Current Quality Assurance Practices in Web Archiving

Brenda Reyes Ayala. University of North Texas. brenreyes@gmail.com

Research problem: Considerable knowledge gap - Practicioners do not know if and how their peers are conducting a QA process and generally do not share this information. If they exist, QA procedures are often not publicly available and not thoroughly documented, if at all.

Research question: What are the current QA practices in the web archiving community?

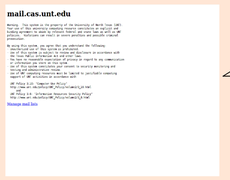
Background

Quality in a web archive can be defined using the following two aspects (Masanes 2006)

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

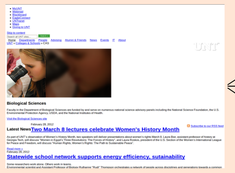
Example

Bad



- Error message from original site
- Desired content cannot be played back

Better



- Content was captured and is present
- Original appearance of website is lost
- Multimedia content missing or unplayable

Best

- Content was captured and is present
- Original appearance is preserved
- Multimedia content present and playable

Methodology

1. Document analysis and email communication

   three publicly available documents examined
   eight institutions were contacted about QA

2. Interviews

   two interviews with web archiving staff

Results

	BnF	Michigan (CDL)	Archive-It	BL (WCT)	NLA (PANDAS)	Denmark (NAS)	Switzerland	LoC
Conduct broad crawls	√			√	√			
Conduct focused crawls	√	√	√	√	√	√	√	√
Distinguish between high-priority and low-priority sites		√						√
Include click-through review of archived sites		√	√	√		√	√	√
Compare harvested site to live site		√	√	√			√	
Documents errors and problems on a specific system		√	√	√	√			√
Create metadata records for each site		√	√					
Review and analyze crawl reports	√	√	√	√		√		√

QA survey

Recognizing that QA in Web Archiving is still an ongoing activity, UNT has designed a survey to find out more about current QA practices. It has been distributed widely in order to ensure the maximum number of respondents, since many institutions that conduct web archiving are not members of the IIPC

Take our survey!

http://bit.ly/13Dpc7

Conclusions and Next Steps

- Web Archiving is still in its infancy. Therefore, more research is needed to get a general overview of the state of QA processes within the Web Archiving community
- End result of survey will be a white paper, which will include a discussion of QA and anonymized survey results
- Possible development of framework for quality assurance
- Exploration of what QA processes can be successfully automated without sacrificing quality and which ones cannot

UNT Libraries