One Site Among Many: Stanford and Collaborative Technical Development for Web Archiving

Nicholas Taylor
Web Archiving Service Manager
Stanford University Libraries

PASIG 2016
March 11, 2016
overview

- web archiving opportunity gaps
- situation of SUL web archiving
- APIs + community (technical) development

“LAX on take off” by Doug under CC BY-NC-ND 2.0
OPPORTUNITY GAPS

“Mind The Gap” by R-P-M under CC BY-NC-ND 2.0
link rot + content drift

Andrew Jackson: “Ten years of the UK Web Archive”
a centralized enterprise

a centralized enterprise

minimal local preservation

evolving web
opportunities for research
WHAT ARE WE DOING?
web archaeology (SLAC)

WorldWideWeb SLAC Home Page

SLAC 24 Dec 1993

Use the WorldWideWeb (WWW) service to gain access to a wide range of information at SLAC and elsewhere around the globe. Emphasized text like this is a hypertext link.

You may view WWW information through GUI or line-mode browsers. At least most SLAC pages have been tested on the Mosaic WWW X Window System browser. Note that over time links may move around on a page, migrate to others, or be removed entirely, due to the dynamic nature of the Web.

SLAC Information

People and organizations:
people at SLAC, particle physics people and institutions.

Library:
SPIRES-HEP, Current PPP List, Books, SLACweek glossary, other databases

Physics Preprint Bulletin Boards (full-text postscript):
today, yesterday, last seven days, week before that, anytime.

Seminars:
today, tomorrow, this week, next week, anytime.

Conferences:
this month, next month, next year, next summer, all future, let me search.

News:

oldweb.today: “WorldWideWeb SLAC Home Page”
building + integrating infrastructure

capture

access

preservation

discovery

SDR
APIS + COMMUNITY DEVELOPMENT
web archiving lifecycle
## functional overlap

<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</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>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>
smaller, modular components
community seed

National Digital Platform Projects funded in August 2015

Systems Interoperability and Collaborative Development for Web Archiving (LG-71-15-0174-15): The Internet Archive, working with partner organizations University of North Texas, Rutgers University, and Stanford University Library will undertake a two-year research project to explore techniques that can expand national web archiving capacity in several areas.
API candidates

- capture tool/proxy interconnect
- capture tool management
- data import/export
- query + extraction
- integrity audit + repair
- descriptive metadata

- logs + analytics
- renderings/derivative formats
- federated data delivery
- federated replay
- federated full-text search
let’s combine forces