the site to look for problematic content	6	2
Perform a test crawl of the site	6	2
Run a program to discover links that might be missed by the crawler	4	3
Identify potential crawler traps	3	4
Check for robots.txt exclusions	3	4
View and analyze the crawl logs	3	4
Other method of assuring quality before a crawl begins	3	4
Total	36	

Most participants (64.3%) viewed QA as a manual process that involves human effort, as opposed to an automated or semi-automated process to be carried out by a software tool. Over a third of participants (34.7%) viewed QA as both a manual and technical process.

Table 6: *Do you generally conduct QA as a: _____?*

Answer	Frequency	Percent	Cumulative Percent
Manual process that involves a person looking at and navigating the site	32	65.3	65.3
Technical process to be done in an automated or mostly-automated fashion by a software tool	0	0	65.3

Both	17	34.7	100.00
Total	49	100.00	

Most participants who used automated or semi-automated methods to assure the quality of their web archives used a variety of tools. Crawl reports produced by web crawlers were by far the most popular tool used with over three quarters of the participants saying they used it to conduct QA. Other popular tools were the QA features within Archive-It, as well as other pre-built, "off the shelf" tools such as Xenu Link Sleuth, PhantomJS, and HTTPFox. A few of the respondents used custom-built tools and systems that were specific to their institutions.

Of the crawl reports used during the QA process, the most widely used were the ones created by Archive-It and Heritrix.

Table 7: *Type of crawl reports used during the your QA process*

Answer	Frequency	Rank
Archive-It crawl logs and reports	5	1
Heritrix crawl logs and reports	5	2
Hosts Report	4	3
Seeds Report	3	4
Crawl summary/crawl report	3	4
Response code report	2	5
MIME Report	2	5
Other logs and reports	2	5
All crawl reports generated by Heritrix	1	6
Source Report	1	6
Total	28	

The respondents identified crawl reports as an important part of the QA process. They are used for a variety of reasons, but the most popular use is to ensure that all the sites of interest were captured. Because the frequencies exhibit low variance, it is safe to assume these uses are all equally important to participants.

Table 8: *How crawl reports are used during the QA process*

Answer	Frequency	Rank

To make sure all the necessary sites were captured	3	1
To find out if extraneous or unnecessary content is being captured	2	2
To check the crawl status to see if the crawl is running well, or if there are any problems	2	2
To identify crawler traps	2	2
To check the size of crawls	2	2
To see if adjustments need to be made to the crawl scope	1	3
Total	12	

The most popular method for assessing the quality of a site was viewing using the Internet Archive's Wayback Machine, followed by viewing it using a proxy server.

Table 9: *If you do manual QA, how you review specific sites? (Check all that apply)*

Answer	Frequency	Rank
View the site using the Wayback Machine	32	1
View the site using a proxy server	17	2
View the site in a browser, no specific information about what platform they are using	4	3
Other type of QA	4	3
View the site within a pre-built web archiving system	3	4
Access the crawl logs and reports to see if the site was captured properly	1	5
Total	61	135.6

Most respondents collected some sort of information about their archived sites. The most oft-collected information was: if content were missing from the site, if the site's appearance resembled the original, depth a user could navigate within a site, if the site's multimedia resources could be successfully played back, and if JavaScript were functioning correctly.

Table 10: *If you do manual QA, what kind of information do you collect about a site? (Check all that apply)*

Answer	Frequency	Rank
If content is missing from site	43	1
If the site's appearance resembles the original	39	2
Depth you can navigate to within a site	33	3

If the site's multimedia resources can be played back	32	4
If Javascript is functioning correctly	31	5
Statistical data (site URI, size of captured site, MIME types, response codes, etc)	22	6
If the site is still present on live web	15	7
Priority of the captured resource	8	8
Other information	7	9
Response not given	2	10
Total	232	

Most participants indicated that they only sometimes compare the captured site to the live target site.

Table 11: *How often do you compare the captured site to the live target site?*

Answer	Frequency	Percent	Cumulative Percent
Always	15	30.6	30.6
Sometimes	32	65.3	95.9
Never	2	4.1	100.00
Total		100.00	

When asked whether they did QA on every site or on a sample of sites, most participants responded that they did QA for every captured site, while a significant percentage (38.8%) indicated that they used a sampling method.

