Building Web Archiving Technology, Together

Nicholas Taylor
Web Archiving Service Manager
Stanford University Libraries

Web Archives 2015: Capture, Curate, Analyze
November 13, 2015
overview

• **why** build together?
• community for collaborative work
• **APIs** for collaborative work
WEIRD — MY CODE'S CRASHING WHEN GIVEN PRE-1970 DATES.

EPOCH FAIL!

not a programmer
aspiring OSS contributor

GitHub: “nullhandle (Nicholas Taylor)”
studying the landscape
a centralized enterprise

a centralized enterprise

minimal local preservation

evolving web
opportunities for preservation
opportunities for research
not the only one
a response

**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.
COMMUNITY

"Why we love Peckham, P1020468crop" by Eye magazine under CC BY-NC-SA 2.0
community analysis

SAA Web Archiving Roundtable

Archive-It Partners

IIPC

models of software production
(irrespective of license)

• **sole** source
  – *single developer*

• **closed** source
  – *team/corporate dev; no outside contributions*

• **club** source
  – *pool resources for solo/team/corporate dev*

• **community** source
  – *direct and distributed community participation*

• **open** source
  – *grassroots, democratic, meritocratic participation*

*Tom Cramer: “Collaborative Open Source Software Production & APIs”*
club source examples

• Archivematica, AtoM (Artefactual)
• ArchivesSpace (Lyrasis)
• Bitcurator (Educopia)
• Fedora (DuraSpace)
• JHOVE (OPF)
• LOCKSS (Stanford University)
• Omeka (George Mason University)
community source examples

In a Nutshell, Project Hydra...

... has had 11,717 commits made by 105 contributors representing 73,546 lines of code

... is mostly written in Ruby with a low number of source code comments

... has a well established, mature codebase maintained by a very large development team with stable Y-O-Y commits

... took an estimated 19 years of effort (COCOMO model) starting with its first commit in October, 2009 ending with its most recent commit 3 days ago

In a Nutshell, Blacklight...

... has had 2,887 commits made by 67 contributors representing 14,774 lines of code

... is mostly written in Ruby with an average number of source code comments

... has a well established, mature codebase maintained by a large development team with increasing Y-O-Y commits

... took an estimated 4 years of effort (COCOMO model) starting with its first commit in October, 2009 ending with its most recent commit 2 days ago
community architecture

- privileges **community** over code
- recognizes **distribution** of investment
- embraces community **diversity**
- models **open** processes and governance
- encourages **varied contributions**
- serves **community needs**
success of a standard

• **capture**: DeDuplicator, Heritrix, python-heritrix, SiteStory, WAIL, WARCreate, WarcMITMProxy, WarcProxy, Webrecorder, wget, Wpull

• **access**: OpenWayback, pywb, warc-proxy, WarcManager, Wayback Machine, Web Archive Discovery, WebArchivePlayer

• **utilities**: JHOVE2, JWAT, Megawarc, pylibwarc, WARCAT, Warcbase, warctools, Web Archive Commons
web archiving lifecycle
missed opportunities?

<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>
smaller, modular components
smaller projects do better

small projects (<$1 million)  
large projects (>$$10 million)  

Standish Group: “Chaos Manifesto 2013: Think Big, Act Small”
## IIPC community interest in APIs

<table>
<thead>
<tr>
<th>contribution type</th>
<th>% of respondents</th>
<th># of respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>help define functional requirements</td>
<td>94%</td>
<td>15</td>
</tr>
<tr>
<td>contribute use cases</td>
<td>81%</td>
<td>13</td>
</tr>
<tr>
<td>help define technical details</td>
<td>69%</td>
<td>11</td>
</tr>
<tr>
<td>help schedule and run meetings</td>
<td>19%</td>
<td>3</td>
</tr>
<tr>
<td>implement and test</td>
<td>6%</td>
<td>1</td>
</tr>
</tbody>
</table>

Andrea Goethals: “Results of the Web Archiving API Survey of IIPC Members”
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