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
I suspect I need some metadata…?

It’s faster to collect it than describe it.

How do we get {this project} into Primo?

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

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

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

What’s going on with the repository?

Nobody is going to fill out this form.

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

Which schema should we use?

Let’s push the boundaries do some cool stuff.

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

Metadata is a bottleneck.

It’s faster to collect it than describe it.

It’s faster to collect it than describe it.

It’s faster to collect it than describe it.

It’s faster to collect it than describe it.

It’s faster to collect it than describe it.

It’s faster to collect it than describe it.

It’s faster to collect it than describe it.

It’s faster to collect it than describe it.

It’s faster to collect it than describe it.

It’s faster to collect it than describe it.

It’s faster to collect it than describe it.

It’s faster to collect it than describe it.

It’s faster to collect it than describe it.

It’s faster to collect it than describe it.
How do we move forward?

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
PROJECT ACTIVITIES & OUTCOMES

Getting there:
• Research-driven
• Work Process
• Timeline
• Activities
Stages of Work

1. Planning, diagnostics, process design, survey, stakeholder identification
2. Benchmark schemas, guidelines, Emory systems, mine the analytics
3. Boil it down to core fields: Decide what’s required, recommended
4. Create mappings across common EUL schemas
5. Develop content entry guidelines, examples, local vocabularies
6. 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

- 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

- Atom
- DAMS
- DC Library Profile
- DC Qualified
- DC Simple
- DDI Lite
- LA Application Profile v3
- AD - DACS Single Level
- Emory CTI
- ETDs
- FADGI
- FGDC
- Google Scholar
- HTML 5 Header
- ISO 19115 Core
- ISO 19115 Extended
- Keep Audio
- MODS - DLF Aquilar
- MODS 3.4
- NLM
- OAI PMH
- Open Emory
- PBCore
- P - Digital Image Archive
- Princeton Core
- RDA Core
- RDA LC Core
- RSS 2.0
- Schema.org
- TEI Header
- TMS/Carlos
- UMass Amherst
- Univ North Texas
- VRA Core 4
- Yale Discovery Metadata

Legend:
- Required
- Required If
- Recommended
- Optional/Unspecified
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)
APPENDIX: SUMMARY OF DATA COLLECTED

Top Elements and Benchmarking Process
System Interfaces
Analytics
## High Level Statistics - Elements

<table>
<thead>
<tr>
<th>Category</th>
<th>Quantity</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Element Sets/Standards Reviewed</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emory Instances</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total Elements</strong></td>
<td>1340</td>
<td></td>
</tr>
<tr>
<td>Emory: Elements</td>
<td>245</td>
<td>18%</td>
</tr>
<tr>
<td><strong>Total Required Elements</strong></td>
<td>251</td>
<td>19%</td>
</tr>
<tr>
<td>Emory: Required Elements</td>
<td>85</td>
<td>35%</td>
</tr>
<tr>
<td><strong>Total Required if Applicable Elements</strong></td>
<td>291</td>
<td>22%</td>
</tr>
<tr>
<td>Emory: Required if Applicable</td>
<td>21</td>
<td>9%</td>
</tr>
<tr>
<td><em><em>Total Recommended</em> Elements</em>*</td>
<td>70</td>
<td>5%</td>
</tr>
<tr>
<td>Optional/Unspecified Elements</td>
<td>728</td>
<td>50%</td>
</tr>
<tr>
<td><strong>Recommended/Optional Elements</strong></td>
<td>740</td>
<td>55%</td>
</tr>
<tr>
<td>Emory: Recommended/Optional</td>
<td>139</td>
<td>57%</td>
</tr>
</tbody>
</table>
## Top Element Concepts/Tags - Clarified

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

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

*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*
# Top Elements – Compared Across Standards

