Florida International University
FIU Digital Commons

Works of the FIU Libraries | FIU Libraries

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

 Part of the Animal Sciences Commons, Biochemistry, Biophysics, and Structural Biology Commons, Biodiversity Commons, Biology Commons, Higher Education Commons, Marine Biology Commons, Microbiology Commons, Plant Sciences Commons, and the Terrestrial and Aquatic Ecology Commons

Recommended Citation
Cooper, L. Bryan and Perez-Martinez, Margarita, "Discovery & Born-Digital Archiving: Open Source Systems for Preservation and Access" (2015). *Works of the FIU Libraries*. 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.

DISCOVERY & BORN-DIGITAL ARCHIVING: OPEN SOURCE SYSTEMS FOR PRESERVATION AND ACCESS

EVERGLADES EXPLORER

EE.FIU.EDU

L. BRYAN COOPER; MARGARITA PEREZ MARTINEZ

DECEMBER 3, 2015

EVERGLADES EXPLORER – EE.FIU.EDU

1 DISCOVERY SYSTEM; 1 SEARCH ENGINE; 1 INTERNET ARCHIVE; WHY?

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.

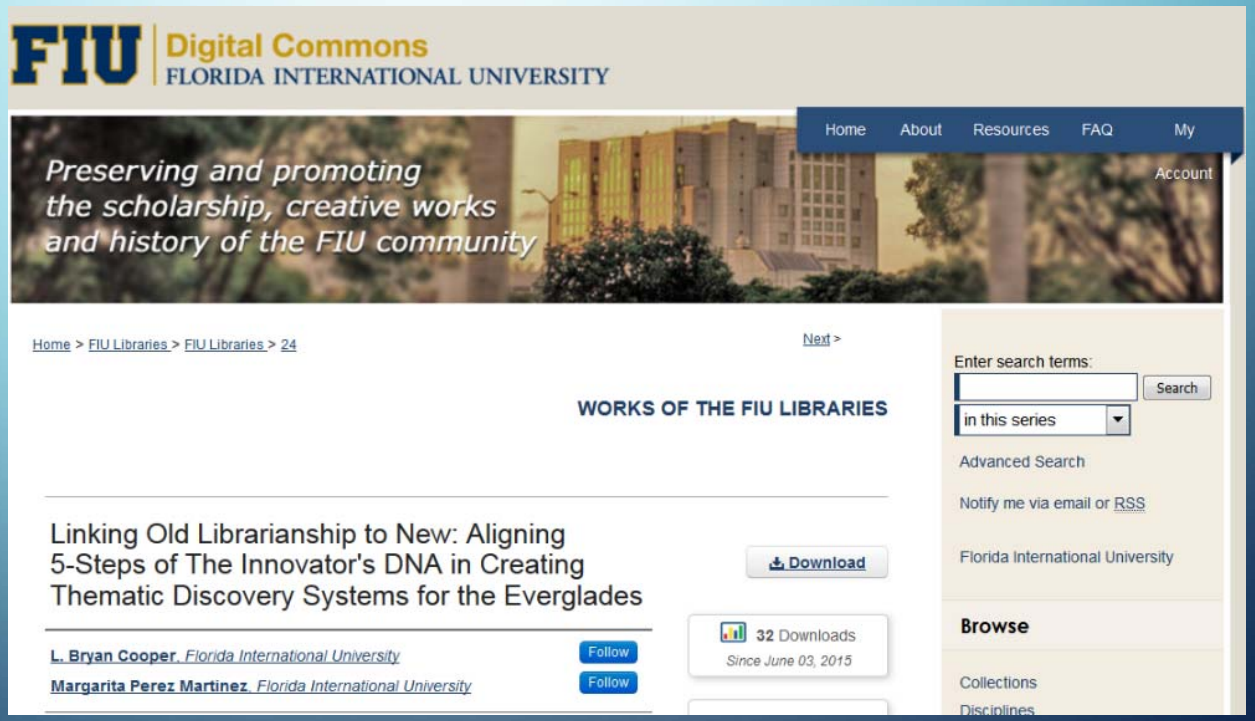
- **And 5th Law – adapt to change – freeing up and syndicating metadata and data – preparing for enriching evolving discovery systems**

SILOS -- FIU LIBRARIES & BEYOND

- 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)

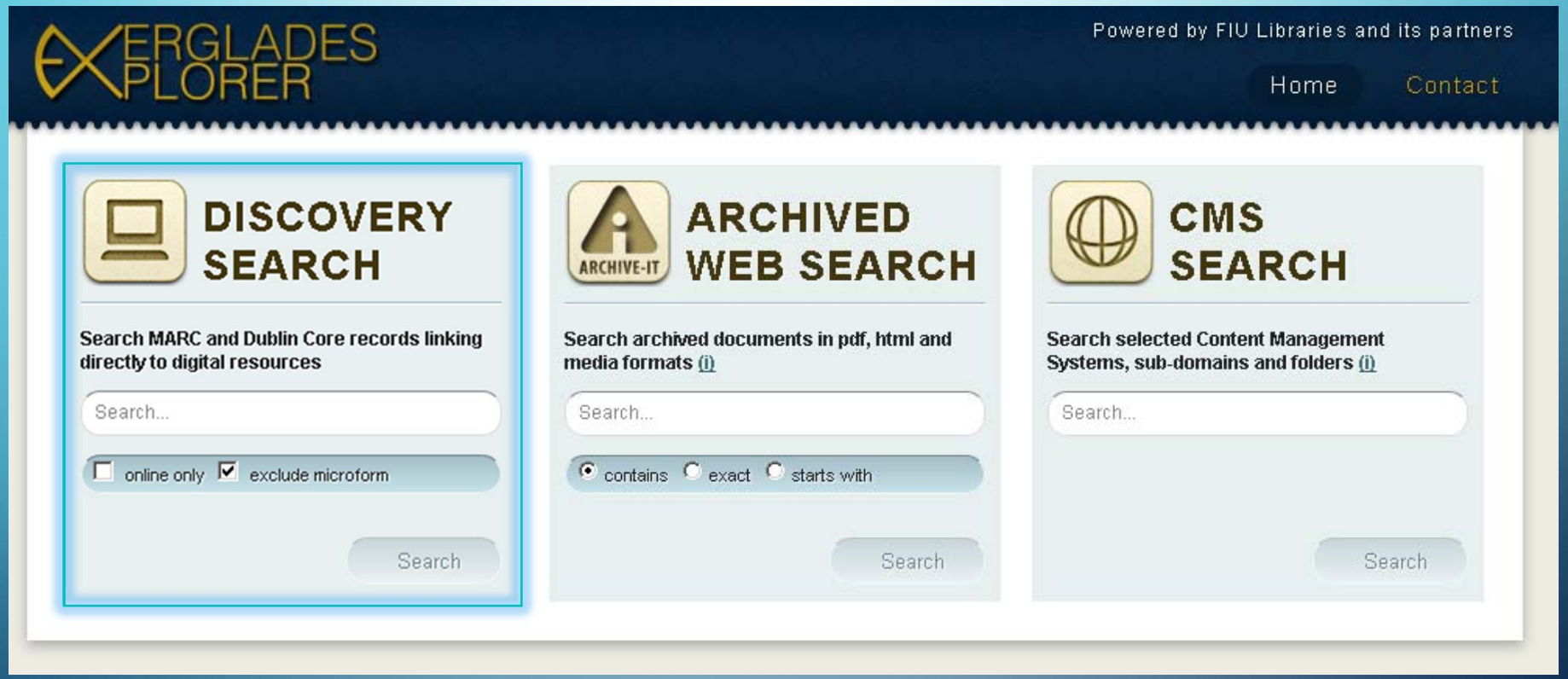
FLA 2015 ANNUAL CONFERENCE POSTER SESSION