Table 12: *Do you try to do QA on every site you capture or do you use a sampling method?*

Answer	Frequency	Percent
Do QA on every captured site	30	61.2
Use a sampling method: do QA on only a sample of all the captured sites	19	38.8
Total	49	100.00

Of those participants that used a sampling method, most indicated that the process of selecting a sample

depended on the types of sites they were archiving. Others chose a random subset of sites on which to do QA.

Table 13: *Please describe how you determine which sites will be sampled. Also, typically, how large is your sample?*

Answer	Frequency	Rank
QA process depends on the type of site	6	1
QA is implemented on a random subset	5	2
QA is implemented for a subset of sites, but only the most important ones	4	3
QA is implemented for a subset of sites selected for any other reason	3	4
QA is implemented on a subset determined by time constraints	2	5
QA is implemented on all the seeds	2	5
Response not given	2	5
Total	24	

During the QA process, most participants reviewed the entire site, while a smaller number checked only the seed or homepage.

Table 14: *What part of a site do you review during the QA process?*

Answer	Frequency	Rank
The entire site, including all domains and subdomains	23	1
Only the seed or homepage	12	2
Depends on the site	8	3
Other part of a site	5	4
Only a specific subdomain	4	5
Only part of the site that is related to a specific topic	4	5
Home page and another page	1	6
Total	57	

If participants found they could not do an optimal crawl of a site, they aimed for a “good enough” approach to quality by aiming to capture it as best as possible, even if the end result was not perfect.

Many participants also noted the problems with the capture or attempted to recrawl problematic parts of the site.

Table 15: *If you found that you could not do an optimal crawl of a site, how do you address this problem?*

Answer	Frequency	Rank
Aim to capture the site as best as possible, even if it is not perfect	46	1
Note the problems with the capture	30	2
Recrawl the problematic parts of the site	26	3
Manually patch the problematic content	17	4
Discard the capture	13	5
Other approach	4	6
Total	156	

When participants performed focused crawls, most (63.3%) treated all sites with equal priority, while some assigned a higher priority to a smaller set of sites.

Table 16: *Within a focused crawl (where you only capture sites from a specific list rather than an entire domain), do you treat all sites equally?*

Answer	Frequency	Percent
Yes, we treat each site equally	31	63.3
No, we assign higher priority to a smaller set of sites which must be captured as well as possible	18	36.7
Total	49	100.00

Most participants noted any problems with their archived resources in a spreadsheet, or within specialized software such as Archive-It or NAS.

Table 17: *Where do you note problems with the quality of your crawl? (check all that apply)*

Answer	Frequency	Rank
In a spreadsheet	*20*	*1*
Within specialized software such as Archive-It or Net Archive Suite	*18*	*2*
Other, please specify	*8*	*3*
In a database	*7*	*4*

In a wiki	*5*	*5*
Total	*58*	

Of the problems that participants encountered with the quality of an archived site, the most common one was a wrong representation of the site, followed closely by missing content that was not captured. A few participants used the “Other” category to add their own errors, such as problems with software that blocked websites, redirects, difficulty in capturing rich media, and problems with JavaScript and Flash that make playback of the archived site difficult.

Table 18: *Of the errors that you encounter during the QA process, please rank them in order of their frequency, from those that occur most often to those that occur least often or not at all.*

Type of error	Rank
Wrong representation of the site. Ex: video content not playing correctly, menus that do not display well, or layout problems.	1
Missing content. Refers to intellectual content, not layout or appearance.	2
Access or playback errors. Ex:"Resource not found" or "Resource not available."	3
Other type of error	4

In most institutions that carried out web archiving activities, those responsible for QA were the same people involved in implementing a crawl, such as crawl engineer or web archiving specialist. Fewer institutions had dedicated QA staff.

