12-3-2015

Discovery & Born-Digital Archiving: Open Source Systems for Preservation and Access

L. Bryan Cooper  
*Florida International University*

Margarita Perez-Martinez  
*Florida International University, perema@fiu.edu*

Follow this and additional works at: [https://digitalcommons.fiu.edu/glworks](https://digitalcommons.fiu.edu/glworks)

Part of the [Animal Sciences Commons](https://digitalcommons.fiu.edu/glworks), [Biochemistry, Biophysics, and Structural Biology Commons](https://digitalcommons.fiu.edu/glworks), [Biodiversity Commons](https://digitalcommons.fiu.edu/glworks), [Biology Commons](https://digitalcommons.fiu.edu/glworks), [Higher Education Commons](https://digitalcommons.fiu.edu/glworks), [Marine Biology Commons](https://digitalcommons.fiu.edu/glworks), [Microbiology Commons](https://digitalcommons.fiu.edu/glworks), [Plant Sciences Commons](https://digitalcommons.fiu.edu/glworks), and the [Terrestrial and Aquatic Ecology Commons](https://digitalcommons.fiu.edu/glworks)

**Recommended Citation**

[https://digitalcommons.fiu.edu/glworks/28](https://digitalcommons.fiu.edu/glworks/28)

This work is brought to you for free and open access by the FIU Libraries at FIU Digital Commons. It has been accepted for inclusion in Works of the FIU Libraries by an authorized administrator of FIU Digital Commons. For more information, please contact dcc@fiu.edu.
Live “sand-box” for play and continuing education; and to prototype systems addressing needs of Everglades Partners; e.g.

* Loss of material at local government agencies (SFWMD)
* Inability to expose collections to the larger world (National Park)
* Local MARC records not “set” with OCLC (MDPL)
* Need to get scientific data out of silos – federal mandate (LTER)
* Gov agency problems with CMS – search engine limitations
* Inability to find historical research near scientific government reports
RELEVANT TECHNOLOGY CHANGE

• OPAC & Technology Challenges – Google Search Ideal

• Federated Search Systems (Using APIs) to pull metadata from diverse silos Peaked 2009-2011)

• Internet Archive WARC files – ISO Standard in 2009

• Discovery Systems Ascendant – 2011
LIBRARIANS APPROACH TECHNICAL SERVICES, ASKING....

I found a report on website that is something the library needs to collect -- or used to collect and still should.

***

Saw a bibliographic record somewhere else, pointing to an electronic resource, and wanted to get that into our catalog or discovery system so FIU end-users could more easily locate.
OUR PROJECT IS RELATED TO:
NATIONAL DATA TRENDS -- LIBQUAL...MOST RECENT REPORT

• Survey data reveals **insatiable demand** for access to an exponentially growing field of information.

• “No library can ever have sufficient content that would come close to satisfying this appetite.”

• And “...our discovery tools are not quite maximizing the value libraries can deliver. There is a lot of room for improvement in this area.

  *Martha Kyrilldou -- LibQUAL*
OTHER REASONS? PROFESSIONAL PHILOSOPHY?

RANGANATHAN’S LAWS

• 1st Law – resources are for use – fostering access as prime directive

• 2nd and 3rd Laws – every user his/her document (focus on knowing your users); and every record or document its user (focus on item – and making it easier to connect it with the user wanting it.)
RANGANATHAN’S LAWS

• 4th Law – Save Time of the User

Join in one place what is in numerous different search systems, using the 3 search systems.

• And 5th Law – adapt to change – freeing up and syndicating metadata and data – preparing for enriching evolving discovery systems
When we started our project, there were silos where Everglades researchers would have to perform repetitive searches if performing a complete literature review.

