EMORY
LIBRARIES &
INFORMATION
TECHNOLOGY

Metadata Working Group: Descriptive Metadata Guidelines Project Update

July 2014

Metadata Working Group

- Library Cabinet-level group; re-chartered summer 2013
- Library Cabinet: leadership body including
  - School library directors
  - Woodruff Library Sr. Directors
  - Emory Center for Digital Scholarship; Library & Information Technology Services representatives
- Metadata Working Group's broad charge:
  - Identify and promote best practices
  - Unified organizational metadata strategy
- Current focus: metadata for digital collections* (non-MARC)

**Unique, Emory-created collections of digitized or born-digital content, intended for delivery to an Emory Libraries-supported repository or discovery tool*

Environmental Scan of Metadata (EUL, ECDS): Autumn 2013

- Inventory – Key Statistics
  - **60+** digital projects/collections*
    - **10+ business units**
    - **16+ data entry tools**
    - **10+ search systems**
    - **13+ technical standards**
    - **10+ years of silo-ed activity**
    - **10Ks+ items not findable**
  - **Multiple environments** for discovery, preservation, and presentation
  - **0** central standards/guidelines

- Metadata User Profiles
  - Diverse skills/needs
    - Digital content creators
    - Subject Matter Experts/Service Owners
    - Archivists/preservationists
    - Catalogers
    - (others) …
- Challenges
  - Organizational restructuring/ staffing changes
  - Major Library system platforms in flux
  - History of silo-ed, project-level practices

Metadata: User Comments

I suspect I need some metadata…?

How do we get crawled by Google/Scholar/ Images?

Nobody is going to fill out this form.

It's faster to collect it than describe it.

Our data entry staff don't understand how to fill out these fields.

What's going on with the repository?

How do we get {this project} into Primo?

I digitize things all day long and no one ever sees them.

Which schema should we use?

Metadata is a bottleneck.

Let's push the boundaries do some cool stuff.

How do we map our {schema X} metadata to {schema Y} for {system Z}?

We've also got a lot of stuff on {the shared drive}…

Approaching best practices

Actionable targets in a time of change

Managing complexity

How do we move forward?

Guidelines for Descriptive Metadata: Project Scope (Jan 2014)

1. **Based on current local and industry practices, identify baseline descriptive metadata standards for non-MARC, digital collections*/ projects:**
   - A ***set of core metadata fields***, required and/or recommended for all projects
   - Schema-independent: can be mapped to common standards (MODS/DC/VRA etc.)
   - Focused on end-user discovery needs in core Library systems – explaining what each field means, how it is utilized
2. **Provide content standards for fields:**
   - How to populate and format metadata entries for core fields
   - Accessible to a broad audience, including non-specialists
   - Including clear, specific usage examples (and how to troubleshoot)
3. **Provide public-facing, web-based delivery of metadata guidelines**
   - Searchable, browse-able, cross-referenced
   - Including introductory material, FAQs, and tips

What do we mean by “core” metadata?

- **We can always do more, but we shouldn’t do *less***
  - *Specific schemas, systems, and local practices may have stricter requirements, but local schemas should not have looser requirements*
- Critical to end-user discovery and access
- Commonly required in relevant standards (local, external)
- Frequently utilized in Emory system interfaces for search, sort, filter, browse
- Broadly applicable: scales across multiple content types and standards; enhances interoperability
- Framework to build upon (tiered approach)

How can we use core metadata?

