Information sharing about Columbia University Library’s recent web archiving conference

Web Archiving Collaboration: New Tools and Models

Anna Perricci
Columbia University Libraries

2015 Archive-It Partner Meeting
August 18, 2015
• This conference was part of a set of projects within a Mellon funded project

• About 10 slide decks on associated Mellon supported web archiving collaborations are available on SlideShare: http://www.slideshare.net/annaperricci
We had a conference
(it was pretty awesome)
Conference: why & how

• This conference was a follow up to incentive awards program for improving web archiving tools

• Goals:
  – Share results of funded projects and foster conversations about application & extensibility of work done
  – Provide information about other new collaborative models for web archiving

• Space limitations - Day 1 and Day 2
  – Videos: https://www.youtube.com/playlist?list=PLf1Dab4lwQhBpFRB1dpUnKLglmM2iScj

• Co-organizers
  – Bob Wolven, Alex Thurman and Anna Perricci
  – Thanks also to many volunteers and those kindly conscripted
Keynote address

https://www.youtube.com/watch?v=kr82McjQGyQ&index=1&list=PLf1Dab4lwQhBpFRB1dpUnKLgImM2iScj
We need radical collaborative models to do a better job preserving & making available born digital materials
Incentive award winners: Preserving online law content

https://www.youtube.com/watch?v=t_qZ4hNtmyw&index=2&list=PLf1Dab4lwQhBpFRB1dpUnKgLmM2iScjl
Free Law Project
Free Law Project & developments

Free Law Project goals:

• Put entirety of US Case Law online for the public for free

• Develop web-based legal research tools

• Support academic research on legal corpora

• See CourtListener: https://www.courtlistener.com

Results from grant

• Expanded capacity of Juriscraper to harvest opinions on appellate court websites & involve wider community in scraping work
  – https://github.com/freelawproject/juriscraper
  – Over 200 web pages checked every weekday, harvest when changes detected
  – Federal circuit courts and Supreme Court; now covers all state courts of last resort & most intermediate appellate state courts

• Capture oral arguments

• Collaborative work with RECAP
Perma.cc

- Premise: https://perma.cc/about
  - Per http://dx.doi.org/10.2139/ssrn.2329161
    “approximately 70% of all links in citations published between 1999 and 2011 no longer point to the same material. Broken links in journal articles undermine the citation-based system of legal scholarship by obscuring the evidence underlying authors’ ideas.”

- “Any author can go to the Perma.cc website and input a URL. Perma.cc downloads the material at that URL and gives back a new URL (a “Perma.cc link”) that can then be inserted in a paper.”

- There are some more steps to read about on their website, including some stats: https://perma.cc/stats

Results

Better APIs for greater extensibility of tools and wider use of technology in other settings / for other services
New collaborative models

https://www.youtube.com/watch?v=ntAP064ZdBM&list=PLf1Dab4lwQhBpFRB1dpUnKLgIM2iScjI&index=3
New York Art Resources Consortium (NYARC)

- NYARC is a consortium of museum libraries at the Frick Collection, the Museum of Modern Art and the Brooklyn Museum

- With Mellon funding they are preserving and making available websites on themes ranging from artists websites to their own institutions’ websites to resources for restitution related research

- Cutting edge work on access via a discovery layer is in progress

- They are leaders in work on quality assurance
  - Excellent resource created by NDSR resident: http://wiki.nyarc.org/web-archiving/quality-assurance
Collaborative collecting with Ivy Plus
International Internet Preservation Consortium (IIPC) collaborative collections

- Access working group of the IIPC is leading an initiative to collaboratively collect websites on topics of international interest

- Complications and challenges arose particularly in determining collecting themes and given variations in laws governing intellectual property

- Archive-It chosen as tool for creating collection

- No spoiler alert: watch the video to see historic moment in IIPC’s collaborative collecting
Lunch

