Considerations for Strategic Web Archive Collection Development

Nicholas Taylor
Web Archiving Service Manager
Stanford University Libraries

Curating Web Archives: Who Cares for Content?
May 23, 2014
web archiving lifecycle
## curator tools

<table>
<thead>
<tr>
<th></th>
<th>Appraisal and Selection</th>
<th>Scoping</th>
<th>Data Capture</th>
<th>Storage and Organization</th>
<th>QA and Analysis</th>
<th>Metadata / Description</th>
<th>Access / Use / Reuse</th>
<th>Preservation</th>
<th>Risk Management</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Archive-It</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AtN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BCWeb</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDL WAS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DigiBoard</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Islandora WARC Solution Pack</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Netarchive Suite</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PageFreezer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UNT Nomination Tool</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WCT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
appraisal and selection
we are few

- **70** web archiving initiatives on Wikipedia
- **313** Archive-It partners
- **33** CDL WAS subscribing institutions
how much archived?

- dmoz: 79%
- bitly: 16%
- del.icio.us: 68%
- Yahoo: 19%

selection determines preservation
subject expertise
traditional collecting
collecting compared

traditional
• published
• one-time, up-front curation
• rivalrous, usable by a local service population
• comprehensive
• many copies
• purchase/license
• finite acquisition

web archives
• public
• ongoing curation
• non-rivalrous, potentially usable by anyone
• representative
• few copies
• permissioned/sanctioned
• contingent acquisition
how others collect
necessary but not sufficient

- align with organizational mission
- support research and teaching
- preserve institutional legacy
- consider history and geography
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.”
what not to collect
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
- collect publicly available content
- anticipate value to researchers
- collect content, not links to content
- target specific resource or format types
- enable designated research
thank you!

Nicholas Taylor
ntay@stanford.edu

“stanford dish at sunset” by Dan under CC BY-NC-SA 2.0