- **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/

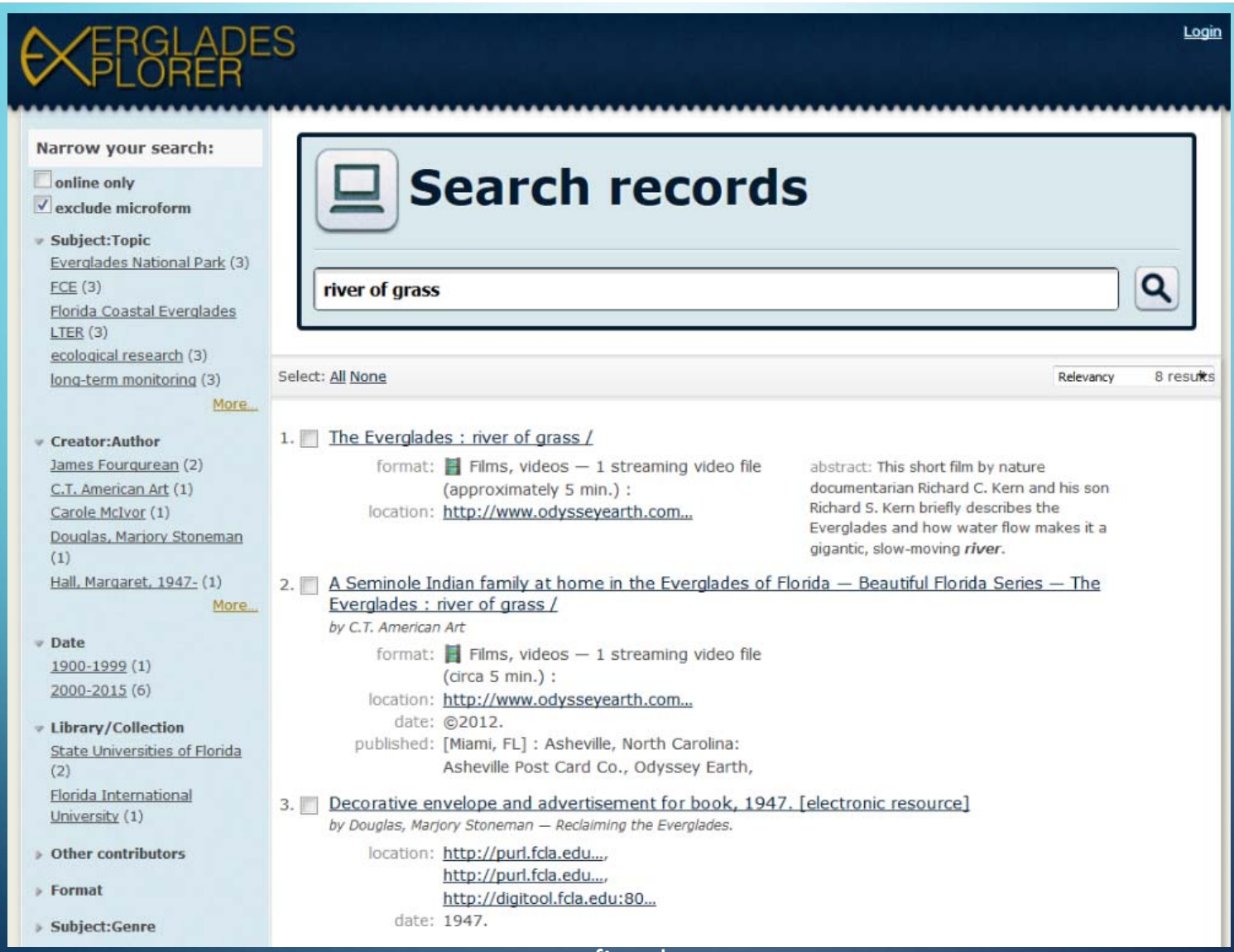
DISCOVERY SYSTEM

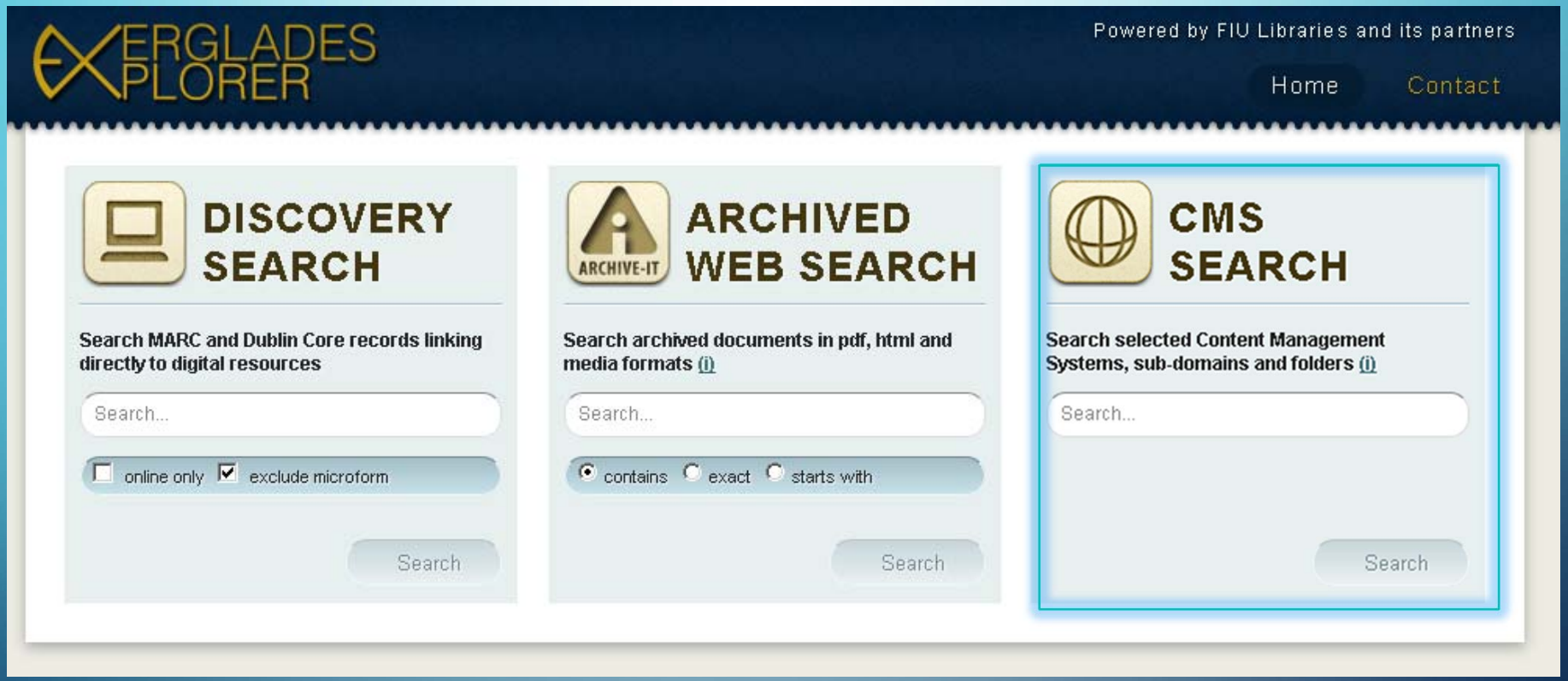


HTTP://EE.FIU.EDU/

EVERGLADES EXPLORER METADATA

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


EVERGLADES EXPLORER
Login
Narrow your search:
online only
exclude microform
Subject:Topic
Everglades National Park (3)
FCE (3)
Florida Coastal Everglades LTER (3)
ecological research (3)
long-term monitoring (3)
More...
Creator:Author
James Fourqurean (2)
C.T. American Art (1)
Carole McIvor (1)
Douglas, Marjory Stoneman (1)
Hall, Margaret, 1947- (1)
More...
Date
1900-1999 (1)
2000-2015 (6)
Library/Collection
State Universities of Florida (2)
Florida International University (1)
Other contributors
Format
Subject:Genre
Search records
river of grass
Select: All None
Relevancy
8 results
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.
published: [Miami, FL] : Asheville, North Carolina: Asheville Post Card Co., Odyssey Earth,
3. Decorative envelope and advertisement for book, 1947. [electronic resource]
by Douglas, Marjory Stoneman — Reclaiming the Everglades.
location: http://purl.fcla.edu...,
http://purl.fcla.edu...,
http://digitool.fcla.edu:80...
date: 1947.