Image source: http://www.omgcatsinspace.com/post/38940531331
Lightning Talks

https://www.youtube.com/watch?v=jqrC0Ygdc-M&list=PLf1Dab4lwQhBpFRB1dpUnKgLmM2iScjI&index=4
Lightning talks: Michael Lissner on RECAP
Lightning talks: Jefferson Bailey on Archive-It Researcher Services
Lightning talks: Dragan Espenschied on Rhizome’s web archiving work
Lightning talks: Dan Chudnov on Social Feed Manager
Lightning talks: Jack Cushman on Tools for Time Travel
Incentive award winners: New platforms

https://www.youtube.com/watch?v=6h8MohBSETl&index=5&list=PLf1Dab4lwQhBpFRB1dpUnKLGlmM2iScj
Warcbase: Building a scalable platform on HBase and Hadoop
Warcbase

Warcbase is an open-source platform for storing, managing, and analyzing web archives using current “big data” infrastructure and tools (e.g. HBase for storage, Hadoop for data analytics)

-further applications on ‘wimpy hardware’ (Raspberry Pi) also demonstrated for personal digital archiving

For more information on Ian’s work: [http://ianmilligan.ca/](http://ianmilligan.ca/)

For more information on Warcbase: [https://github.com/lintool/warcbase](https://github.com/lintool/warcbase)
Archiving transactions towards an uninterruptible web service
Webmasters can benefit from web archiving!
Archiving transactions towards an uninterruptible web service

Goal: create and leverage existing tools so when a web resource is unavailable due to some interruption of a key service an archived copy will be provided to an end user

• Web archives can serve as a value-added collection to motivate web archiving as a tool for day-to-day IT operation

For more information see: https://www.cs.vt.edu/node/7650
Incentive award winners: New curation tools for web archivists

https://www.youtube.com/watch?v=yeuk_vIOXc
w&list=PLf1Dab4lwQhBpFRB1dpUnKLgImM2iScjl
&index=6
Tools for managing seed URIs
Tools for managing seed URIs

- Results: developed tool to enable curators to evaluate and detect when their web archives are off topic or discover new seed sites to include in collections

- For more information see: https://github.com/yasmina85/offtopic-Detection
Visualizing Digital Collections of Web Archives
Visualizing Digital Collections of Web Archives

Results: developed tool for showing how a single web page changes over time

For more information see:
https://github.com/machawk1/ArchiveThumbnails,
http://thumbnails.cs.odu.edu:15421,
https://github.com/machawk1/ArchiveThumbnails/blob/master/CalendarResults.jsp
Exploring a national collaborative model for web archiving

https://www.youtube.com/watch?v=6f1Vj5cNAIU&index=7&list=PLf1Dab4lwQhBpFRB1dpUnKLglmM2iScj
Day 2: smaller group discussions

• Capture beyond traditional crawling
  – subtopics: transactional archiving; direct ingest of publisher files; web recording

• Collection development / Descriptive metadata & access
  – subtopics: collaborative collection development; institutional archives; metadata workflows; non-document collections

• Tools/APIs: integration into systems and standardization
  – subtopics: development of tools/APIs; integration into vendor and local systems; goal of network of standardized components

• Analysis of web archive datasets
  – subtopics: tools; researcher perspectives

• Preservation of scholarly & legal record / Long-term preservation
  – subtopics: preservation of scholarly record; reference rot; long-term preservation
It was a really busy and productive day but we only have one picture
Check out the slides and videos and let us know what you think!

- See the slides, watch the videos:
  https://library.columbia.edu/bts/web_resources_collection/Conferences/program.html

Feedback welcome!

Attendees of the conference are especially encouraged to keep us posted on resulting work.
Thanks!

Anna Perricci
Columbia University Libraries
anna.perricci@gmail.com
@AnnaPerricci

Unless otherwise noted, all images in this presentation are from the CUWARC Flickr album: https://flic.kr/s/aHskfjd54s