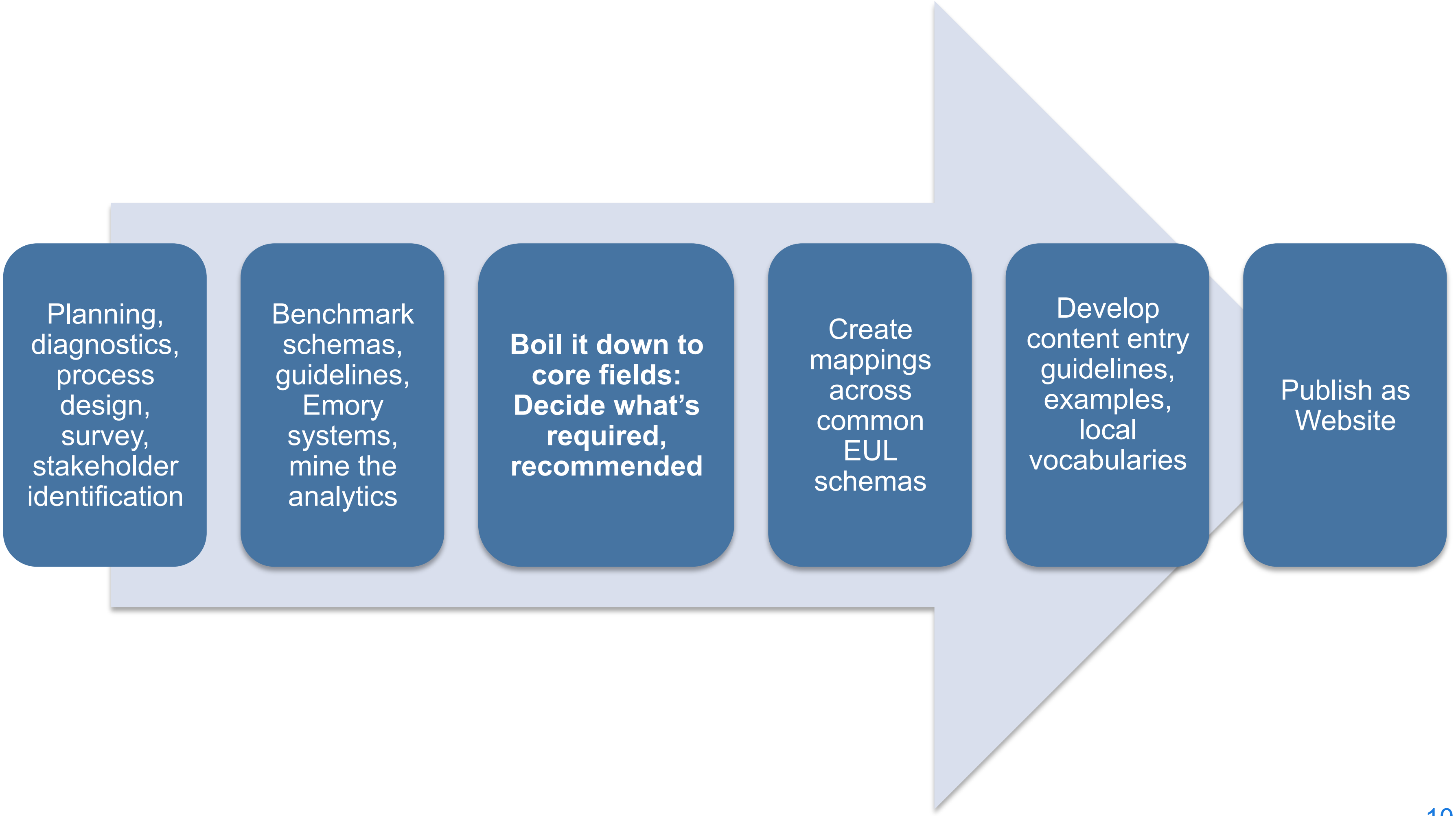
- **Digitization**: minimal descriptive metadata if standards not otherwise specified
- **Repository ingest/collection assessment:** retrofitting collections with no/limited metadata
- **Repository architecture:** re-usable mappings; indexing; local vocabularies
- **Born-digital content creators, faculty/scholars/researchers:** guidance for non-specialists to provide core data at time of content creation
- **Primo ingest:** potential to streamline common non-MARC mappings, data formatting
- **Outsourcing to vendors:** more efficient RFP and QA processes
- **Collaboration/sharing:** publishing metadata standards and mappings assists external collaborators
- **Alma adoption (future state):** standards can be applied to non-MARC processes with future support of MODS/other schemas
- **System migrations (future state):** core element mappings facilitate migration to new systems
- **Schema evolutions (future state):** maintaining core mappings across multiple schemas will benefit metadata utilization over time