- Aleph catalog
- FIU Digital Collections
- Hathi Trust
- FCE LTER site housing data sets and articles
- National Park Service website and hidden museum catalog
- SOFIA
- GIS data
- PALMM Everglades Collections
- Miami-Dade Public
- FIU/LTER partners like Odyssey Earth
- Local Publications (born-digital)
Linking Old Librarianship to New: Aligning 5-Steps of *The Innovator’s DNA* in Creating Thematic Discovery Systems for the Everglades

http://digitalcommons.fiu.edu/glworks/24/
DISCOVERY SYSTEM
# EVERGLADES EXPLORER METADATA

<table>
<thead>
<tr>
<th>Partner Institution</th>
<th>Num. Rec.</th>
<th>Metadata Schema</th>
</tr>
</thead>
<tbody>
<tr>
<td>Everglades National Park</td>
<td>5</td>
<td>MARC</td>
</tr>
<tr>
<td>HathiTrust</td>
<td>85</td>
<td>MARC</td>
</tr>
<tr>
<td>SUS Libraries Integrated Library System (ILS) - Aleph</td>
<td>402</td>
<td>MARC</td>
</tr>
<tr>
<td>Miami Dade Public Library</td>
<td>164</td>
<td>MARC</td>
</tr>
<tr>
<td>FIU GIS Center</td>
<td>12</td>
<td>EML</td>
</tr>
<tr>
<td>Florida Coastal Everglades Long Term Ecological Research (FCE LTER)</td>
<td>125</td>
<td>EML</td>
</tr>
<tr>
<td>South Florida Information Access (SOFIA)</td>
<td>250</td>
<td>FGDC</td>
</tr>
<tr>
<td>Publication of Archival Library &amp; Museum Materials (PALMM) digital collections</td>
<td>3,450</td>
<td>Dublin Core</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>4,493</strong></td>
<td><strong>XC</strong></td>
</tr>
</tbody>
</table>
Search records

river of grass

1. The Everglades: river of grass
   format: Films, videos — 1 streaming video file (approximately 5 min.)
   location: http://www.odysseyearth.com...
   abstract: This short film by nature documentarian Richard C. Kern and his son Richard S. Kern briefly describes the Everglades and how water flow makes it a gigantic, slow-moving river.

2. A Seminole Indian family at home in the Everglades of Florida — Beautiful Florida Series — The Everglades: river of grass
   by C.T. American Art
   format: Films, videos — 1 streaming video file (circa 5 min.)
   location: http://www.odysseyearth.com...
   date: @2012.

   by Douglas, Marjory Stoneman — Reclaiming the Everglades.
   location: http://purl.fcla.edu...
   http://purl.fcla.edu...
   http://digitool.fcla.edu:80...
   date: 1947.
CMS SEARCH
ACROSS SELECT AGENCY AND PARTNER SITES USING GOOGLE API

- South Florida Information Access (SOFIA)
- South Florida Water Management District (SFWMD)
- Southeast Environmental Research Center (SERC) Water Quality Monitoring Network
- The Everglades Foundation Channel
- Everglades Digital Library
- Everglades Foundation
- Everglades NPS's channel
- Eyes on the Rise
- Florida Coastal Everglades (FCE) LTERNET
- Florida Keys National Marine Sanctuary
- Library of Congress Reclaiming the Everglades
- National Park Service
- Odyssey Earth
- SFWMD America's Everglades project site
WHAT HAVE WE LEARNED?
METADATA CHALLENGE

Everglades Explorer

Extensible Stylesheet Language Transformations (XSLT)

Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)
METADATA TRANSFORMATION

Extensible Stylesheet Language Transformations (XSLT)

Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)

Low-barrier mechanism for repository interoperability
HARVESTING OVERVIEW
Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)
February 2015, Library and Information Technology Association (LITA) panelists say Top Technology Trends include enhancing **discoverability** (Enis, 2015)

Making content accessible **where the search originates** (e.g. Google, Google Scholar, WorldCat, DPLA, Europeana) creates value for digital libraries and users (Enis, 2015)

Repositories contributing to aggregators can experience **increased site visits from 55-109 per cent** (DPLA, n.d)

[ee.fiu.edu](http://ee.fiu.edu)
METADATA AGGREGATOR
COOL PROJECT...
A portal that delivers students, teachers, scholars, and the public to incredible resources, wherever they may be in America.

A platform that enables new and transformative uses of our digitized cultural heritage.