<table>
<thead>
<tr>
<th>Element/Concept</th>
<th>ETD</th>
<th>OE</th>
<th>Keep</th>
<th>DAMS</th>
<th>DC LAP</th>
<th>MODS Aquifer</th>
<th>DDI Lite</th>
<th>ISO 19115 Core</th>
<th>TEI Header</th>
<th>EAD DACS</th>
<th>RDA Core</th>
<th>VRA Core</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>O</td>
</tr>
<tr>
<td>Identifier</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>O</td>
<td>R</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>M</td>
<td>C</td>
<td>O</td>
</tr>
<tr>
<td>Creator</td>
<td>M</td>
<td>M</td>
<td>C</td>
<td>C</td>
<td>O</td>
<td>C</td>
<td>O</td>
<td>?</td>
<td>R</td>
<td>C</td>
<td>C</td>
<td>O</td>
</tr>
<tr>
<td>Subject (aggregated)</td>
<td>M</td>
<td>O</td>
<td>?</td>
<td>O</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>O</td>
<td>M</td>
<td>O</td>
<td>R</td>
<td>?</td>
</tr>
<tr>
<td>Date (aggregated)</td>
<td>?</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>O</td>
<td>M</td>
<td>O</td>
<td>M</td>
<td>O</td>
<td>M</td>
<td>C</td>
<td>O</td>
</tr>
<tr>
<td>Subject/Topic/Keyword</td>
<td>M</td>
<td>O</td>
<td>?</td>
<td>O</td>
<td>C</td>
<td>C</td>
<td>O</td>
<td>M</td>
<td>O</td>
<td>R</td>
<td>?</td>
<td>R</td>
</tr>
<tr>
<td>Type</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>O</td>
<td>M</td>
<td>O</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>M</td>
</tr>
<tr>
<td>Rights/Access</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>C</td>
<td>M</td>
<td>O</td>
<td>?</td>
<td>O</td>
<td>M</td>
<td>?</td>
</tr>
<tr>
<td>Language</td>
<td>M</td>
<td>R</td>
<td>M</td>
<td>O</td>
<td>C</td>
<td>C</td>
<td>?</td>
<td>M</td>
<td>O</td>
<td>M</td>
<td>C</td>
<td>?</td>
</tr>
<tr>
<td>Contributor</td>
<td>M</td>
<td>O</td>
<td>M</td>
<td>N/A</td>
<td>C</td>
<td>C</td>
<td>O</td>
<td>?</td>
<td>O</td>
<td>M</td>
<td>C</td>
<td>O</td>
</tr>
<tr>
<td>Description</td>
<td>M</td>
<td>R</td>
<td>?</td>
<td>O</td>
<td>R</td>
<td>R</td>
<td>O</td>
<td>M</td>
<td>O</td>
<td>M</td>
<td>C</td>
<td>O</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Types of Sources Reviewed:</th>
<th>Types of Content Represented:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• National/international standards</td>
<td>• Still Images</td>
</tr>
<tr>
<td>• Technical/XML schema/XSD</td>
<td>• Manuscripts/archival collections</td>
</tr>
<tr>
<td>• Content standards/Element sets</td>
<td>• Audiovisual resources</td>
</tr>
<tr>
<td>• Web publishing/content syndication practices</td>
<td>• Articles/analyzed publications</td>
</tr>
<tr>
<td>• Scholarly Publishers</td>
<td>• Websites</td>
</tr>
<tr>
<td>• Major consortia/agencies</td>
<td>• Data sets</td>
</tr>
<tr>
<td>• Emory projects/repositories</td>
<td>• Geospatial data</td>
</tr>
<tr>
<td>• Other libraries</td>
<td>• Full text</td>
</tr>
</tbody>
</table>
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

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

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

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

<table>
<thead>
<tr>
<th>System</th>
<th>Options</th>
<th>Element/Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finding Aids, Dataverse, Open Emory, ETDs</td>
<td>Relevance</td>
<td>Mixed: N/A?</td>
</tr>
<tr>
<td>Dataverse</td>
<td>Global ID</td>
<td>Identifier</td>
</tr>
<tr>
<td>Dataverse, Primo</td>
<td>Most Recently Released; Production Date; Creation Date</td>
<td>Date; Date Created; Date Produced</td>
</tr>
<tr>
<td>Primo, Dataverse</td>
<td>Popularity; Most Downloaded</td>
<td>Usage Statistics</td>
</tr>
<tr>
<td>Primo, ETDs</td>
<td>Title</td>
<td>Title</td>
</tr>
<tr>
<td>Primo, ETDs</td>
<td>Author</td>
<td>Creator</td>
</tr>
</tbody>
</table>
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
  - PID/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)