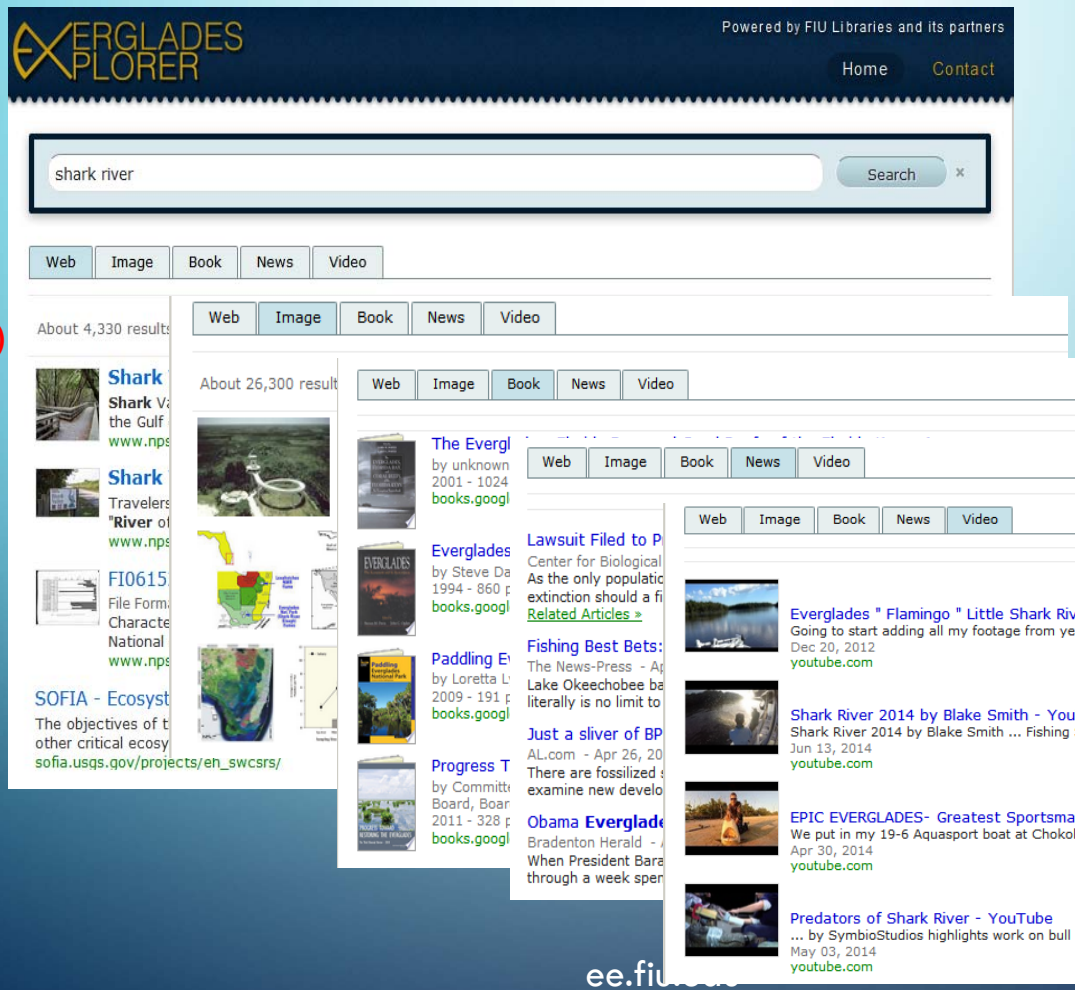


HTTP://EE.FIU.EDU/

CMS SEARCH

ACROSS SELECT AGENCY AND PARTNER SITES USING GOOGLE API

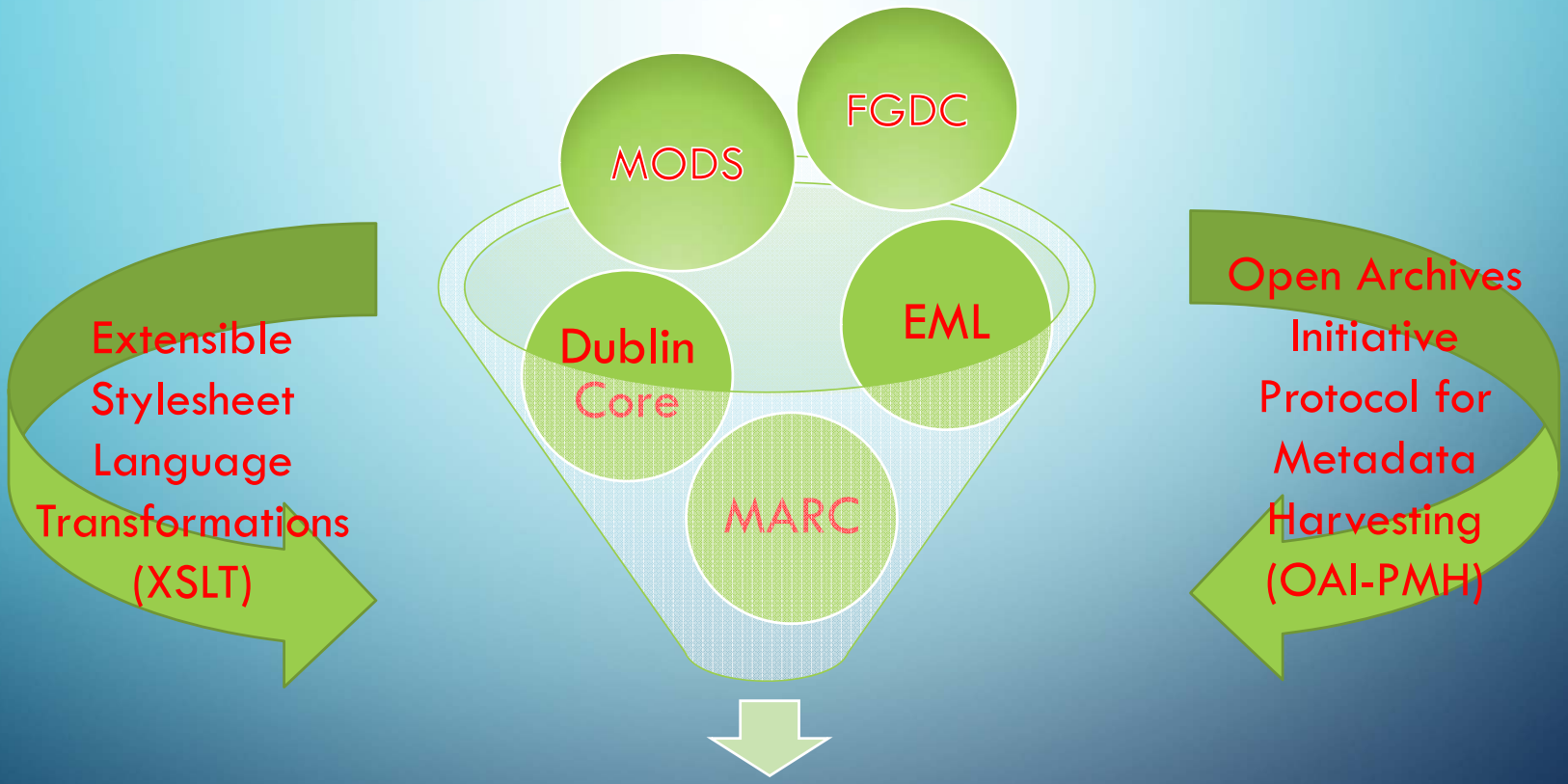
- South Florida Information Access (SOFIA)
- South Florida Water Management District (SFWMD)
- Southeast Environmental Research Center (SERC) Water Quality Monitoring Network
- The Everglades Foundation Channel



