Web Archiving at K-State: Archive-It

4 December 2014
Web Editors’ meeting
Cliff Hight, University Archivist
Kansas Archive-It Consortium Intro

- Developed during late 2013 and 2014
- Members include:
  - Kansas Historical Society
  - K-State
  - KU (including Dole Institute of Politics)
  - Washburn University
  - Emporia State
  - Fort Hays State
Outline for Rest of Presentation

• Why preserve web content?
• What tools are we using?
• How do archived crawls look?
• What will we do next?
Why Preserve Web Content?

- Content is moving there
  - Paper used to be main medium
  - Some never sees paper today
Why Preserve, II

- Historical interest
  – See websites throughout time
- Visual history of web design
Why Preserve, III

• Potential for research (Think K-State 2025!)
  – Preserves content that often disappears later
  – Government information in the digital era
  – Machine access for types of “big data” analysis

Sources for further reading:


• Oxford Internet Institute, “Using Web Archives: A Futures Perspective,” February-June 2011, see: http://www.oii.ox.ac.uk/research/projects/?id=85.
• Archive-It partners, 2006-2014
What Tools Are We Using?

• Internet Archive (http://archive.org/)
  – Home of collections for video, live music, audio, texts, TV news, and more
  – Home of software collection (http://archive.org/details/software)
  – Home of Internet Arcade (http://archive.org/details/internetarcade)
  – Home of Wayback Machine (http://archive.org/web)
  – Home of Archive-It service (the one that matters for this presentation, (http://www.archive-it.org/)}
What Tools, II

- Archive-It tools
  - Heritrix Web Crawler collects content
    - Written in Java by Internet Archive and others, open and free
    - Writes content to WARC files, dedups documents, etc.
  - Umbra Browser Automation Tool also captures
    - Allows preservation of dynamic components of sites
- NutchWAX for full text searching
- Wayback Machine for viewing and access
- Solr for metadata searching
How Do Archived Crawls Look?

- Example from Emporia State University
  - [http://www.archive-it.org/organizations/892](http://www.archive-it.org/organizations/892)
How Crawls Look, II

- K-State’s site on Wayback Machine
  - [http://web.archive.org/web/*/k-state.edu](http://web.archive.org/web/*/k-state.edu)
What Will We Do Next?

- Determine what to crawl
  - K-State sites
  - Collection strength areas (cooking, agriculture, Kansas life and culture, military history, consumer movement, etc.)
  - Possibly create a web-based nomination form

- Make content publicly available

- Publicize availability
Thanks!

Questions?

Library: www.lib.k-state.edu
Special Collections: www.lib.k-state.edu/special-collections
Archive-It collections: www.archive-it.org/organizations/890

My email: chight@ksu.edu