Getting there:

- Research-driven
- Work Process
- Timeline
- Activities

PROJECT ACTIVITIES & OUTCOMES

Stages of Work



Planning, diagnostics, process design, survey, stakeholder identification

Benchmark schemas, guidelines, Emory systems, mine the analytics

Boil it down to core fields: Decide what's required, recommended

Create mappings across common EUL schemas

Develop content entry guidelines, examples, local vocabularies

Publish as Website

Group Task Analysis

- Created simple Dublin Core records for Emory Center for Digital Scholarship's Battle of Atlanta project
  - Test of "minimal"/core metadata with varying experience levels
  - Inconsistent results
    - Interpretation of element meanings
    - Data encoding and formatting issues

→ Identified opportunities for local vocabularies/boilerplate content

→ Affirmed need for documentation and local content standards (not just a set of elements)

2014 Stakeholder Survey: Selected Statistics

- Respondent roles:
  - Collection or service owner (32%)
  - Stakeholder (Other) (20%)
  - Creator (16%)
  - Systems administrator/developer (12%)

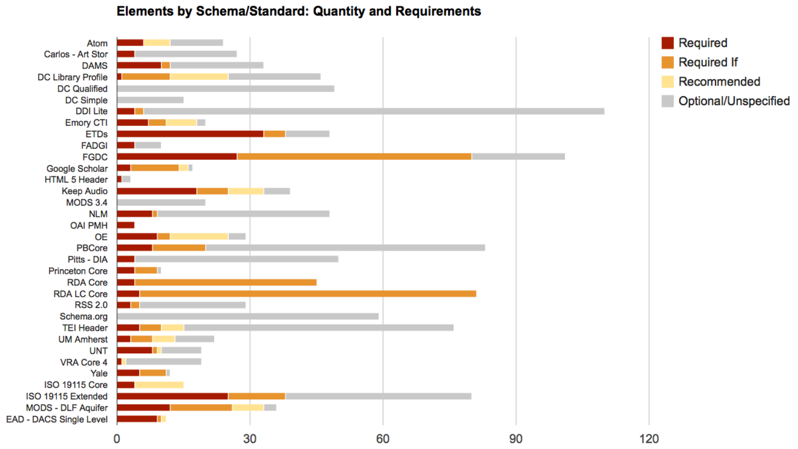
- Seen as most helpful:
  - Organizational standards (68%)
  - Quality control (64%)
  - Schema selection (60%)
  - Controlled Vocabularies Support (60%)
  - Documentation/training (60%)
  - System integration (60%)

- Most engagement with:
  - Dublin Core (Simple) (75%)
  - MODS (50%)
  - Custom schemas (42%)

- Metadata is created by:
  - Staff (in same unit) (60%)
  - Graduate students (60%)
  - Content creators/scholars (40%)
  - Staff (outside unit) (28%)
  - Undergraduate students (12%)
  - No one (12%)