- 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

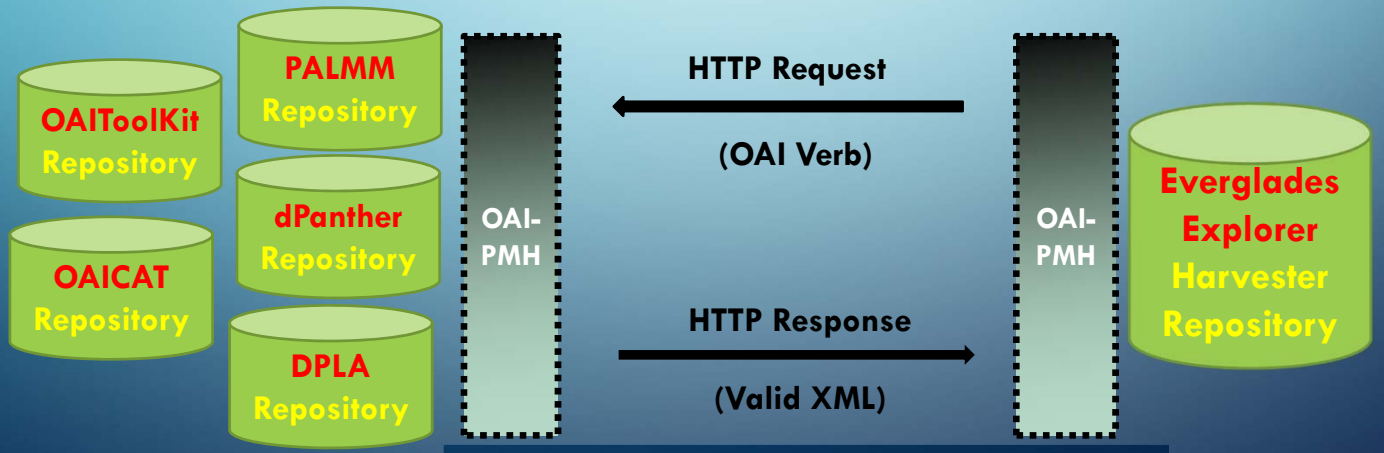
ee.fiu.edu

EVERGLADES EXPLORER

METADATA TRANSFORMATION

Extensible Stylesheet Language Transformations (XSLT)

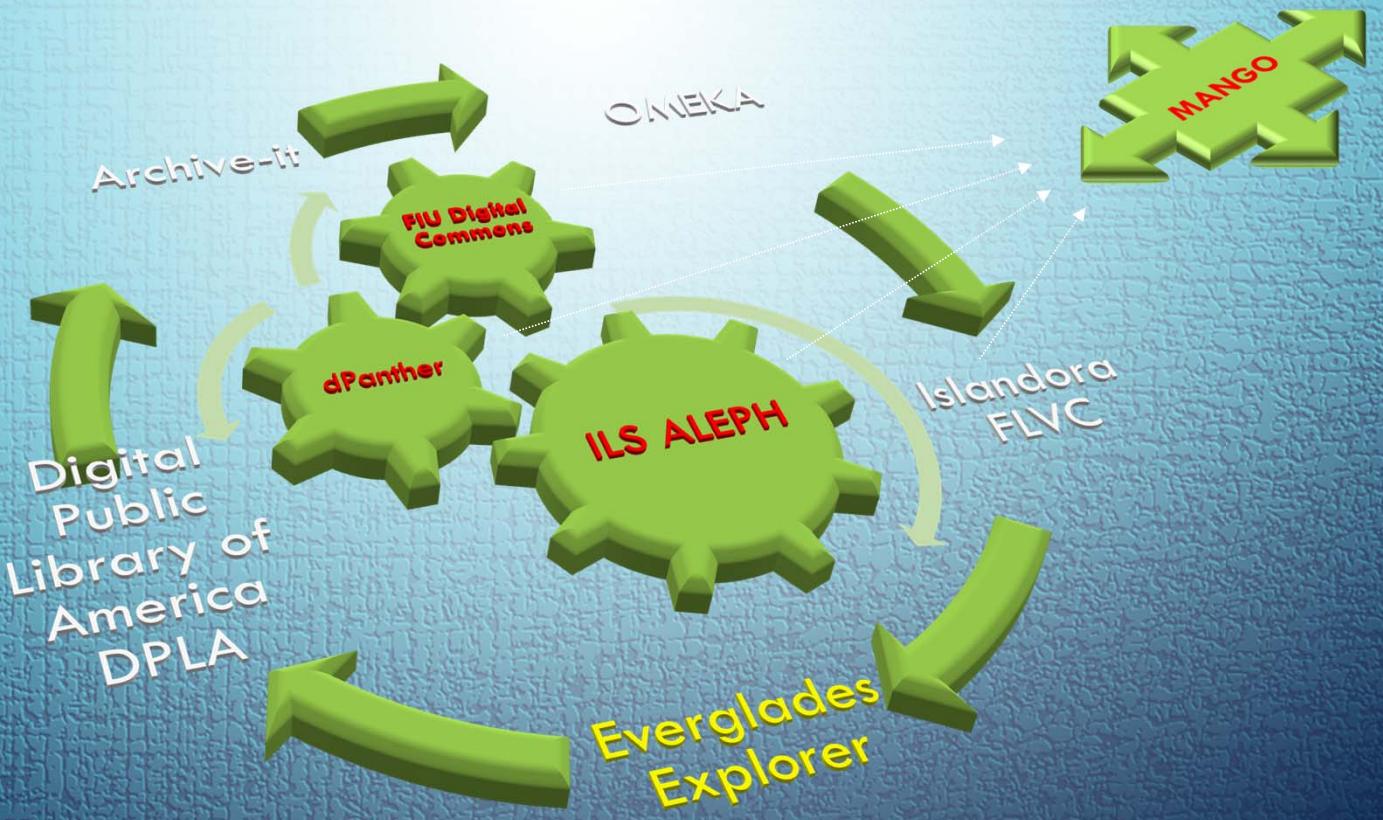
Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)



Low-barrier mechanism for repository interoperability

HARVESTING OVERVIEW

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



HARVESTING IMPACTS

February 2015, Library and Information Technology Association (LITA) panelists say Top Technology Trends include enhancing ***discoverability*** (Enis, 2015)

Making content accessible ***where the seach 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)*

METADATA AGGREGATOR COOL PROJECT…

DIGITAL PUBLIC LIBRARY OF AMERICA (DPLA)

