Boiling the Ocean, Together: Web Archive Collection Development in a Global Context

Nicholas Taylor
Web Archiving Service Manager
Digital Library Systems and Services

Chalk Talk
May 12, 2014
Distributed COLLECTION EFFORTS
by the numbers

- **70** web archiving initiatives on Wikipedia
- **313** Archive-It partners
- **33** CDL WAS subscribing institutions
broad but shallow
national domains

“Europe blank map.png” by Wiki-vr under public domain

“AustraliaNZ-blank.png” by Chug under CC BY-SA 3.0

Wikipedia: “List of Web archiving initiatives”
selective archiving

- Human Rights Web Archive (Columbia)
- CyberCemetery (North Texas)
- Elections Web Archives (Library of Congress)
- Labor and the Left (NYU Tamiment)
- Ukraine Conflict (Archive-It)
- NC State Government (NC Archives, Library)
- Michigan Historical Collections (UM Bentley)
- Health and Medicine Blogs (NLM)
how much archived?

- DMOZ: 79%
- Del.icio.us: 68%
- Bitly: 16%
- Yahoo!: 19%

WE ARE COLLECTING

What
topical collections

Middle East Politics

African Politics

Digital Games
government information
institutional legacy
How OTHERS COLLECT

“2009 san diego comic-con: comics, still an elemental part of the con” by george ruiz under CC BY 2.0
necessary but not sufficient

“In principle, the collection development policy for the Tamiment Library’s Web Archive parallels that of the Tamiment Library as a whole (labor and radicalism)”

In practice, this is complicated by (a) the enormous size and variety of born digital materials within Tamiment’s collecting scope...and (c) resource restraints. Thus the Library will not only have to carefully appraise materials, but to set priorities and limitations.”
necessary but not sufficient

- align with organizational mission
- support research and teaching
- preserve institutional legacy
- consider history and geography
sufficient-y

- collect within subject area
- focus on at-risk content
- collect content previously collected in print
- limit to particular types of organizations
sufficient?

- consider what others are collecting
- don't aim to be comprehensive (if you can’t be)
- complement existing strengths
- prefer current and/or unique content
- mind resource constraints
- anticipate value to researchers
- collect content, not links to content
- only collect publicly available content
- only target specific resource or format types
- enable designated research
COLLECTION CONSIDERATIONS
copyright and access policy
FERPA
social media
archiving proscriptions
offramps

“Pathfinder Panorama” by NASA Solar System Exploration under Public Domain
technical challenges
cost modeling
collection use cases

- outreach and education
- persistent citation
- documenting spontaneous events
- preserving citizen journalism
- saving at-risk content
- litigation risk mitigation
- capture related resources
- records management
research use cases

• how incumbent candidates talk to local constituencies
• inter-link graph of websites in different languages
• policies and practices of public health NGOs
• Honduran government websites after 2006 coup
• file format analysis for preservation planning
• prevalence of semantic markup
• most commonly-used JavaScript libraries
• digital archaeology of GeoCities
• resource persistence in Egypt Revolution social media

“Web Archiving Use Cases” by Emily Reynolds (2013)
Questions for DISCUSSION

“Question #2” by Robert S. Digby under CC BY-NC-ND 2.0
where do we start?

1. maintain awareness of topical web archives
2. promote and facilitate access to existing web archives
3. provide curatorial assistance for collaborative projects
4. enhance our collection dev policies for web archives
5. evaluate local/global gaps to build unique collections
how do we do it?

• how can we maintain awareness, facilitate discovery, and promote use of other web archives?
• what is relative importance of collection development policies, at-risk nature of the content, research use case tangibility, what others are collecting, etc.?
• how to maximize value of existing web archives and those we create, for Stanford and larger community?
• what are the key elements and optimal approach for creating our own collection development policies?
thank you!

Nicholas Taylor
ntay@stanford.edu