SAA Research Forum, 28 July 2021

Observing changes in EAD Tag Usage to Support Discovery, 2013-2021

Merrilee Proffitt
Senior Manager
OCLC Research Library Partnership

Bruce Washburn
Principal Architect
OCLC Research

- Partnership with California Digital Library, OCLC, University of Virginia Library, and Chain Bridge Group
- Explore and create the blueprint for implementing a national finding aid network that is community-driven, -sustained, and –governed
- Find out more at https://confluence.ucop.edu/display/NAFAN/Building+a+National+Finding+Aid+Network

University of California CDL California Digital Library

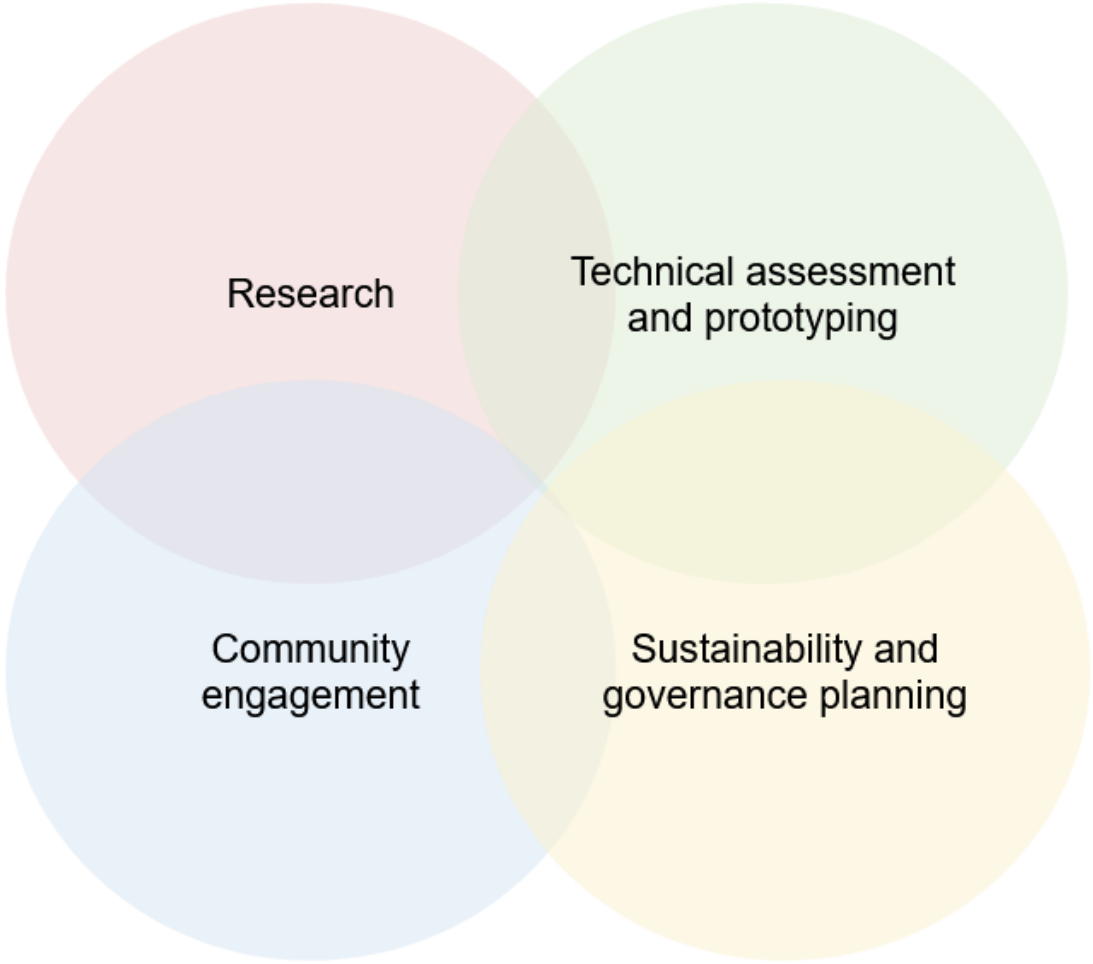
SHIFT Collective equity by design

Chain Bridge Group