Table 19: *Who is responsible for conducting the QA process?*

Answer	Frequency	Rank
People who worked on implementing the crawl, such as crawl engineers or web archiving specialists	40	1
Dedicated QA staff	14	2
Volunteers/students	12	3
Those who requested the site be crawled	4	4
Total	70	

Discussion

In this section we discuss the findings of the survey and the conclusions we draw from them.

Do web archiving institutions view quality assurance primarily as an automatic or a manual process?

Web archivists view the process of QA as both a manual process, which requires a human to inspect a captured resource, and a technical process to be done in an automated or mostly-automated fashion by a software tool. We have already mentioned that manual QA is labor-intensive, complicated, and requires staff to receive special training; however, our data shows that most web archivists believe that manually looking at a site is important. We surmise there are several reasons for this:

- Lack of tools to automate the QA process: we have identified only one automated QA process, which is implemented by the Library of Congress in conjunction with the Internet Archive. This process, described in a later section, is not completely automated, but it makes manual QA easier and more efficient for web archivists
- Lack of tools to make the manual QA process quicker, more efficient, or less labor-intensive

How do institutions assure the quality of an individual site?

In our survey we found that respondents employ a wide variety of strategies and tools when archiving Internet resources; however, it is possible to speak of a “typical” or “common” QA process that many institutions employ. Here we offer a description of a typical QA process.

- 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

What tools do institutions use to do QA when web archiving? Do they rely mostly on existing systems and tools (such as Archive-It, WAS, Heritrix crawl reports) or do they implement their own systems?

The results of our survey indicated that most institutions involved in web archiving rely on pre-built systems. The most popular and widely-accepted tools are those that were developed by the Internet

Archive, notably the Heritrix crawler and the Archive-It system. The Heritrix web crawler stores archived resources in the Web Archive (WARC) file format, which allows many captured sites to be stored in a single, highly-compressed file. This is useful for institutions that collect many sites and have space limitations. In addition, Heritrix allows for ample customization that allows users to narrow the scope of a crawl in order to capture only desired content. It also creates crawl reports that are highly informative; our participants named them as the most popular tool when performing automated or semi-automated QA.

Crawl reports were another useful tool identified by our respondents. The primary crawl reports used were those generated by Archive-It and the Heritrix crawler itself. They contain statistical information such as the size of the crawl, the time a crawl took to complete, the number of seeds successfully crawled, and the number of hosts crawled. Web archiving institutions use crawl reports for a variety of reasons, but the most popular use is to ensure that all the sites of interest were captured. Other popular tools were the QA features within Archive-It, as well as other pre-built, "off the shelf" tools such as Xenu Link Sleuth, PhantomJS, and HTTPFox.

Archive-It's QA interface helps web archivists assess the quality of a crawled site. From the interface, users can view and inspect an archived site and also view media resources that could not be played back using the Wayback Machine. Internal scripts check if the archived site is missing elements such as links, CSS or JavaScript files. The user then has the ability to run a patch crawl to attempt to capture the missing content.

Xenu Link Sleuth is a software tool that analyzes the structure of a web site. Though it was designed primarily to find broken links in a site, it can be used to generate an XML sitemap and analyze a site's information architecture, which can be very useful for web archiving. Phantom JS is a headless WebKit scriptable with a JavaScript API. Several web archiving institutions employ PhantomJS to create screen captures for their archived web sites. HTTPFox is an add-on to the Firefox browser that monitors and analyzes all incoming and outgoing HTTP traffic between the browser and the web servers. It can be used when doing QA to check that a captured site is functioning correctly.

Our results indicate that most institutions do not implement their own web archiving systems, but rather rely on pre-existing components and full systems. This is not surprising, as designing a complete web archiving system from scratch to include capabilities for selecting relevant resources for preservation, capturing and storing them, and providing for their access is a herculean task. Some large organizations such as the Library of Congress and the National Library of Australia implement their own custom web archiving solutions. Sometimes several institutions will work together to produce a shared web archiving infrastructure, such as the collaboration between the British Library and the National Library of New Zealand to create the Web Curator Tool (WCT) (The Web Curator Tool Project, 2014). However, these are generally not representative of the majority of cases.

What kind of information do institutions collect about an individual site during the QA process?