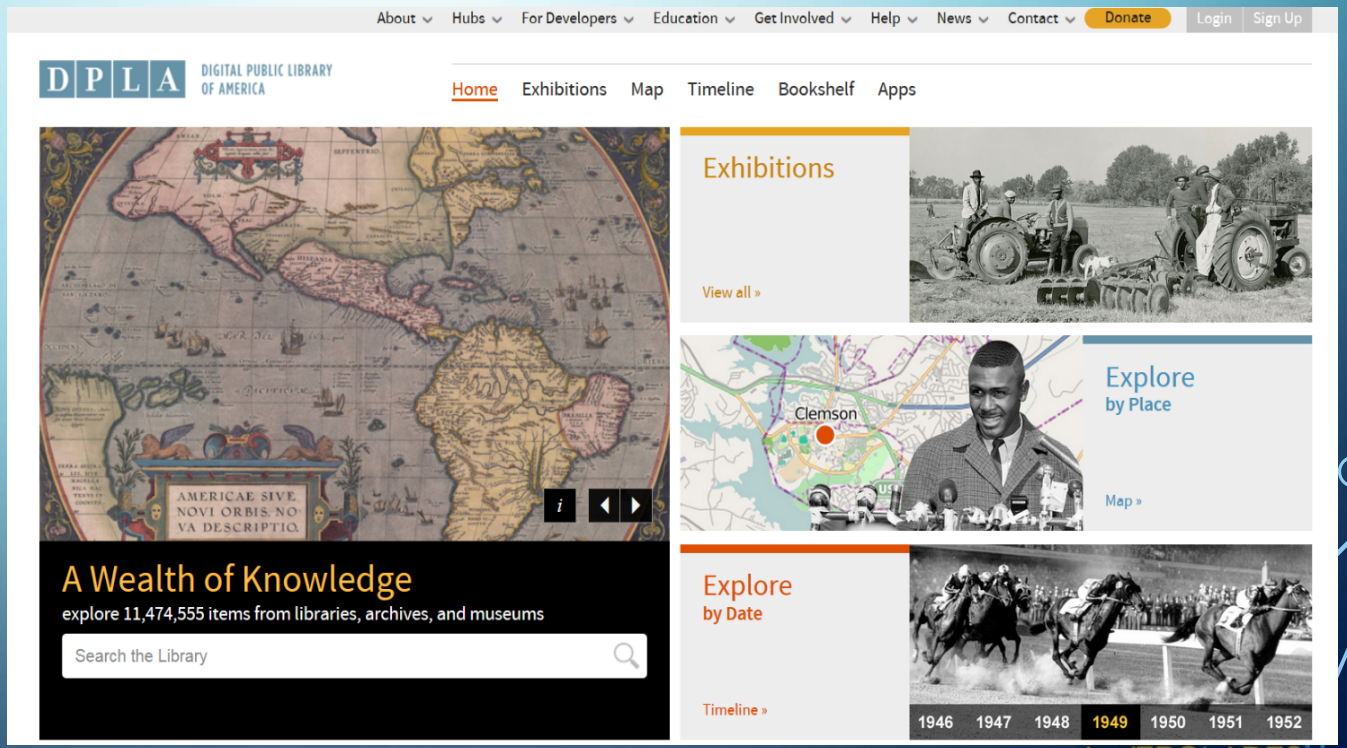
HTTP://DP.LA/

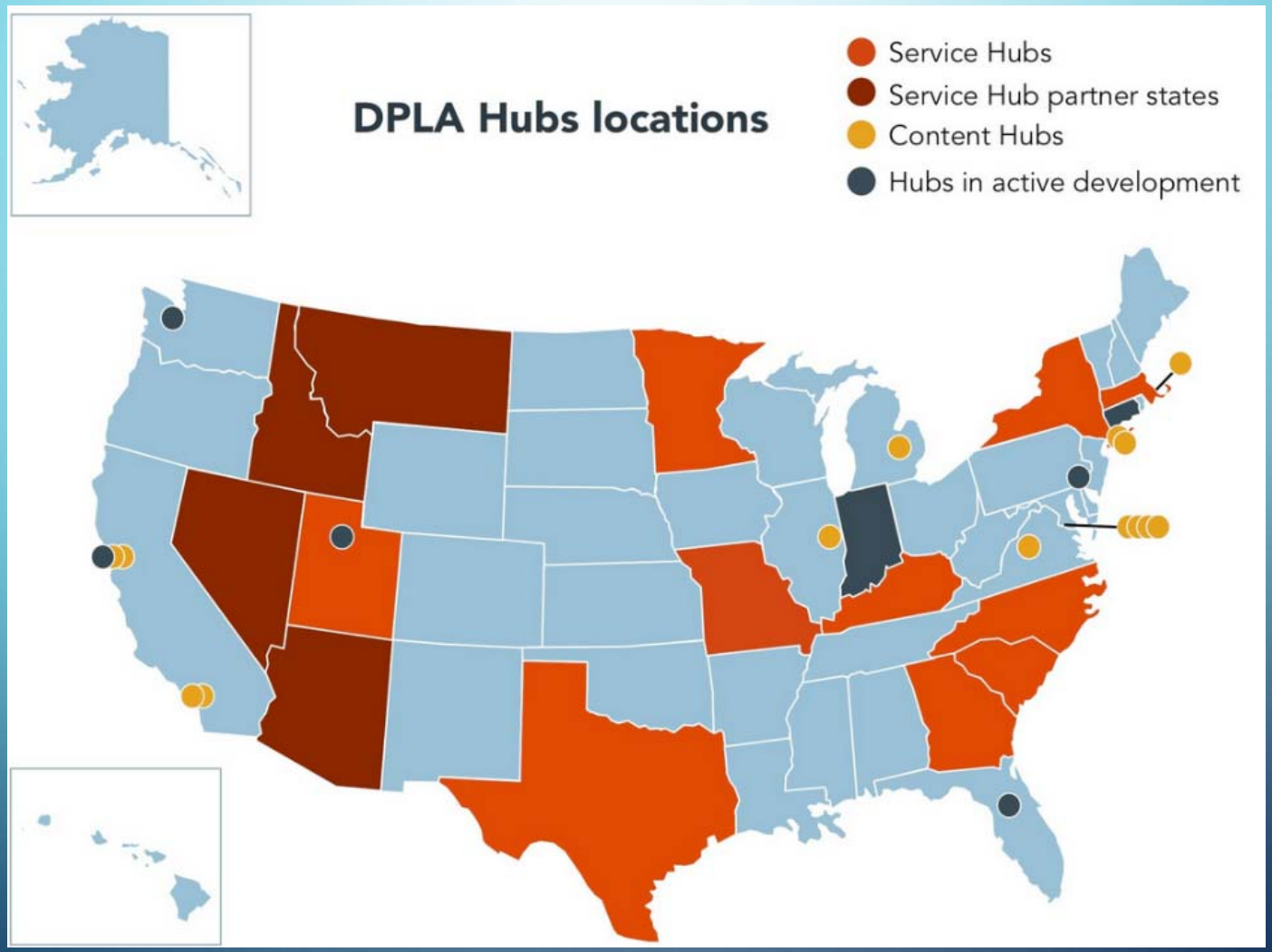
A *portal* that delivers students, teachers, scholars, and the public to incredible resources, wherever they may be in America.

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.





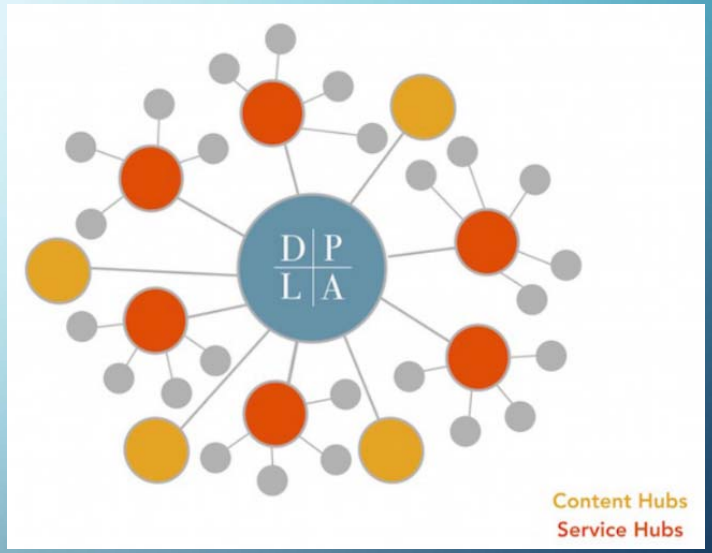
https://digitalpubliclibraryofamerica.atlassian.net/wiki/download/attachments/524354/DPLA-states-map-20141007.jpg?version=1&modificationDate=1415646248263&api=v2/