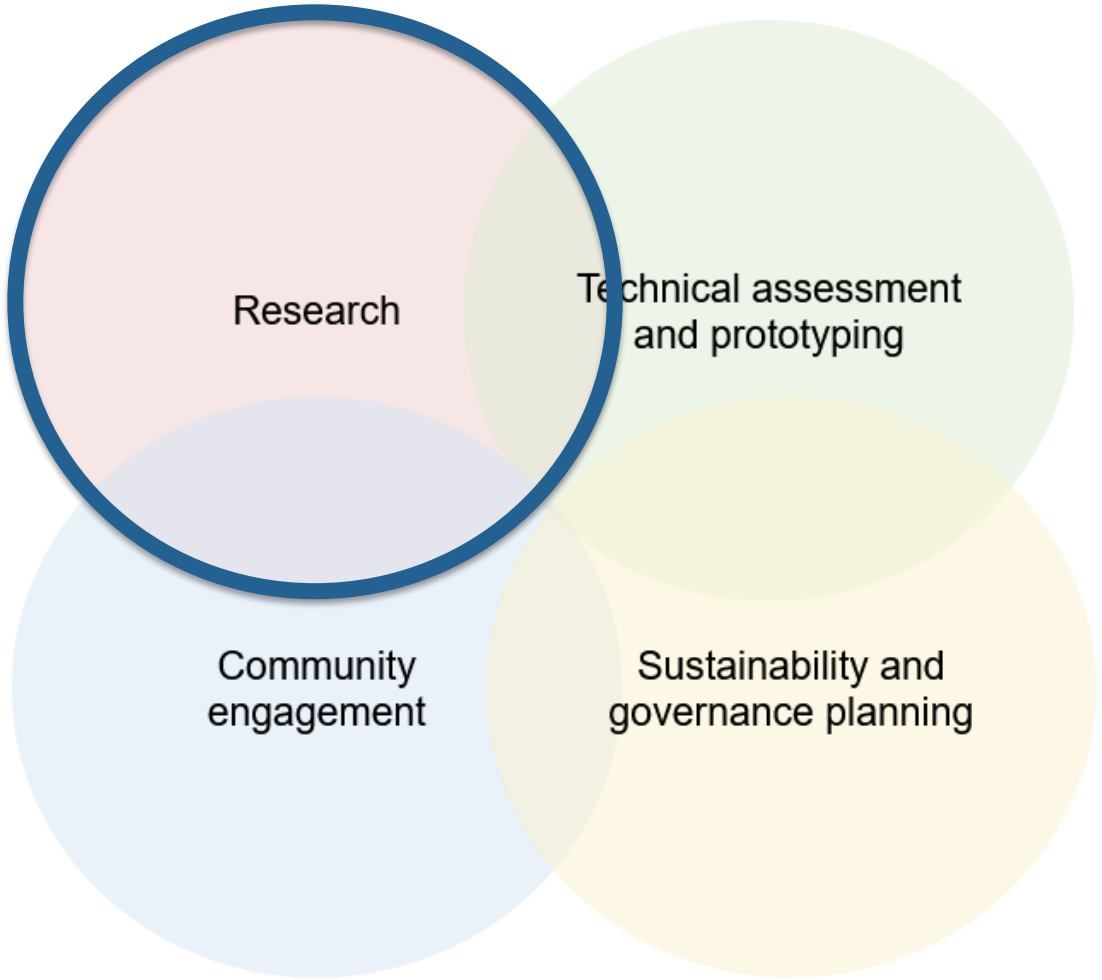
University of Virginia Library

Institute of Museum and Library Services

NAFAN Activity Streams



OCLC's Primary Role: Research

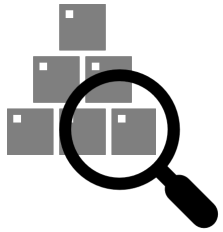


Research Summary

Qualitative: Gain an updated understanding of the needs of both researchers and cultural heritage institutions regarding finding aid aggregation.

Research Summary

Qualitative: Gain an updated understanding of the needs of both researchers and cultural heritage institutions regarding finding aid aggregation.



Quantitative: Evaluate an aggregation of finding aids in order to:

- scope the network's functionality to the finding aid data
- lay the groundwork for data remediation and expanded network features.