We found that most institutions collect some type of information about the sites they archive. The most popular information collected was: if content was missing from the site, if the site's appearance resembled the original, the depth to which a user could navigate within a site, if the site's multimedia resources could be successfully played back, and if JavaScript was functioning correctly. If we assume that most web archivists are recording quality problems in a spreadsheet or a database, as our results indicated, this means that they are in all likelihood manually typing all the required information. This is an area that would benefit from some automation.

Is QA implemented for every site or for only a subset of sites?

Most institutions indicated that they perform QA for every single site that is captured. This approach is extremely time-intensive for web archivists, as it requires specialized skills and knowledge, and seems an almost impossible task given a limited amount of time and resources.

In light of these answers, we wish to highlight two significant issues. First, the nature of a collecting institution and its web archives will naturally influence each respondent's answer. A large institution that captures thousands of sites might not be able to implement QA for all of its captured sites, while a smaller institution with smaller collecting goals and a topic-centered web archive might have the resources at hand to perform QA for all of its sites. Second, there is a difference between the concept of a *site* and the concept of a *seed.* A seed is a single URI fed to a crawler, whereas a site is a set of web pages served from a single web domain. For example a seed might be the URI http://www.unt.edu, while a *site* would include the seed URI as well as the many related pages linked to/from the seed URI, such as http://www.library.unt.edu/ and http://www.unt.edu/about-unt.htm. We surmise that the large number of respondents who answered that they implement QA for all sites might stem from a misunderstanding of the differences between the two concepts arising from a lack of clarity or agreement in the web archiving documentation.

How do institutions deal with crawl problems that might negatively affect the quality of their sites?

The most popular answer was "aim to capture the site as best as possible, even if it is not perfect." This points to a key issue: since perfect quality is impossible or practically unattainable, the strategy for a web archivist is to settle for "good enough" quality. The standard for "good enough" varies with each institution. We hypothesize that the size and collection goals of each institution play a significant role in its measure of quality. For example, a large national library that archives its entire national domain may have a lower acceptable standard of good enough quality, since it is impossible to check the quality of every captured site on very large datasets. But a smaller institution that collects a limited number of websites might have higher quality standards, since it is feasible to check the quality of each individual website. Similarly, a user of web archives might also have quality standards that differ significantly from those of the collecting institutions or those of another user. For example, a computer science researcher interested in analyzing the size and volume of a web archive might settle for lower quality in individual sites, while a legal scholar might have very high quality standards that must be met for an archived site to count as a valuable resource.

What are the most serious quality problems that institutions encounter when archiving sites?

1. wrong representation of the site
2. missing content
3. access or playback errors
4. other

We return now to the definition of quality put forward by Masanès (2006). We noted earlier that this definition is very centered on the technological tools needed to archive websites. We found that it does not do a good job of describing the most important quality problems in web archiving. For example, Masanès' definition does not account for access or playback errors that, though unrelated to crawling, may still affect the perceived quality of a site. Also, what if the archived site was not crawled, but captured via an alternate mechanism, or simply deposited electronically at the archiving institution? In this case, concepts such as "target perimeter" and "horizontally complete" would not apply.

As an alternative, we advocate a more general, abstract definition of quality within web archiving which is technology-independent. Defined this way, quality in a web archive would consist of:

- 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, the archived resource is itself the original resource or a copy of it. Consequently, 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.
- Completeness: The archived resource contains all its constituent elements.
- Coherence: The archived resource integrates diverse elements in a logical and consistent manner.
- Integrity: The data elements that constitute the captured resource are uncorrupted and error-free.

With the exception of integrity, these dimensions of quality in a web archive are not absolute or discrete. Rather, they represent a continua of measurement. For example, a site may have a high or low measure of correspondence. It could be argued that perfect quality is unattainable in a web archive.

We can now use this model to describe our findings. If we bracket “access or playback” errors as technical issues that occur due to a bad connection or problems with a server, we are left with the two most important errors: wrong (replay quality) representation and missing content (capture quality). Capture quality is a measure of the coherence, completeness, and integrity of an archived site, whereas replay quality is a measure of its correspondence. The seriousness of these problems points to a need for the development of standards and tools to facilitate the QA process.