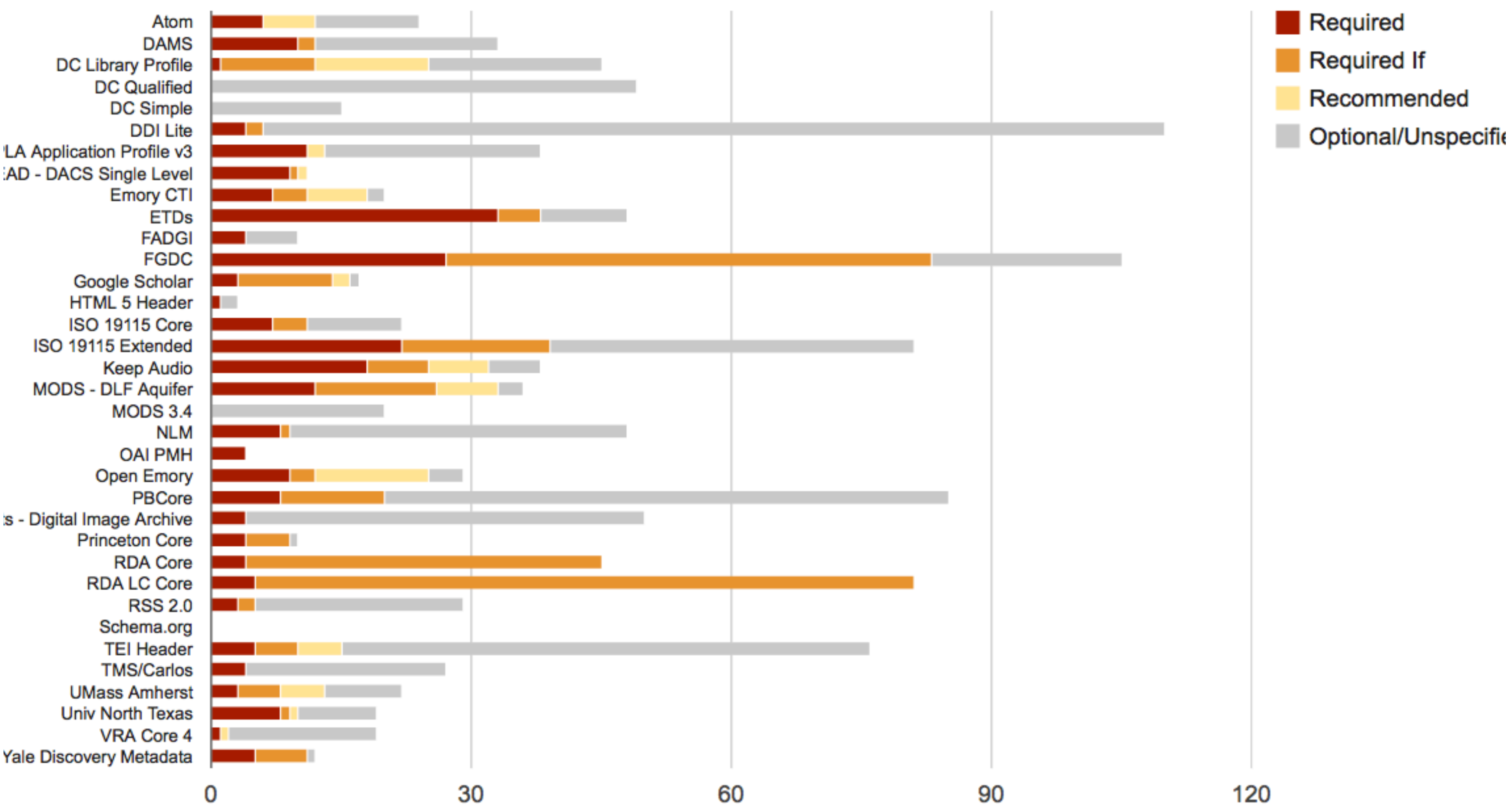
Research: Benchmarking

- Structured comparison of 34 schemas/standards reflecting Emory content types
  - Emory/local
  - Formal XML/XSD/ISO standards
  - National/international/consortial
  - Other libraries
  - Major content standards
- For each standard, reviewed:
  - Element names, meanings
  - Required-ness
- Concept analysis of element names and meanings
- Quantitative analysis:
  - Required in % of schemas
  - Raw counts


Elements by Schema/Standard: Quantity and Requirements

- □ MODS, Dublin Core, VRA, TEI…
- □ DDI, FGDC, ISO 19115...
- □ RDA Core, DACS, DLF Aquifer...
- □ DPLA, FADGI, NLM...
- □ ETDs, Open Emory, Keep, Pitts...
- □ RSS, Google Scholar, Schema.org...

