July 2010 NDIIPP Partners Meeting Breakout Session

Challenges in Web Archiving
Breakout Session #9
NDIIPP Partners Meeting
Wednesday July 21, 2010
4:15 p.m. – 5:30p.m

Presenters: Martha Anderson, Library of Congress
Andrea Goethals, Harvard University
Abbie Grotke, Library of Congress
Gildas Illien, National Library of France
Kris Carpenter Negulescu, Internet Archive
Mark Phillips, University of North Texas
Tracy Seneca, California Digital Library

Attendees: 19

Issues Addressed: Challenges

- The amount of time needed to crawl, identifying content and performing Quality Review.

- Challenges with partners and budget cuts, the rate for each collection.

- Website rules constantly change. i.e. Youtube content capture rules, spam, sites that require a log on.

- Not all data can be or will be crawled, the volume of content being stored, building a digital repository, data being lost during a crawl, what pages on the website will be captured during the crawl that week, moving content, navigating through massive amounts of information.

- 70% of the information being captured is generated by individuals and communities. What is the criteria? Is the content good?

- Ongoing accessibility of harvest data, making archives more usable.

- Blending content from multiple locations, being able to search individually and collectively.

- Integrating content into a repository.

- The way the files are packaged – ARC format/zipped. The tools needed to read the files, collection selection, clarification of the selections

- Motivating people to nominate, getting people to believe in archiving, keeping the researchers engaged (they only want the data and not the entire website), they don’t want private conversations archived, archiving networking sites (personal data being available in a public archive).

- Team resources.

- Providing access for researchers.

- Harvesting and doing full in house crawls.

- Networking sites and privacy

Strategies

- Broaden the ‘End of Life’ with regards to collections.

- Change crawl frequencies.

- Enable partners to drive their own archiving programs.

- Partner with researchers to identify new approaches.

- Using what the Librarian’s are currently using.

- Using subject specialists.

- Identify a set of metrics for materials in web archiving.

- Create a process of understanding web archives.

- Format for better storage – using the PREMIS storage model