Library of Congress: a case study in automated QA

We have previously identified the lack of tools available for automating the QA process. We would now like to discuss the QA implementation done by the Library of Congress in conjunction with the Internet Archive. This information presented was obtained through a meeting with the Library of Congress’ web archiving staff and during the course of a two-week summer internship at the Internet Archive headquarters in San Francisco. It is important to note that this is not an automated QA process, significant human effort is still involved. However, it is as close to an automated QA system as we currently have.

In recent years, the Library of Congress (LOC) has shifted from theme-based web archiving to frequency based web archiving. Crawls of web sites are conducted on a weekly, monthly, and annual basis using the Heritrix crawler. The tools used for this QA process include Hadoop, Pig scripts, and PhantomJS. The following is a description of a weekly crawl:

1. Precrawl. Once the seedlist has been determined, a preliminary crawl will begin. The precrawl only touches the home page of each seed. The goal is to detect any SURT or seed issues in the crawl. Any possible problems are communicated to the LOC web archiving team, and the seedlist and crawl settings will be adjusted accordingly.
2. Production crawl. The production crawl runs for seven days.
    a. Detect any SURT and seed issues. These are reported within 24 hours to the LOC Web Archiving Team.
    b. Generate CDX and WAT files for the crawl. A WAT file contains metadata for each WARC file and is extremely useful for data analysis.
    c. Log any problematic content such as spam, crawler traps, and link farms.
3. Day 7: Automated QA
    a. Perform browser analysis. This step produces a list of missing files that need to be captured.
    b. Perform link analysis on WAT files. This step produces a list of missing embedded content that need to be captured.
    c. Add all the missing content to the crawler frontier.

4. Patch crawl. Takes place over the course of two to five days
    a. Identify whether the quality problem is a replay issue or a capture issue.
    b. Crawl seeds in frontier.
5. Human QA. Curators browse the archived content, view it in proxy mode, and check the YouTube reports to see if videos have been properly captured.

Using browser analysis for automated QA

This process uses PhantomJS, which is a browser emulator and headless WebKit. Browser analysis is done only for the more important seeds, which are usually specified by the LOC.

1. Load archived site into browser using Wayback (usually in proxy mode)
2. Take a screenshot of the site's homepage
3. Record the response codes obtained
4. Generate a report for each URL, which are then synced to a central directory; Reports are in HAR format (HTTP archive format)
5. Parsing tool (JSON) extracts 404 response codes from reports and adds them to Heritrix for a future patch crawl
6. Load all screenshots into a CoolIris wall (interface for browsing through images). This allows a human curator to look at the archived web site, inspect it, and detect any possible problems.

Using link analysis with WAT files for automated QA

1. Extract the outlinks and embedded content (embeds) from the WARC files. Embeds include content such as images, link tags, and anchor text.
2. Run a Pig scripting job to extract and parse the WAT files
3. Get triples for the outlinks. These triples are in the format:

```
source link/resource --- destination resource --- link type
```

4. Store the triples into buckets depending on link types
5. Focus on embedded objects, such as CSS and JS files. We know source URL was crawled (that's why it is in the WAT file) but was destination resource crawled?
6. Load all URLs that were crawled from the index.cdx file (use canonicalized version of the URL)
7. Canonicalize all destination URLs
8. In Hadoop, do a JOIN operation between the two lists: crawled resources vs. embedded destination resources, and look at what has been left out. These are the resources that have not been crawled, but should be
9. URLs that were not crawled become candidates for a patch crawl
10. Calculate a measure of crawl completeness: # of embeds crawled / total # of embeds. For example, 3,000 embedded URLs crawled / 10,000 embedded URLs present = 30% crawl completeness

When the process was first implemented, Internet Archive staff discovered that the browser analysis and link analysis processes each discovered *different* resources. They determined that both processes are necessary to improve the quality of a crawl. Furthermore, in our communications with the staff at the Internet Archive, they noted that their automated QA process has succeeded in greatly increasing capture quality, but not replay quality. As one staff member put it "Human QA is best for detecting replay problems."