Benchmarking: Elements



Elements by Schema/Standard: Quantity and Requirements

Research: System Profiles

- Reviewed and logged:
  - Interface options
    - Structured search
    - Browse
    - Facet/filter
    - Sort
  - General indexing
  - General requirements/ optimization notes
  - General schema support

1. Dataverse
2. Emory Theses & Dissertations
3. Fedora (general)
4. Finding Aids
5. The Keep
6. LUNA
7. OpenEmory
8. Pitts Digital Image Archive
9. Portfolio Server (DAMS)
10. Primo

Research: Summary of System Analytics

- Reviewed 3 months' data (Feb-May) as available:
  - LUNA
  - Primo
  - ETDs
  - Open Emory

- Investigated potential metadata patterns in:
  - Top 25 pages by URL
  - Top 25 search terms

- Metadata Activity Patterns:
  - Creators/Personal Names
  - Collection Names
  - Titles
  - Subjects
    - Topics/keywords
    - Time Period/Culture ("Roman")
    - Geographic Names
  - Identifiers/PIDs
  - School/Program Names
  - Contributors (thesis committees; donors)

Top Element Concepts – By the Numbers

Ranking Methodology:

- Pulled most-required elements across **all** standards (greater than 15%)
- Averaged system UI utilization (structured search/browse /facet/sort)
- Averaged analytics activity
- Created weighted scoring metrics for criteria:
  - 2x for standards' required-ness
  - 1x for system UI utilization
  - .5x for analytics

- **Title**
- **Identifier**
- **Subject** (aggregated)
- **Creator** (aggregated)
- **Location** (aggregated)
- **Date** (aggregated)
- **Subject/Topic/Keyword**
- **Collection**
- **Type**
- **Rights/Access**
- **Location/Institution/Affiliation**
- **Location/Institution/Repository**
- **Language**
- **Description**
- **Contributor** (aggregated)

Top Elements and Benchmarking Process

System Interfaces

Analytics

APPENDIX: SUMMARY OF DATA COLLECTED

High Level Statistics - Elements

Element Sets/Standards Reviewed	**34**	
Emory Instances	7	21%
Total Elements	**1340**	
Emory: Elements	245	18%
Total Required Elements	**251**	**19%**
Emory: Required Elements	85	35%
Total Required if Applicable Elements	**291**	**22%**
Emory: Required if Applicable	21	9%
Total Recommended* Elements	70	5%
Optional/Unspecified Elements	728	50%
Recommended/Optional Elements	**740**	**55%**
Emory: Recommended/Optional	139	57%

Top Element Concepts/Tags - Clarified

Element/Concept	Meaning/Examples
Title	Title/name of resource being described (generally consistent meaning across standards)
Identifier (aggregated)	Unique identifier: filename, URL/URI, accession/call number, standard record number (ISBN)
Creator (aggregated)	Creator name (primary creation role such as author, photographer, researcher)
Subject (aggregated)	Aggregate for multiple types of subject terms (keyword, topic, geographic, name) – except genre
Location (aggregated)	Aggregate for parent library/museum/institution: generally an institution name where things are located/housed/served from. May also be be a URL for primary digital access point.
Date (aggregated)	Aggregate: includes date of creation, date of publication, other types of dates
Subject/Topic/Keyword	Narrower: focused on topics or uncontrolled keywords (not geographic, names)
Type	Type of original content (map, data set, image, text); e.g. MODS Type of Resource
Collection	Name of a collection that the item is a part of (e.g. Langmuir African American Photographs)
Rights/Access	Statement or condition of usage or access
Location/Institution/Repository	Parent library/museum/institution – often physical holdings-related; owner or steward
Language	Language of the content
Contributor (aggregated)	Aggregate of various contributor types (secondary role: donor, funder, research contributor vs. PI)
Description (aggregated)	Aggregate of different descriptions (abstract, table of contents, general/misc. description)

