Current Quality Assurance Practices in Web Archiving

Brenda Reyes Ayala, Mark Phillips, and Lauren Ko
University of North Texas
Brenda.Reyes@unt.edu, Mark.Phillips@unt.edu, Lauren.Ko@unt.edu

IIPC General Assembly
May 22, 2014
Paris, France
Research Questions

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

- 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 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?
- 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 impact the quality of their site?
- What are the most serious quality problems that institutions encounter when archiving sites?
Methods

- Document Analysis
- Email Communications
- Interviews & Meetings.
- Survey
  - Distributed to institutions involved in Web Archiving, both IIPC members and non-members
  - 54 completed responses
  - 23 colleges & universities, 25 national institutions, 6 other type of institutions
  - close-ended questions analyzed with SPSS, a common statistical analysis package
  - open-ended questions analyzed using thematic analysis method by authors
  - 90% conduct QA for their archived sites
Do web archiving institutions view QA primarily as an automatic or a manual process?

<table>
<thead>
<tr>
<th>Answer</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual process that involves a person looking at and navigating the site</td>
<td>32</td>
<td>65.3</td>
</tr>
<tr>
<td>Technical process to be done in an automated or mostly-automated fashion by a software tool</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Both</td>
<td>17</td>
<td>34.7</td>
</tr>
<tr>
<td>Total</td>
<td>49</td>
<td>100.00</td>
</tr>
</tbody>
</table>
How do institutions assure the quality of an individual site?

What a typical QA process looks like:
1. QA is done after the sites are captured
2. QA is done manually: This involves a person that looks at the archived version of the site and assesses its quality
3. View the site using the Internet Archive’s Wayback Machine
4. QA is done on every captured site. Also, the entire site is put through the QA process, not just the homepage or specific domains
5. Quality problems are noted, either in a spreadsheet or in another system such as a database
6. QA is done by the same person who implemented the crawl, such as a crawl operator or engineer
What kind of tools do institutions use to do QA when web archiving?

1. Heritrix crawl logs and reports
2. Other tools used: Xenu Link Sleuth, PhantomJS, and HTTPFox
3. QA features within Archive-It

- most institutions do not implement their own web archiving systems, but rather rely on pre-existing systems
- some large organizations such as national libraries implement their own custom web archiving solutions, but 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?

1. If content was missing from the site
2. If the site’s appearance resembled the original
3. The depth a user could navigate to within a site
4. If the site’s multimedia resources could be successfully played back
5. 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 single 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. It requires specialized skills and knowledge.
How do institutions deal with crawl problems that might negatively impact the quality of their site?

<table>
<thead>
<tr>
<th>Answer</th>
<th>Frequency</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aim to capture the site as best as possible, even if it is not perfect</td>
<td>46</td>
<td>1</td>
</tr>
<tr>
<td>Note the problems with the capture</td>
<td>30</td>
<td>2</td>
</tr>
<tr>
<td>Recrawl the problematic parts of the site</td>
<td>26</td>
<td>3</td>
</tr>
<tr>
<td>Manually patch the problematic content</td>
<td>17</td>
<td>4</td>
</tr>
<tr>
<td>Discard the capture</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>Other approach</td>
<td>4</td>
<td>6</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Type of error</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wrong representation of the site. Ex: video content not playing correctly, menus that do not display well, or layout problems. <em>Replay Quality</em></td>
<td>1</td>
</tr>
<tr>
<td>Missing content. Refers to intellectual content, not layout or appearance. <em>Capture Quality</em></td>
<td>2</td>
</tr>
<tr>
<td>Access or playback errors. Ex: &quot;Resource not found&quot; or &quot;Resource not available.&quot;</td>
<td>3</td>
</tr>
<tr>
<td>Other type of error</td>
<td>4</td>
</tr>
</tbody>
</table>
Conclusions & Future Work

- 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.
- Final whitepaper with full survey results will be released in June (UNT Digital Library).