Conclusion

Our research identified quality as an important issue in web archiving and quality assurance as a process that almost all web archiving institutions undertake. Quality assurance processes usually take place after the capture phase and employ various pre-built tools from the Internet Archive, usually the Archive-It platform and the Heritrix web crawler. Most institutions also note any quality problems they encounter, though only a minority use the information to recrawl the site or attempt to manually patch the crawled content.

Our survey results indicated that quality assurance is currently undertaken manually at most places, necessitating a significant time commitment from web archivists as well as specialized training, knowledge, and skills. The scope of many captures, as well as the small size of most web archiving teams makes manual QA impractical. We identify a need for the development of tools to automate the QA process, or at least some portions of it, and present the approach taken by the Internet Archive and the Library of Congress as a possible way forward.

References

Braun, V. & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*, 77-101.

Grbich, C. (2013). *Qualitative Data Analysis: An introduction* (2nd ed.). London, England: Sage Publications.

International Internet Preservation Consortium. (2012). Missions & Goals. *International Internet Preservation Consortium*. Retrieved January 31, 2014, from http://netpreserve.org/about-us/mission-goals.

Kobus, M. (2009, April 4). QA Process [Message posted to IIPC Web curators mailing list]. Retrieved from http://iipc.simplelists.com/curators

Masanès, J. (2006). *Web archiving*. Berlin, Germany: Springer.

The Web Curator Tool Project (2014). *The Project Team*. Retrieved June 13, 2014, from http://webcurator.sourceforge.net/team.shtml

University of North Texas Libraries (2014). CyberCemetery. UNT Libraries: *CyberCemetery Home*. Retrieved August 12, 2014, from http://govinfo.library.unt.edu/

Voorburg, R. (2010, September) Improving quality assurance for selective harvesting. Presented at the IIPC Access Working Group Meeting,Vienna, Austria.

Voorburg, R. (2010, May). QA of Webharvests @ KB.nl. Presented at the IIPC General Assembly, Singapore.

Appendix

RQ = Research Question
SQ = Survey Question

RQ	SQ	code_abbreviation	code_name	description
	Can apply to any answer	response_not_given	Response was not given.	
Do web archiving institutions view Quality Assurance primarily as an automatic or a manual process?	Do you generally conduct QA as a manual process, an automatic process, both, or other type of process?	manual_qa	QA as a manual process	Involves a person looking at and navigating the archived site to check its quality
		automatic_qa	QA as an automatic process	Technical process to be done in an automated or mostly-automated fashion by a software tool
		both_manual_auto_qa	QA as both an automatic and a manual process	Use of both manual QA (Involves a person looking at and navigating the archived site to check its quality) and automatic QA (technical process done by a software tool)
How do institutions assure the quality of an individual site?	Describe the process you use to assure quality before a crawl begins	browse_site	Browse the site to look for problematic content	Describes any activity that involves accessing a site and clicking around to look for problems. Takes place before the crawl begins
		robots_exclusions	Check for robots.txt exclusions	Look at robots.txt to see what parts of a site are excluded from being crawled. Takes place before the crawl begins

		test_crawl	Perform a test crawl of the site	Process that takes place before capture in order to insure quality. Refers to a preliminary crawl of the site that is conducted before the actual crawl in order to detect possible problems
		link_discovery	Run a program to discover links that might be missed by the crawler	Process that takes place before capture in order to insure quality. Involves deploying program to extract links that might be missed by the crawler, and then adding them
		id_crawler_traps	Identify potential crawler traps	Process where problematic content that could trap the crawler is identified. Takes place before the crawl begins
		adjust_scope_rules	Adjust scope rules to deal with problematic content	Process where a crawler's scope rules are adjusted to allow for better capture quality. Can take place before or during a crawl
		analyze_crawl_logs	View and analyze the crawl logs	Process where crawl logs are viewed and analyzed in order to check for problems. May occur at any stage of the QA process
		other_qa_before_capture	Other method of assuring quality before a crawl begins	Method for assuring QA before a crawl begins that is not covered by the above codes. Can also cover