HARVESTING ROLES

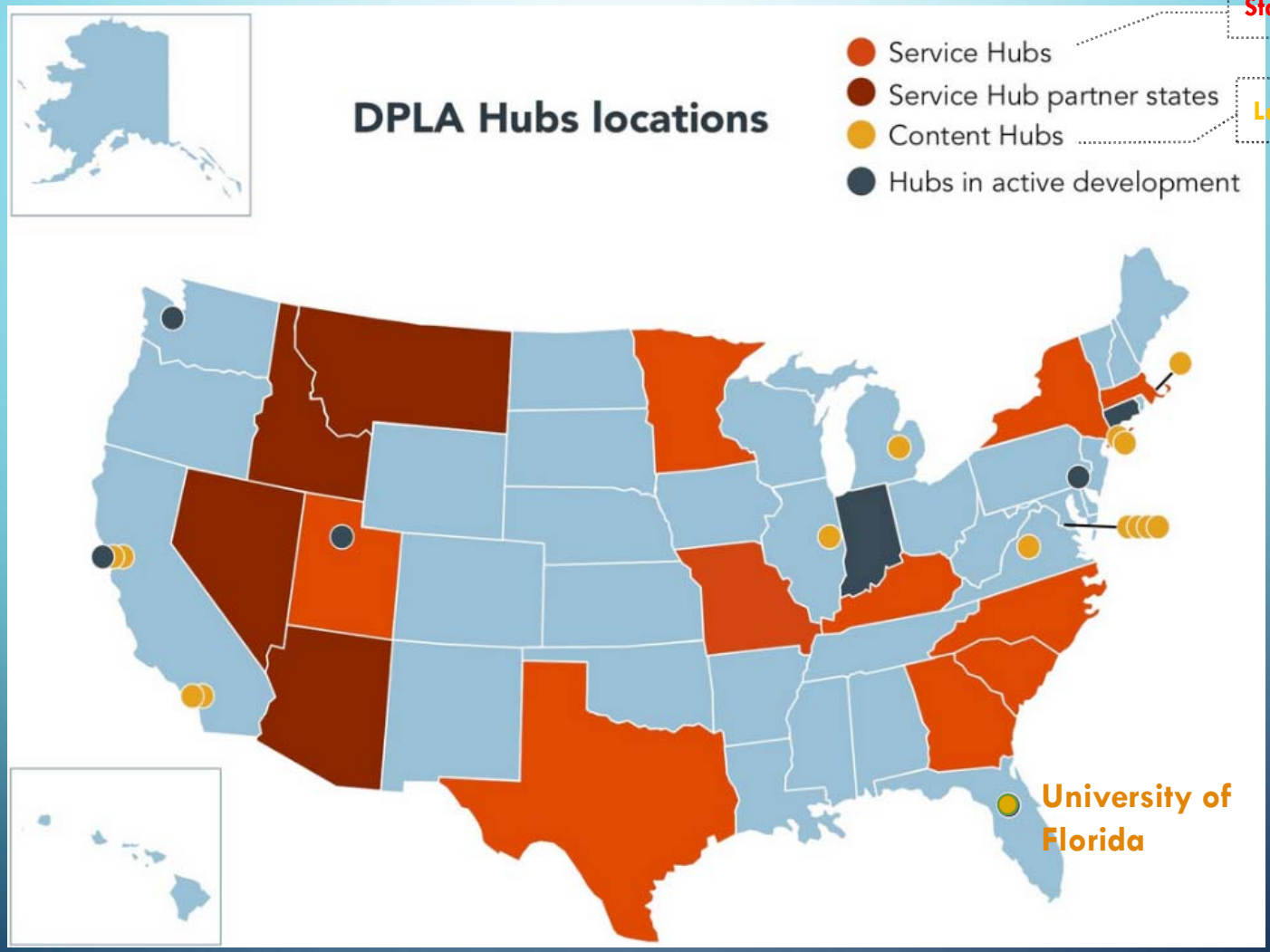
Service Providers or Service Hubs **then make OAI-PMH service request to harvest that metadata (State or regional projects)**