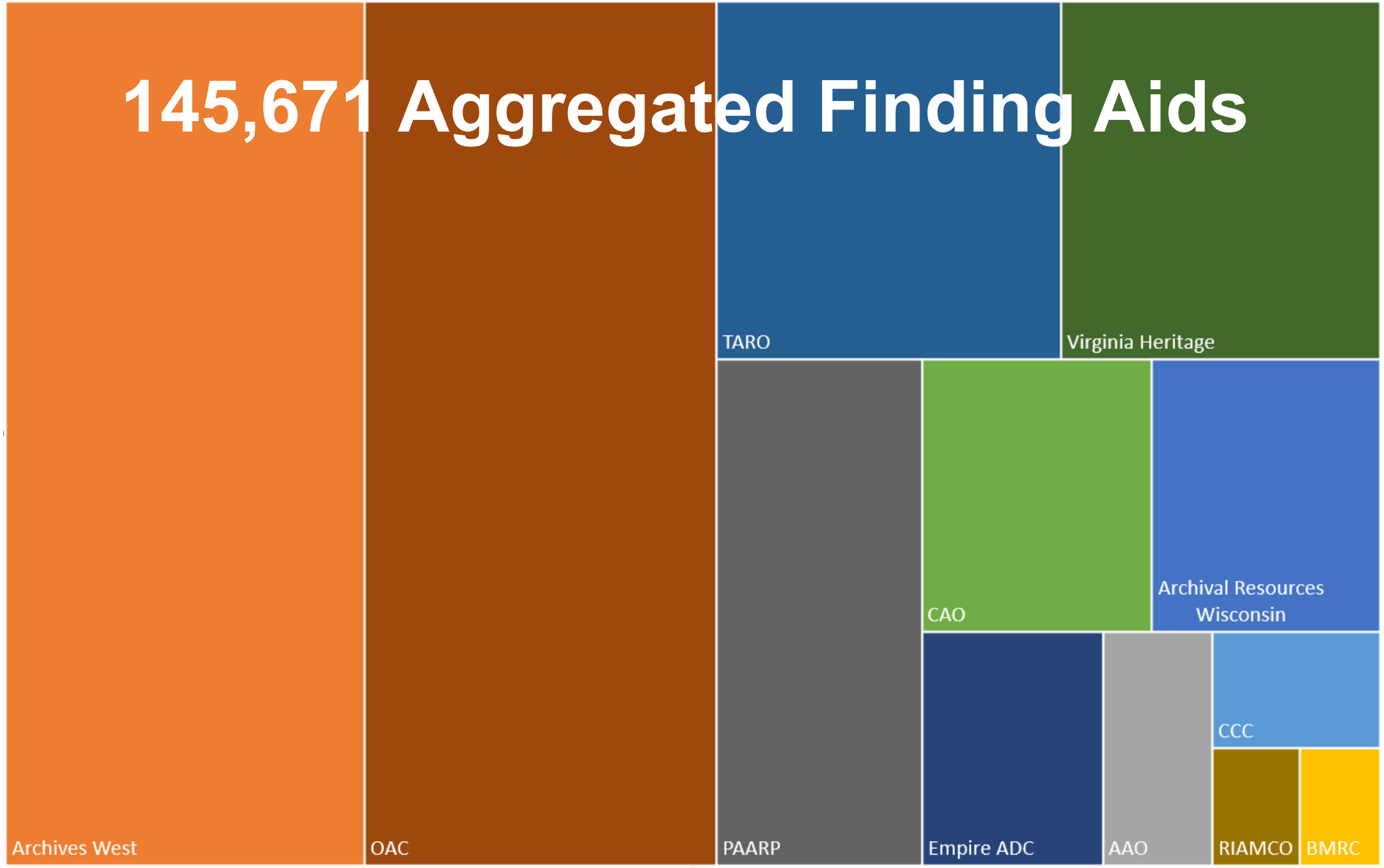
Research Summary

Qualitative: Gain an updated understanding of the needs of both researchers and cultural heritage institutions regarding finding aid aggregation.

Quantitative: Evaluate an aggregation of finding aids in order to:

- scope the network's functionality to the finding aid data
- lay the groundwork for data remediation and expanded network features.

145,671 Aggregated Finding Aids

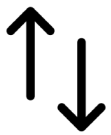


Dimensions for analysis

Search: all discovery systems have a keyword search function; many also include the ability to search by a particular field or element.

Browse: many discovery systems include the ability to browse finding aids by title, subject, dates, or other facets.

Results display: once a user has done a search, the results display will return portions of the finding aid to help with further evaluation.



Sort: once a user has done a search, they may have the option to reorder the results.

Facet: once a user has done a search, they may have the option to narrow the results to only include results that fall within certain facets.