Top Elements - Overall

Rank	Element/Concept	Req'd in % of Schemas	EU System Interfaces Utilizing	Analytics Activity – in % of Systems	Weighted Score
1	Title	76%	22%	50%	57%
2	Identifier	50%	13%	50%	40%
3	Subject (aggregated)	21%	55%	75%	39%
4	Creator	24%	61%	50%	38%
5	Location (aggregated)	29%	32%	75%	36%
6	Date (aggregated)	35%	50%	0%	34%
7	Subject/Topic/Keyword	15%	48%	75%	31%
8	Collection	15%	36%	50%	26%
9	Type	32%	13%	25%	25%
10	Rights/Access	41%	0%	0%	23%
11	Location/Institution/Affiliation*	3%	28%	75%	20%
12	Location/Institution/Repository	21%	24%	0%	19%
13	Language	26%	8%	0%	17%
	Description	26%	8%	0%	17%
	Contributor	12%	12%	50%	17%

Used in System Interfaces = aggregated utilization across % of systems (search options, facets, browse, etc)
Analytics Activity – user activity: search or browse usage of element category, in % of systems reviewed
Weighted rank counts schema requirement percentage 2x; UI options 1x; analytics .5x

EMORY
LIBRARIES &
INFORMATION
TECHNOLOGY

Top Elements – Compared Across Standards

Element/Concept	ETD	OE	Keep	DAMS	DC LAP	MODS Aquifer	DDI Lite	ISO 19115 Core	TEI Header	EAD DACS	RDA Core	VRA Core
Title	M	M	M	M	M	M	M	M	M	M	M	O
Identifier	M	M	M	M	O	R	O	O	O	M	C	O
Creator	M	M	C	C	O	C	O	?	R	C	C	O
Subject (aggregated)	M	O	?	O	C	C	O	M	O	R	?	O
Location (aggregated)	M	O	R	M	C	M	?	?	?	M	?	O
Date (aggregated)	?	M	M	M	O	M	O	M	O	M	C	O
Subject/Topic/ Keyword	M	O	?	O	C	C	O	M	O	R	?	R
Type	M	M	M	M	O	M	O	?	?	?	M	O
Collection	?	?	M	M	?	?	O	?	O	?	?	?
Rights/Access	M	M	M	M	C	M	O	?	O	M	?	O
Language	M	R	M	O	C	C	?	M	O	M	C	?
Contributor	M	O	M	N/A	C	C	O	?	O	M	C	O
Location/Institution/ Repository	?	N/A	?	M	?	O	?	?	?	M	?	O
Description	M	R	?	O	R	R	O	M	O	M	C	O

M= mandatory; C = conditional; R = recommended; O = optional; ? = unclear

Missing the Cut (Lower Rankings)

- Format (aggregated concept for physical/digital carrier characteristics)
  - Physical or digital characteristics (size, extent, dimensions, medium, file type, duration)
  - Borderline element: scoring impacted by analytics
- Genre
  - Specific sub-categories of Type (e.g. score, postcard, letter, black and white photograph)
- Publication Event information
  - Publisher, Date of Publication, Place of Publication, parent source/citation
- Specific subject sub-types
  - Geographic/geospatial information
- Granular roles for Creators/Contributors
- Granular Date types
- Content Status
- Related/relationships
- Audience
- Metadata record information

Element Sources Reviewed - Types

Types of Sources Reviewed:

- National/international standards
- Technical/XML schema/XSD
- Content standards/Element sets
- Web publishing/content syndication practices
- Scholarly Publishers
- Major consortia/agencies
- Emory projects/repositories
- Other libraries

Types of Content Represented:

- Still Images
- Manuscripts/archival collections
- Audiovisual resources
- Articles/analyzed publications
- Websites
- Data sets
- Geospatial data
- Full text

List of Standards Reviewed (34)