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

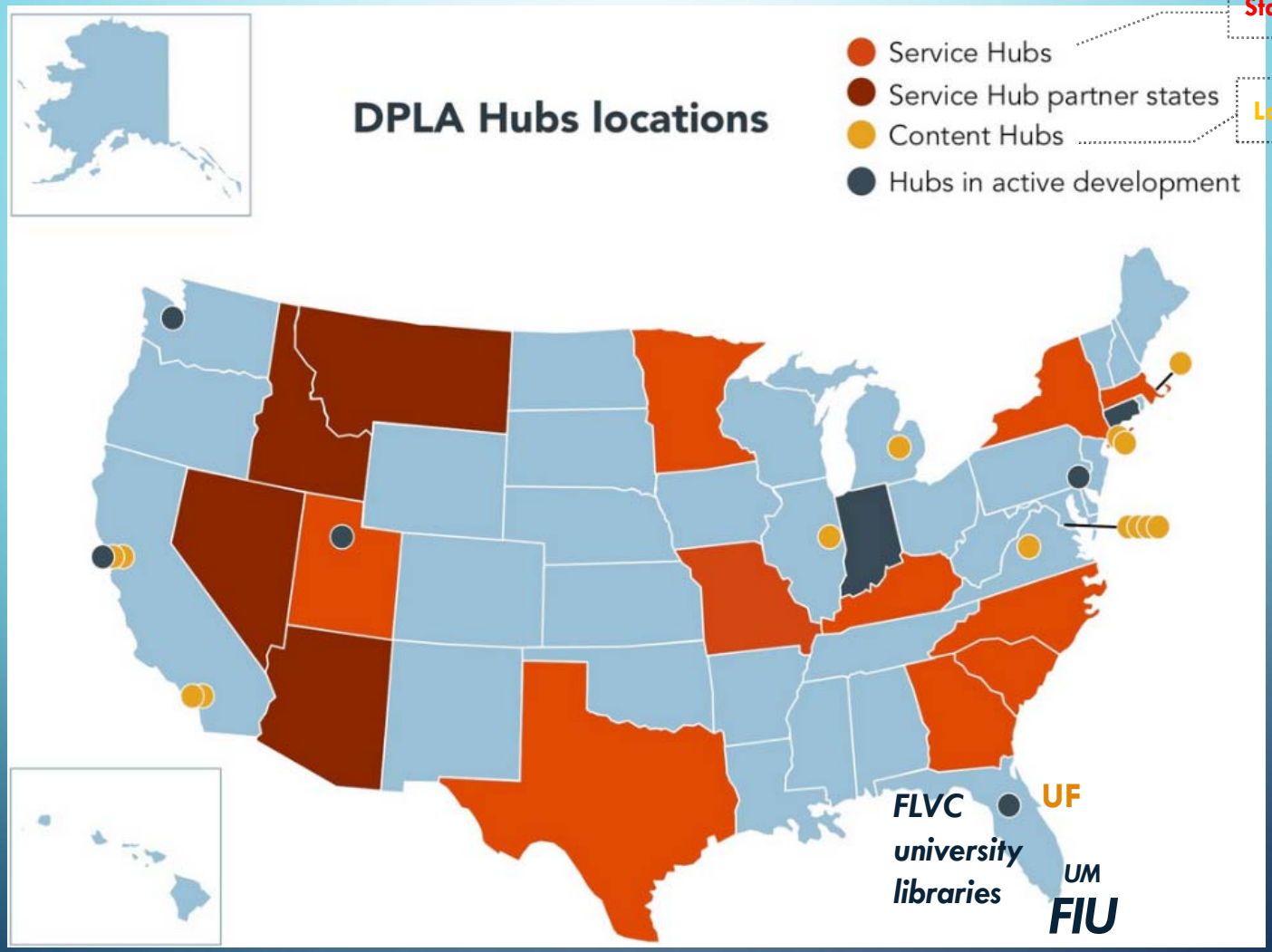
DPLA DIGITAL PUBLIC LIBRARY OF AMERICA



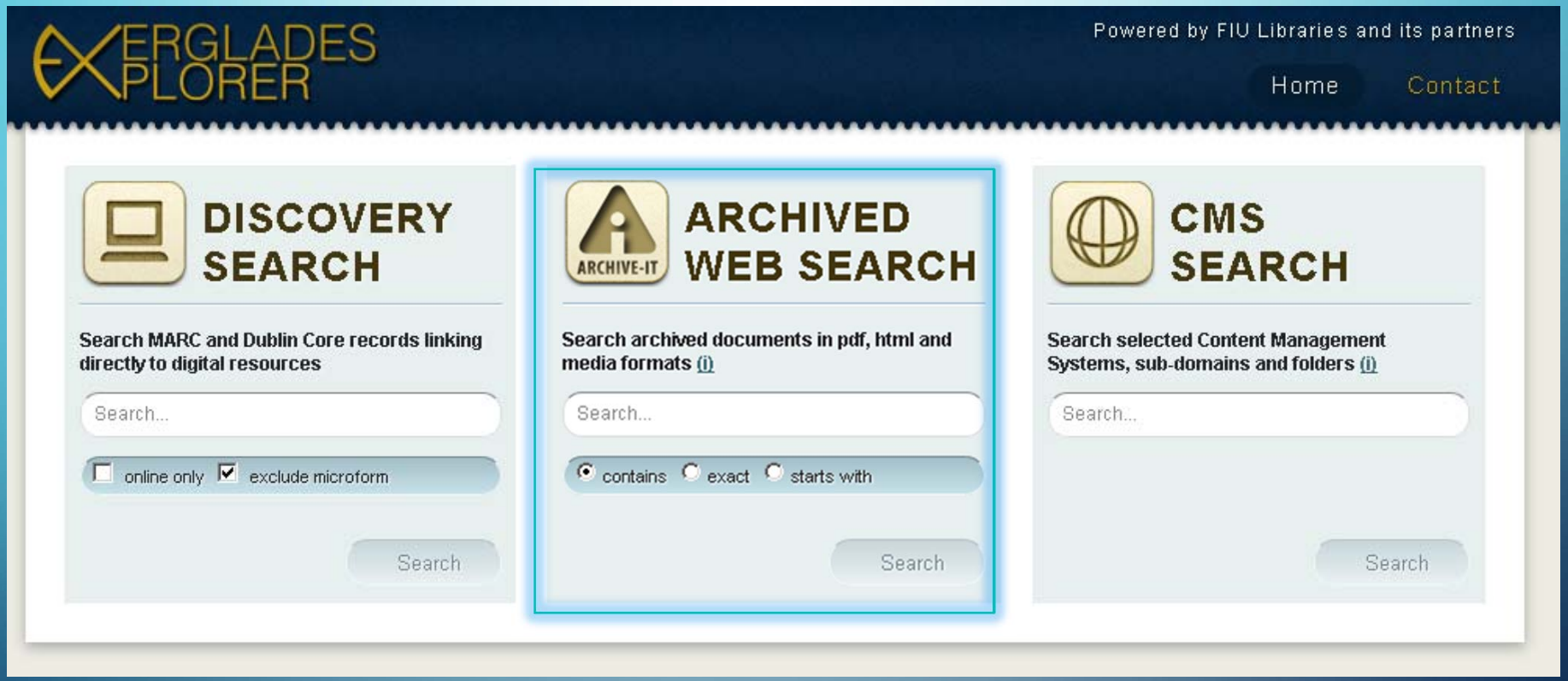
Retrieved from http://dp.la/info/hubs/



https://digitalpubliclibraryofamerica.atlassian.net/wiki/download/attachments/524354/DPLA-states-map-20141007.jpg?version=1&modificationDate=1415646248263&api=v2/



https://digitalpubliclibraryofamerica.atlassian.net/wiki/download/attachments/524354/DPLA-states-map-20141007.jpg?version=1&modificationDate=1415646248263&api=v2/



HTTP://EE.FIU.EDU/

ARCHIVE-IT

OPEN –SOURCE SOFTWARE & HOSTING PROVIDED BY INTERNET ARCHIVE (BREWSTER KAHLE)

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

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).

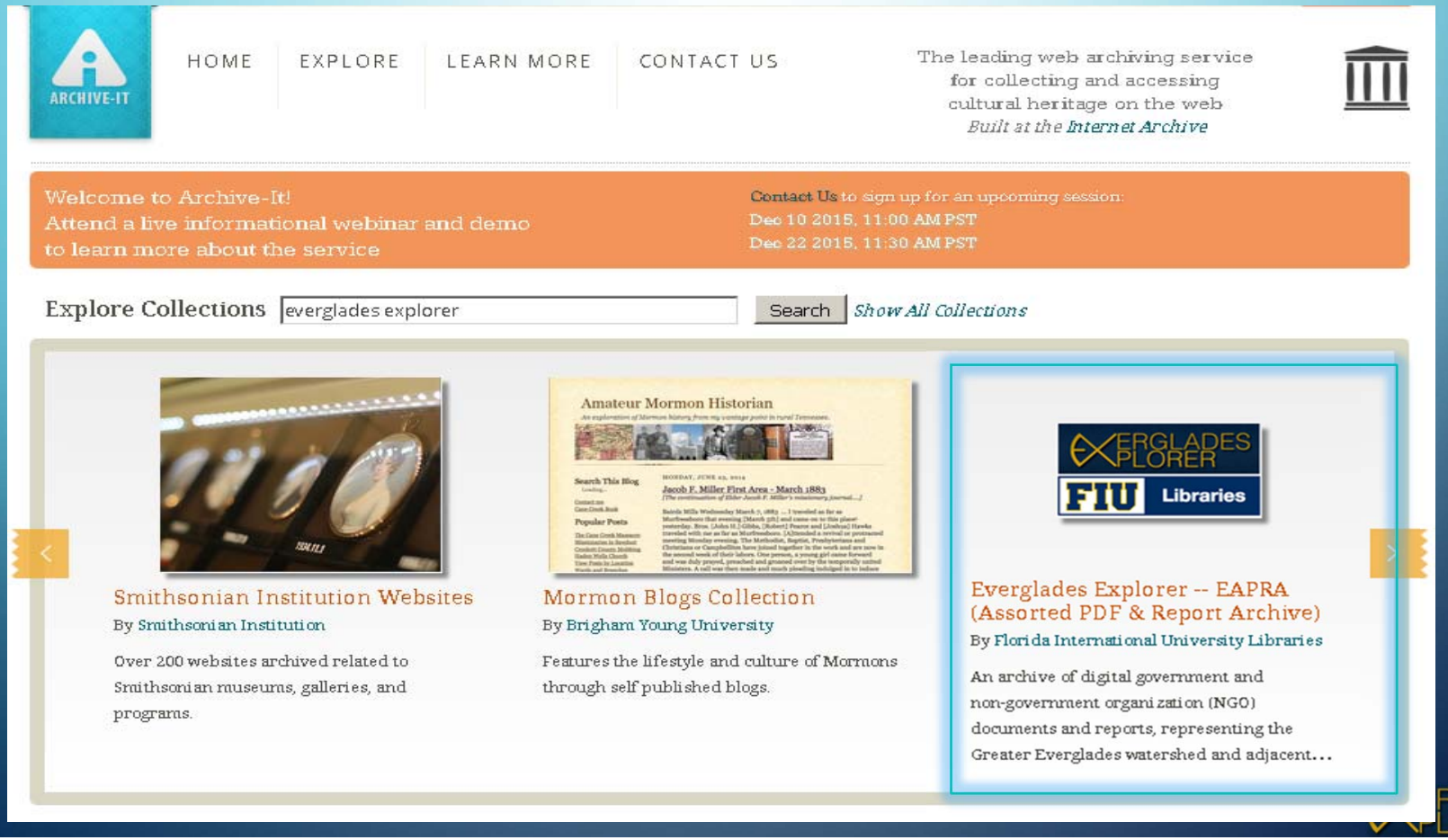
TESTING ACQUISITION OF 3 DIFFERENT FORMATS

- 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

About us | Research | Data | Publications | Student Organization | Education & Outreach | Intranet

What We Do | News | Jobs | Meetings | Personnel | Photos | Visitor Info | About the Everglades | Contact Us

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)

Narrow Your Results

Group Sort By: Count | (A-Z)

ERIC -- Education (11)

Subject Sort By: Count | (A-Z)

Everglades National Park (24)
Big Cypress National Preserve (10)
Estero Bay (10)
Hydrology (8)
Climate Change (6)
Endangered Species (5)
Environmental Education (5)
Lee County (5)
Wildfires (5)
Conservation Land (4)
Estero Bay Agency for Bay Management (4)
Everglades Headwaters (4)
Everglades Restoration (4)
Natural Resources (4)
Science Education (4)
Taylor Slough (4)
Vegetation Report (4)
Water Management (4)
Bibliography (3)
Dry Tortugas National Park (3)
Ecology (3)
Elementary Education (3)
Sea Level Rise (3)
Arnold Committee (2)
Biscayne National Monument (2)

More ▼ Less ▲

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Enter search terms here | Search | Clear

