ICA Africa Programme
Programme pour l'Afrique

ICA Training Programme
Programme de Formation

Digital Records Curation Programme

Week 10:
Web and Social Media Archiving

Week 9 Recap

What did you learn?

- Class on Access and Description
- Seminar on Low Cost Digital Preservation Strategies

Learning Outcomes

At the end of this class, students should be able to:

- understand why some websites and social media account should be preserved
- understand approaches to preserving websites
- develop strategies for preserving social media

Discussion – Websites as Records

Consider if websites are records.

What characteristics do websites have that make them records (or not)?

Why might we want to preserve websites?

What is web archiving?

- Remote harvesting using web crawlers
- The WARC (Web ARChive) format specifies a method for combining multiple digital resources into an aggregate archival file together with related information. The WARC format is a revision of the Internet Archive's ARC File Format that has traditionally been used to store "web crawls" as sequences of content blocks harvested from the World Wide Web. The WARC format generalises the older format to better support the harvesting, access, and exchange needs of archiving organisations.
- Capturing dynamic content: where are the boundaries?
- Appraisal, curation, preservation, access

International Internet Preservation Consortium

- In July 2003, the IIPC was formally chartered at the National Library of France with 12 participating institutions.
- The IIPC is a membership organization dedicated to improving the tools, standards and best practices of web archiving.
- The WARC archival standard, the Heritrix crawler, and the WARC analytic tools are all products of IIPC working groups.

IIPC’s Web Archiving Chain

Toolkit

Acquisition

- An archival-quality crawler capable of web-scale operation.
- A portable database extraction and migration tool for database-driven Deep Web sites.

Focused selection and verification

- Analysis and prioritization tools for dynamic crawl re-focusing.
- User-friendly interfaces for curatorial activities such as selecting, monitoring and verifying archived websites.

Collection storage and maintenance

- File manipulation tools.

Access and finding aids

- An interface for browsing web archive file containers, providing management of the linking environment, URI presentation, and temporal navigation.
- A full-text indexer, scalable to large collections and minimally supporting boolean operators, proximity queries, and the temporal dimension of the archive.
- A query interface generator for archived databases.

How can we archive websites?

- Conifer: https://conifer.rhizome.org/
- The Internet Archive: https://archive.org/
- Screenshots?

Social Media and Memory

- Why preserve social media?
- https://www.youtube.com/watch?v=I-k30ZzFoT0

Group Work – Social Media Preservation Strategies

- Working in two groups, use the State Records NSW guidance to develop an approach to social media preservation for:

- Group 1: A charity that provides services to the homeless
- Group 2: An international mining company
- Group 3: A local authority
- Group 4: A political party

Any questions?