EAD elements that impact discovery

- Dates
  - unitdate
- Extent data
  - extent
- Collection title sources
  - unittitle, titleproper/@type=filing
- Content tags in dsc
  - corpname, famname, function, genreform, geogname, name, occupation, persname, subject
- Content tags in origination
  - corpname, famname, name, persname
- Content tags in controlaccess
  - corpname, famname, function, geogname, name, occupation, persname, subject
- Material type
  - controlaccess/genreform
- Repository
  - repository
- Notes
  - abstract, bioghist, scopecontent

Derived from "Thresholds for Discovery: EAD Tag Analysis in ArchiveGrid, and Implications for Discovery Systems" published in code4lib Journal issue 22.

Defining Thresholds for Discovery

Usage	Threshold
0-50%	Low
51-80%	Medium
81%-95%	High
96%-100%	Complete

Discovery Category	EAD Element	% in 2021	% in 2013
Dates	unitdate	81.89	72.64
Extent data	extent	99.83	70.43
Collection Title sources	unittitle	99.98	99.93
	titleproper @type=filing	0.08	
Content tags in dsc	corpname	2.22	6.3
	famname	0.13	0.95
	function	0.04	0.05
	genreform	3.36	7.24
	geogname	1.39	6.89
	name	0.11	0.39
	occupation	0.06	0.3
	persname	5.02	12.4
	subject	1.03	2.51
origination and content tags	origination	85.31	87.78
	corpname	32.16	
	famname	2.07	
	name	0.32	
	persname	69.18	
archdesc/controlaccess and content tags	controlaccess	83.48	72.89
	corpname	26.26	
	famname	1.9	
	function	0.17	
	geogname	17.24	
	name	0.01	
	occupation	6.05	
	persname	27.85	
	subject	45.78	
Material type	genreform	20.88	
Repository	repository	99.68	99.45
Notes	scopecontent	87.57	84.41
	abstract	84.63	79.2
	bioghist	71.64	70.42

Representation	Threshold
0-50%	Low
51%-80%	Medium
81-95%	High
96%-100%	Complete

Discovery Category	EAD Element	% in 2021	% in 2013		Representation	Threshold
Dates	unitdate	81.89	72.64		0-50%	Low
Extent data	extent	99.83	70.43		51%-80%	Medium
Collection Title sources	unittitle	99.98	99.93		81-95%	High
	titleproper @type=filing	0.08			96%-100%	Complete
Content tags in dsc	corpname	2.22	6.3			
origination and content						
archdesc/controlaccess						
	persname	27.85				
	subject	45.78				
Material type	genreform	20.88				
Repository	repository	99.68	99.45			
Notes	scopecontent	87.57	84.41			
	abstract	84.63	79.2			
	bioghist	71.64	70.42			

"If the archival community continues on its current path, then the potential of the EAD format to support researchers or the public in discovery of material will remain underutilized. Minimally, collection descriptions that are below the thresholds for discovery will hinder their discovery efforts and maximally will remain hidden from view.

Perhaps with emerging evidence about the corpus of EAD, continued discussion of practice, recognition of a need for greater functionality, and shared tools both to create new EAD documents and improve existing encoding, we can look forward to further increasing the effectiveness and efficiency of EAD encoding and develop a practice of EAD encoding that pushes collection descriptions across the threshold of discovery."

https://journal.code4lib.org/articles/8956

More Research Opportunities

- Finding aid links to digital content
- Specifying physical characteristics and genre
- Content element value connections to controlled vocabularies
- Consistency of institutional contact information
- Informing researchers about access to, use of, and reuse of materials

More Research Opportunities

- Finding aid links to digital content
- Specifying physical characteristics and genre
- Content element value connections to controlled vocabularies
- Consistency of institutional contact information
- Informing researchers about access to, use of, and reuse of materials

Finding aid links to digital content

- How many links are there, per finding aid?
- Which EAD elements and attributes include links?
- What types of digital objects are linked?
- Are links a complete URL or are they relative?
- How many links resolve?

Review

Topics Covered:

- The NAFAN “Building a National Finding Aid Network project”
- NAFAN activity streams and OCLC’s role
- Qualitative and Quantitative research objectives
- The NAFAN EAD aggregation
- A Discovery-oriented EAD tag analysis, past and present
- More research opportunities

Contact Us

Merrilee Proffitt

Senior Manager

OCLC Research Library Partnership

merrilee_proffitt@oclc.org

Bruce Washburn

Principal Software Engineer

OCLC Research

bruce_washburn@oclc.org