An advocate for a strong public option in the twenty-first century. DPLA seeks to multiply openly accessible materials to strengthen the public option that libraries represent in their communities.
HARVESTING ROLES

Service Providers or Service Hubs then make OAI-PMH service request to harvest that metadata (State or regional projects)

Data Providers or Content Hubs are repositories that expose structure metadata via OAI-PMH (large institutions)
DPLA Hubs locations

- Service Hubs
- Service Hub partner states
- Content Hubs
- Hubs in active development

State or regional projects)
Large institutions
INTERNET ARCHIVE

- Founded by Brewster Kahle in 1996
- San Fran based digital library with mission: Universal Access to All Knowledge.”
- Ambitious & Very Progressive – an activist organization
- Wayback Machine – 150 billion web captures
- Archive-It web-crawling & hosting service

ee.fiu.edu
ARCHIVE-IT / MILESTONES

• 2002 – open source web crawler Heratrix released
• 2006 – Archive-It launched in 2006
• 2009 ISO standard achieved
• 2013 -- 13 initial partners grows to 238 – Best Practices for Web Archiving published
ARCHIVE-IT.ORG

FEE BASED HOSTING SYSTEM FOR OPEN SOURCE CRAWLS (HERATRIX)

• ARCHIVE-IT.ORG TODAY – 425 Thematic Collections from over 400 Institutions:

  • Colleges & Universities; Law Libraries; Museuems & Art Libraries; National Institutions; NGOs, Local Government, State Archives & Public Libraries.

• LOC Program, K-12 Web Archiving Program
WHY ARCHIVE THE INTERNET?

• Ephemeral coverage of events – Occupy Movement (Occupy Wall Street)

• Earthquake in Japan (Virginia Tech)

• Warfare; Bad Guys – Passenger jet shot down over the Ukraine

• Politics --SFWMD – Ephemeral government documents
FIU — ARCHIVET-IT
EVERGLADES EXPLORER
A WEB ARCHIVAL SEARCH

• Small grant received to build collection / test system ($3,000)

• 125 GB (1/8 TB; plus 3,000,000 URLs (documents).
TESTING ACQUISITION OF 3 DIFFERENT FORMATS

• Currently archiving 15 GB of Data; 242 active “seeds”
• 11,574 documents crawled, including:
  • HTML
  • Jpeg or other image files for maps
  • Video
  • PDFs
SO FAR, OUR LARGEST SINGLE HARVEST NETTED 6,492 PDF DOCUMENTS

- Test harvests included select partner documents
  - FCE LTER
  - National Park
  - SOFIA
  - SFWMD
  - Etc.
BESIDES EE.FIU.EDU – ARCHIVAL SEARCH BOX, CAN USE … HTTPS://ARCHIVE-IT.ORG/ KW EVERGLADES
Florida Coastal Everglades
Long Term Ecological Research

News and Announcements

Follow @fcelter
Find us on Facebook

Newsletters

FCE Newsletter Spring 2015 (PDF, 0.8 MB)
FCE Newsletter Winter 2015 (PDF, 1.7 MB)
FCE Newsletter Fall 2014 (PDF, 3.6 MB)
FCE Newsletter Summer 2014 (PDF, 1 MB)
8 ORGANIZATIONS HAVE COLLECTIONS RELATED TO CLIMATE CHANGE
SELECT LEARNING CURVES....

• Running test-crawls – reviewing reports.

• Click-through and visually inspect archived seeds/documents

• Finding links that resulted in no material acquired

• Page, “News or RSS Feed” and larger URL crawls (folder level; sub-domain; domain).
ESTABLISHING **BEST PRACTICES**
CURATION; TECHNICAL / SYSTEMS BACK-END

• Appraisal & Selection
• Scoping – portions of sites; whole sites
• Data Capture parameters (crawl frequency; types of files (PDF only?))
  • **27% of Archive-It partners harvest only PDFs**
• Quality Assurance; Clean-up/Removal (No Single Best-Practice).
• Storage & Syndication (Metadata; Data)
WHAT ELSE WAS LEARNED....

• Easy to catalog using DC
• DC most efficient for collection/seed level harvests — docs can rely on OCR indexing)
• Test crawls important; render and analyze reports
• Effective video harvest, but at institutional flat page level (Not good w/YouTube)
• One-time vs periodic harvests might be best
• Harvester rights & abilities – “Archive-This” plugin for Mozilla
• Can assign contributor rights or (metadata development).
EXTERNAL METADATA LOAD (SINGLE OR BULK)

• For set of older documents (ERIC from 60s and 70s)

• Located OCLC records in Worldcat or other OCLC index / single or bulk export to Refworks

• Use Refworks to cross-walk to ODS – Open Document Spreadsheet

• Import into Excel / Check; edit fields

• Import into Archive-It
OTHER UNIVERSITIES – COLUMBIA U.