				ambiguous answers.
	Please briefly describe what type of crawl reports you use and how you use them in your QA process			
		Type of crawl reports and logs		
		archiveit_logs_reports	Use Archive-It crawl logs and reports	Describes the use of the crawl logs generated by Archive-It to perform the QA process
		other_logs_reports	Other logs and reports	Describes the use of crawl logs and reports generated by Heritrix or other tools to perform the QA process. Also includes the use of custom-generated reports
		heritrix_crawl_logs	Use Heritrix crawl logs and reports	Describes the use of the crawl logs and other reports generated by Heritrix to perform the QA process
		all_crawl_reports	All crawl reports generated by Heritrix	Describes the use of all crawl logs and other reports generated by Heritrix to perform the QA process
		seeds_report	Seeds report	Describes the use of the Heritrix seed report to perform the QA process. The seeds report is a list of seeds and whether or not the seed was crawled. It also contains the URI to which the seeds was redirected

		hosts_report	Hosts report	Describes the use of the Heritrix hosts report to perform the QA process. The hosts report includes a list of hostnames, the number of URIs crawled per host, and the number of bytes crawled per host
		source_report	Source report	Describes the use of the Heritrix source report to perform the QA process. The source report includes a list of seeds, the hosts that were accessed from that seed, and the number of URIs crawled for each seed-host
		mime_report	MIME report	Describes the use of the Heritrix MIME report to perform the QA process. The MIME report includes a list of MIME types, the number of URIs crawled per type, and the number of bytes crawled per each type
		crawl_report_summary	Crawl summary/crawl report	Describes the use of the Heritrix crawl summary to perform the QA process. The crawl summary includes information about the status and duration of a crawl, the number of seeds crawled/not crawled, the number of hosts crawled, the

				number of URIs, and the total amount of data crawled
		responde_code_report	Response code report	Describes the use of the Heritrix response code report to perform the QA process. The response code report includes a list of response codes and the number of URIs crawled for each code
		How reports are used		
		adjust_scope	To see if adjustments need to be made to the crawl scope	Describes the use of logs and reports to see if the crawl scope needs to be adjusted
		filter_content	To find out if extraneous or unnecessary content is being captured	Describes the use of logs and reports to see if the crawl scope needs to be adjusted
		ensure_capture	To make sure all the necessary sites were captured	Describes the use of logs and reports to see if all the necessary content was captured
		know_hosts_crawled	To know about the specific hosts that were crawled	Describes the use of logs and reports to find out more about the hosts that were crawled
		check_crawl_status	To check the crawl status to see if the crawl is running well, or if there are any problems	Describes the use of logs and reports to see if the crawl is running well, or if there are any problems
		if_media_captured	To see if rich media has been captured	Describes the use of logs and reports to see if rich media (such as video/audio)

				has been captured
		id_crawler_traps	To identify crawler traps	Describes the use of logs and reports to identify possible or actual crawler traps
		check_crawl_size	To check the size of crawls	Describes the use of logs and reports to check the size of the crawl
	If you do manual QA, how do you review specific sites?	view_proxy	View the site using a proxy server	Describes a situation where Wayback has been configured to act as an HTTP proxy server. This prevents Wayback from retrieving any sites or content from the live web
		view_wayback	View the site using the Wayback Machine	Describes a situation where the site is viewed in archival URL replay mode using the Wayback Machine. This allows Wayback to retrieve sites or content from the live web
		view_browser	View the site in a browser, no specific information about what platform they are using	Describes the viewing of a site in a browser, but no information is given about platform, settings, or mode being used
		check_crawl_logs	Access the crawl logs and reports to see if the site was captured properly	Describes use of crawl logs or reports to see if a site was properly captured
		view_pre-built_system	View the site within a pre-built web archiving system	Describes the process of viewing a site using a pre-built, readily available