- Atom
- Emory: Carlos Museum/TMS
- Emory: DAMS
- Dublin Core Library Profile
- Dublin Core - Qualified
- Dublin Core - Simple
- DDI Lite (Statistical Data)
- DPLA Application Profile v3
- EAD - DACS Single Level
- Emory CTI Best Practices
- Emory ETDs
- Emory - Keep Audio
- Emory - Pitts Digital Image Archive
- Emory – Open Emory
- FADGI Embedded Metadata for Still Images
- FGDC (descriptive sections)
- Google Scholar
- HTML 5 Header
- ISO 19115 Core (GIS data)
- ISO 19115 Extended (descriptive sections) (GIS data)
- MODS - DLF Aquifer Guidelines
- MODS 3.4
- National Library of Medicine (NLM)
- OAI-PMH (record header)
- PBCore
- Princeton Core Metadata
- RDA Core
- RDA LC Core
- RSS 2.0
- Schema.org
- TEI Header
- UMass Amherst Metadata
- University of North Texas Metadata
- VRA Core 4
- Yale Discovery Metadata

Systems: Browse Options

System	Option	Element/Concept
Finding Aids, OE, ETD; Primo	A-Z; Creator Name; Author	Creator
Finding Aids, Dataverse, LUNA	Repository; Emory Dataverse; Woodruff Library	Collection, Location
Finding Aids, Pitts DIA, Primo	Browse LCSH; Call Number	Subject
Pitts DIA	Scripture Reference	Related
Pitts DIA	Full Volumes – PDF/JPG	Format
Open Emory	By Journal	Source/Citation
Open Emory, ETDs; Primo	By Subject; Research Field	Subject/Topic/Keyword
Open Emory	By Department	Location/Institution/Affiliation
Open Emory	By Faculty Profiles	(tied to Creators?)
ETDs, OE	Program; Department	Location/Institution/Affiliation; Degree
ETDs	Committee Member	Contributor
ETDs	Year	Date
Primo	Title	Title
Primo	(LCC, DDC, UDC, RVK)	Classification

Systems: Structured Search Options

System	Summarized Options	Element/Concept
Dataverse, Primo, Finding Aids; LUNA	Collections (by name); Repositories	Collection
ETDs	Committee Member	Contributor
Dataverse, Primo, ETDs	Author; Creator; Producer	Creator
Dataverse	Universe	Data Demographic/Data Sample
Dataverse; Primo	Kind of Data; Material Type; Items; (lists of specific types)	Type/Genre
Dataverse	Woodruff Library; Emory Dataverse	Collection; Location/Institution/ Repository; Affiliation
ETDs, Dataverse, Primo	Program Year; Production Date; Dist. Date; Last [X] years…	Date
Dataverse, ETDs	Description; Abstract; Table of Contents	Description
Dataverse, Primo	Study ID; Other ID; ISBN; ISSN	Identifier
Primo	(selected languages as text values)	Language
Dataverse	Related Publications	Related
Finding Aids; Dataverse; Primo; ETDs	Subject; Keywords; Research Field; (lists of selected broad subject areas); Country/Nation; Geographic Coverage	Subject (incl. keywords); Subject/ Geographic
Dataverse, Primo	Title	Title
Primo	Series	Series
Dataverse	Distributor	Publisher/Distributor

Systems: Search Results’ Facets/Filters

System	Option	Element/Concept
Finding Aids, Dataverse, Open Emory, Primo	Keywords, Country/Nation; Topic Classification; Topic; Subject	Subject, Subject/Geographic, Subject/Topic/Keywords
Open Emory, Primo	Journal; Journal Title	Source/Citation
Dataverse, Primo	Original Dataverse; Collection	Location/Institution/Repository; Collection
Primo	Related	Related
Primo	Library	Location/Institution/Repository
Primo	Genre	Genre
Finding Aids, Pitts DIA, Primo	Digital objects w/in collection; List of images; Image Gallery; PDF; JPG, File size, Format	Format
Dataverse	Distributor	Publisher/Distributor
Dataverse, Open Emory, Primo, ETDs	Production Date; Distribution Date; Year; PNX Date	Date, Date Produced, Date Distributed, Date Created
Primo, Dataverse, Open Emory, ETDs	Creator/Contributor; Author; Committee	Creator, Contributor
Primo	(LCC, DDC, UDC, RVK)	Classification
ETDs	Program	Affiliation/Location/Degree
Primo	Language	Language
Primo	Resource type	Type/Genre