Sites | Search Page Text

Page 1 of 2 (188 Total Results) Next Page ►

Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

Title: South Florida Everglades Restoration

URL: http://cmsdata.iucn.org/downloads/florida.pdf

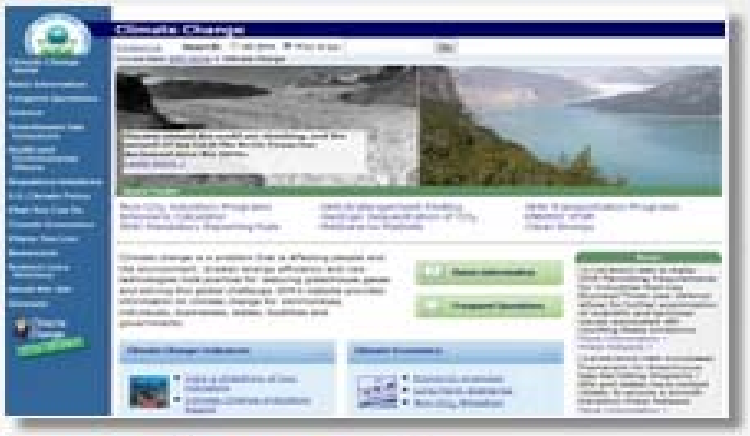
Description: Brief summary of the history of negotiations and agreements between federal and state government and stakeholders of the everglades.
Captured once on Apr 28, 2015
Subject: Environmental law, Everglades National Park, Everglades Restoration
Creator: Ashcraft, Catherine
Publisher: Massachusetts Institute of Technology
Language: English
Coverage: 1988-2005
Format: PDF
Date: 2005
Contributor: Fuller, B.W.

Title: Crusade for the Glades

URL: http://fcelter.fiu.edu/about_us/everglades/general_information/Crusade_for_Glades.pdf

Description: Magazine article about the role of academic scientists and students in Everglades restoration
Captured once on Apr 28, 2015

8 ORGANIZATIONS HAVE COLLECTIONS RELATED TO CLIMATE CHANGE



Climate change and environmental policy

By Stanford University, Social Sciences Resource Group

Stanford University's Social Science Resource Group's collection on Intergovernmental and Non-governmental Organizations that focus on

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.

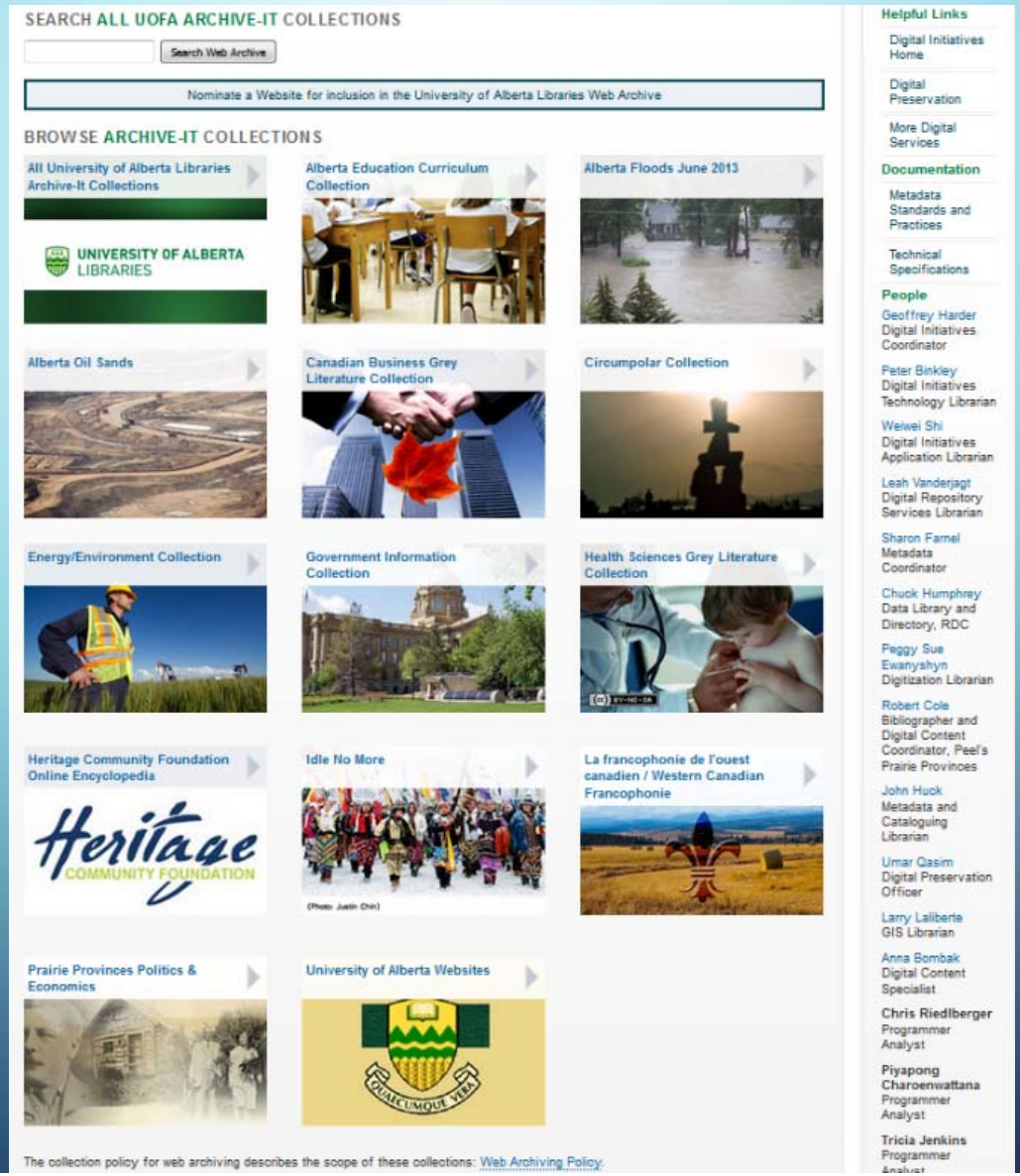
- **Captures university web domains** – focus of University Archives
- **Align Archive-It to existing areas**
  - Collection Development
  - Institutional Academic Specialties /Focus
- **Thematic topics** – global human rights; historic preservation in NYC; NYC religious institutions.

UNIVERSITY OF ALBERTA

- Prairie Politics and Economics
- Government Documents
- Grey Literature – Business & Health Sciences
- Circumpolar Studies
- Education Curriculum Materials

UNIVERSITY OF ALBERTA LIBRARIES

FOCUSED ON DISCOVERY AND FRONT-END DEVELOPMENT



- Merging web archive with Digital Library Systems.

EVERGLADES EXPLORER

MONTANA STATE LIBRARY

ARIZONA STATE LIBRARY

GOVDOCS

- 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.

EVERGLADES EXPLORER

QUESTIONS ?

HTTP://EE.FIU.EDU/

REFERENCES

- Cooper, L. Bryan and Perez Martinez, Margarita, "Linking Old Librarianship to New: Aligning 5-Steps of The Innovator's DNA in Creating Thematic Discovery Systems for the Everglades" (2015). *Works of the FIU Libraries*. Paper 24.
  http://digitalcommons.fiu.edu/glworks/24
- Digital Public Library of America (n.d.) Become a Hub. retrieved from http://dp.la/info/hubs/become-a-hub/
- Enis, Matt. (Feb 5, 2015) Rethinking Privacy at the LITA Top Tech Trends Panel | ALA Annual 2015. library journal. retrieved from http://lj.libraryjournal.com/2015/02/shows-events/ala/lita-members-talk-tech-trends-ala-midwinter-2015/
- Kyrilldou, Martha. "Libqual+: A Project from StatsQual," in Section 1.2 of "Libqual+ 2013 Survey, Association of Research Libraries, Texas A&M."
- Sweetser, Michelle (Marquette University). "Metadata Practices Among Archive-It Partner Institutions: The Lay of the Land." *Archive-It Partner Meeting 2011*. 2011; as found in "The Web Archiving Life Cycle Model."
- "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.
- Wikipedia – "Internet Archive." 11/30/2015

Thank you.