				web archiving system
		view_custom_system	View the site within a custom-built web archiving system	Describes the process of viewing a site using a web archiving system that was developed internally, and is not available to those outside the institution
		other_qa	Other type of QA	Some other type of QA not covered by previous choices
	What part of a site do you review during the QA process?	entire_site	Review the entire site	The entire site is put through the QA process, including all pages and subdomains and their corresponding elements
		subdomains	Review only a specific subdomain	Only a specific subdomain is put through the QA process
		part_site_topic	Only part of a site related to a specific topic	Only a part of a site that is related to a chosen topic is put through the QA process
		seed	The seed or homepage	A site's homepage or seed is put through the QA process
		depends	Depends on the site	The part of a site that is reviewed depends on the context. Different types of sites are reviewed differently
		homepage_plus_one_page	Home page and another page	Only the homepage and another page are put through the QA process
		other_part	Other part of a site	Other part of a site not covered

				by the above options
What kind of tools do institutions use to do QA when web archiving? Do they rely mostly on existing systems and tools (such as Archive-It, WAS, Heritrix crawl reports) or do they implement their own systems?	If you do automated or semi-automated QA, what kind of tool do you use?	use_pre-built_system	Use of pre-built systems and tools for conducting QA	Refers to the use of readily-available tools, platforms, and workflows
		use_custom_system	Use of custom-built systems and tools for conducting QA	Refers to the use of tools, platforms, or workflows that were developed internally, and are not available to those outside the institution
What kind of information do institutions collect about an individual site during the QA process?	If you do manual QA, what kind of information do you collect about a site?	javascript_issues	If Javascript is functioning properly	Refers to whether the Javascript elements of a site are functioning correctly or incorrectly
		appearance_resembles_original	If the site's appearance resembles the original	Refers to checking if the archived site visually resembles the live one
		other_info	Other type of information is gathered	Other information not covered by above categories
Is QA implemented for every single site or for only a subset of sites?	Please describe how you determine which sites will be sampled. Also, typically, how large is your sample?	qa_for_all	QA is implemented for every single site that is captured	QA is implemented for all sites that have been captured
		qa_depends	QA depends on the type of site	They type of QA implemented depends on the type of site.

				Different sites get different types of QA
		qa_important	QA is implemented for a subset of site, but only the most important ones	QA is implemented only for subset of sites deemed most important and which must be captured as well as possible
		qa_random	QA is implemented on a random subset	QA is implemented for a randomly selected subset of sites
		qa_time_constraint	QA is implemented on a subset determined by time constraints	QA is only implemented for as many sites as the person has time to check
		qa_seeds_all	QA is implemented on all the seeds for a crawl	QA is only implemented for the seed pages of a crawl
		qa_other_subset	QA is implemented for a subset of sites selected for any other reason	QA is implemented for a subset of sites selected for any other reason that is not described above
How do institutions deal with crawl problems that might negatively impact the quality of their site?	If you found that you could not do an optimal crawl of a site, how do you address this problem?	recrawl	Recrawl the problematic parts of the site	Involves launching a second crawl that will only capture parts of the site the caused problems during the original crawl
		manual_patch	Manually patch the problematic content	Involves an alternative process of "repairing" an archived site, sometimes by downloading the needed content from an alternative source and adding that to

				the content that has been already archived
		other_approach	Other approach not covered by the options above	Other approach not covered by the options above
What are the most serious quality problems that institutions encounter when archiving sites?	Of the errors that you encounter during the QA process, please rank them in order of their frequency, from those that occur most often to those that occur least often or not at all.	missing_content	If content is missing from site	Refers to missing intellectual content of the site, not its layout or appearance. This content should have been captured in the first place. Examples include missing HTML pages, PDF documents, or video files.Sometimes occurs because of inapporpriate crawl scope, robots.txt exclusions, or problems with permissions
		wrong_representation	Wrong representation of the site	Refers to a display problems associated with a site. Includes problems such as video content not playing correctly, menus that do not display well, or layout problems
		crawler_traps	Errors due to crawler traps	Refers to quality problems caused by crawler traps during the capture
		js_flash_problems	Problems related to Javascript or Flash	Refers to quality problems caused by problems with a site's Javascript or Flash components
		permission_problems	Permission problems	Refers to problems caused by sites that block crawlers or do not

				grant their permission to be captured. For example, authentication permissions, robots.txt exclusions, and blocked crawlers
		other_errors	Other problems not described by the above categories	Other problems not described by the above categories