Systems: Search Results’ Sort Options

System	Options	Element/Concept
Finding Aids, Dataverse, Open Emory, ETDs	Relevance	Mixed: N/A?
Dataverse	Global ID	Identifier
Dataverse, Primo	Most Recently Released; Production Date; Creation Date	Date; Date Created; Date Produced
Primo, Dataverse	Popularity; Most Downloaded	Usage Statistics
Primo, ETDs	Title	Title
Primo, ETDs	Author	Creator

Analytics Review: Methodology

- Pulled 3 months data (Feb-May) for available accounts:
  - ETDs, Open Emory, Primo, LUNA
  - Others identified but not available at time of activity
- Focused on
  - Activity related to specific metadata elements/options
  - URLs, not page titles (not all systems have unique page titles)
  - Searchbox utilization (when available in data)
  - Browse activity
  - Search options' usage (if detectable)
  - Raw search terms > categorized to metadata elements/concepts
- Questions
  - Why do we see relatively low search usage per session on most accounts?
  - Can we invest more resources into better analytics integration (e.g. link-level tracking)?

Analytics: Overall Summary

Most traffic
(ranked by page views):

1. Primo
2. ETDs
3. LUNA
4. Open Emory

Note: due to high volume of activity, even "top" individual URL entries often indicate less than 1% of activity

Browse and search patterns:

- Creators/Personal Names
- Collections
- Titles
- Subjects
  - Subject/topic/keywords
  - Subject/Time Period/Culture ("Roman")
  - Subject/Geographic
- Identifiers/PIDs
- School Names/Program Names
- Contributors (thesis committees; donors)

Analytics: Primo

- **716,677 page views**
- User interface options: "browse search", tabbed search options (advanced and simple), facets, sort
- Top content accessed via URLs include:
  - Login/My Account
  - Multiple search options
  - Query-related URLs
- Search and query-related URLs are 52% of top 25 URLs.
- No presence of "browse" in URL patterns in top 25 (first appears at #27).

- **Site search usage:** not available in Google Analytics
- Search options/UI states noted in top 25 URL parameters:
  - Basic mode (5 instances)
  - Advanced mode (1 instance)
  - Articles tab (2 instances)
  - Combined tab (1 instance)
  - Emory Catalog tab (1 instance)
  - **Repositories tab (1 instance, 0.04%)**
  - Default entry state/no session parameters (1 instance)

Analytics: ETDs

- **Page views: 119,976**
- User interface options: quick search and advanced search; browse; facet; sort
- Top content by URL:
  - Administrative/internal
  - Browse (program, college, author, committee)
  - Specific item records/PIDs
  - Help/support
  - Search
- Browse usage correlates to navigation options vs. search results

- **Searchbox usage: 3%** of sessions
- */search* is #5 out of 25 = Advanced Search
- Top search term categories:
  - Topics/Keywords
  - Personal Names/Creators
  - Program Name/School Name/ Affiliation

Analytics: Open Emory

- **Page views: 33,643**
- Interface options: browse, combined search, facets
- Top content by URL accessed is
  - Administrative/internal
  - General/about/support
  - Faculty/Researcher/Author Profiles
- Profiles accessed via search results or browse from navigation (can't determine from data)
- Search is not in top 25 (first instance of search in URL pattern is at #36)

- **Searchbox usage: 4%** of sessions
- Top search term categories:
  - Full/exact Titles
  - Personal Names/Creators
  - PIDs/identifiers
  - Topics/keywords

Analytics: LUNA

- **Page views: 96,798**
- User interface options: quick/ advanced search; browse; facets
  - Options are configurable on per-collection basis
- Top content accessed by URL includes
  - Homepage
  - Extensive browse activity
    - (Who, What, Where facets)
    - Views of specific collections
  - "search" in URL patterns displays browse screens, or indicates blank search submissions

- **Search usage: 15%** of sessions
  - Highest indicated
- Search term categories:
  - Collection names
  - Titles
  - Subjects
    - Geographic
    - Topics/keywords
    - Time periods
  - Names (donor/contributor)