• **Captures university web domains** – focus of University Archives

• **Align Archive-It to existing areas**
  • Collection Development
  • Institutional Academic Specialties /Focus

• **Thematic topics** – global human rights; historic preservation in NYC; NYC religious institutions.
UNIVERSITY OF ALBERTA

- Prairie Politics and Economics
- Government Documents
- Grey Literature – Business & Health Sciences
- Circumpolar Studies
- Education Curriculum Materials
FOCUSED ON DISCOVERY AND FRONT-END DEVELOPMENT

- Merging web archive with Digital Library Systems.
• Focus on State Documents – often only found online
• Websites of State Agencies
• Counter-balance to decline in submission of print publications
STAFF ALLOCATIONS -- CAN MANAGE W/MINIMAL STAFF -- YET SCALABLE

• **Columbia U.** – 1 staff + some web programming and students for metadata curation.

• **Creighton U.** – 1 Full-time archivist administers

• **University of Alberta** – 1 admin; 40 people actively contributing

• **Montana State Library** – 3 (State Pub Librarian; Metadata Cataloger; Systems/Analyst.

• **NC State Archives** – 4 (2 archives; 2 state library)

• -- *The Web Archiving Life Cycle Model*’ – The Archive-It Team
TYPE OF STAFF AT AN INSTITUTION WORKING W/ARCHIVE-IT – SURVEY

• Archives – 64%
• Library Staff – 42%
• Digital Projects Staff – 30%
• IT Staff – 8%
• Other – Student Workers; Web Development, etc. – 8%

• Source: Sweeter 2011
PARTNER SURVEY REVEALS DESIRES FOR FUTURE

• **56%** want to eventually archive in their own digital library

• **31%** want to allow Archive-It to handle.

• **Others…**
  
  • Incorporate **WARC files into digital repositories**
  
  • Provide **multiple access points** – put metadata & copies of data in multiple gateways (Syndicate/Discovery)

---

*The Web Archiving Life Cycle Model” – The Archive-It Team*
METADATA

• 90% of partners generate collection level metadata.
• 60% generate “seed” level metadata
• 15% generate document level metadata (full-text is fully indexed after 7 days of crawl).
• Currently, only a small minority are cataloging what they harvest in their existing catalog.

— The Web Archiving Life Cycle Model” – The Archive-It Team
COPYRIGHT

• Robots.txt file
• DMCA – takedown provision
• Archive-It will honor requests to remove content
• Most adhere to a library’s federal right to capture under preservation and Fair Use concepts/law.
• If capturing publications funded by tax dollars, additional protection
• Columbia U. 5 out of 783 sites they harvest objected -- .6%
IF ANYONE IS INTERESTED IN USING ARCHIVE-IT IN YOUR RESPECTIVE AREAS....

• I’m available to answer questions, or make department specific presentations.

• In process of working with Archive-It on pricing structure for FIU; should you have interest -- $15,000 will allow harvesting up to 18 million URLs and 1.5TB.

• If Archives, Govdocs, Subject Liaisons, or other units are interested, I would be happy to serve as a resource on the technical function of Archive-It & what we have learned to date.
WE ARE LOOKING FOR VOLUNTEERS
EVERGLADES MATERIAL CURATION / PRESERVATION

• **Google Search:** Help us curate and add sub-domains and directory level URLs to Google Search

• **Archive-It:** Websites, or individual documents/pdfs that should be harvested and archived with metadata.

• **Discovery:** Sets of metadata that should be pulled from a silo – large or small – for aggregation in a thematic gateway.
QUESTIONS?

HTTP://EE.FIU.EDU/
REFERENCES


- “The Web Archiving Life Cycle Model” – The Archive-It Team; Internet Archive, March 2013 Principal Authors Molly Bragg and Kristine Hana; Contributors: Lori Donovan; Graham Hukill; Anna Peterson.