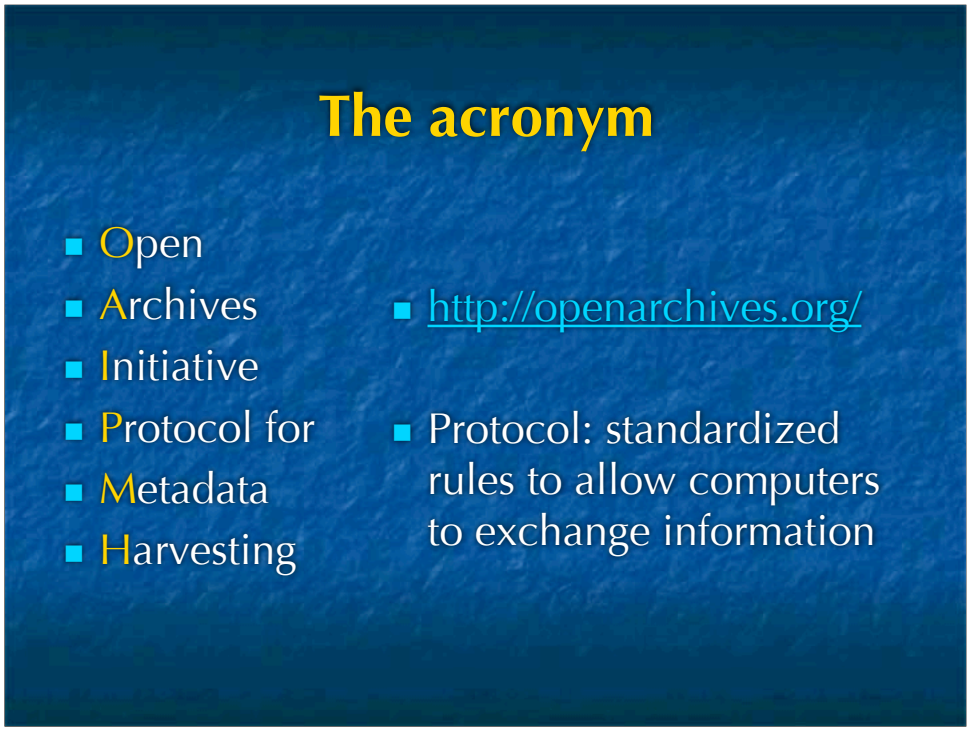
# OAI-PMH

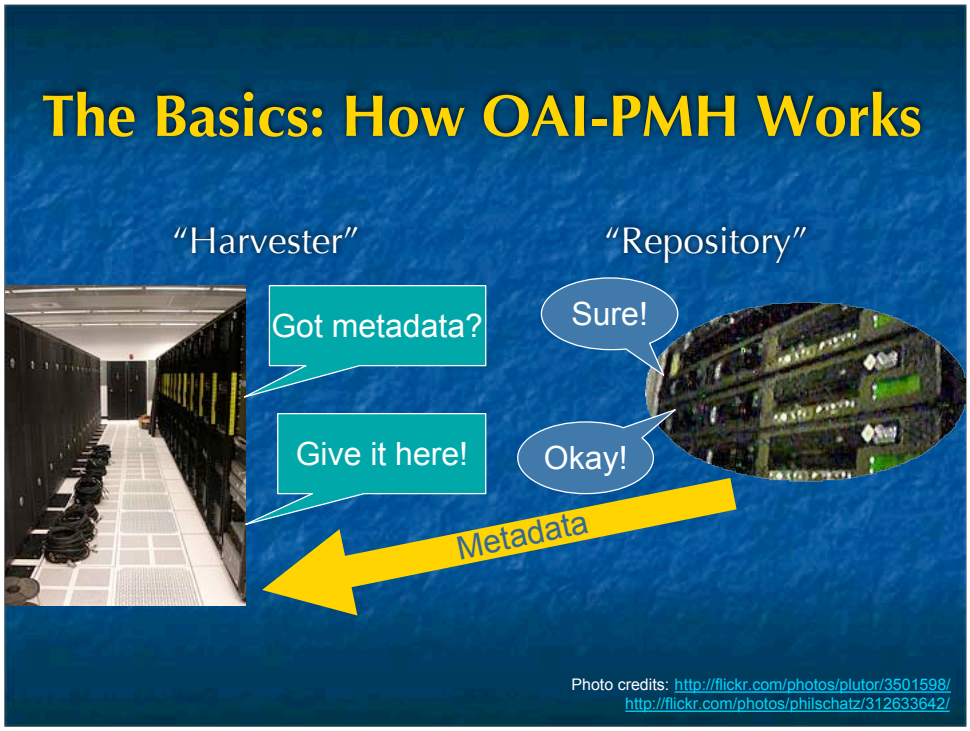
## Making our collections better known

Gail McMillan, Virginia Tech
Dorothea Salo, George Mason
26 January 2007

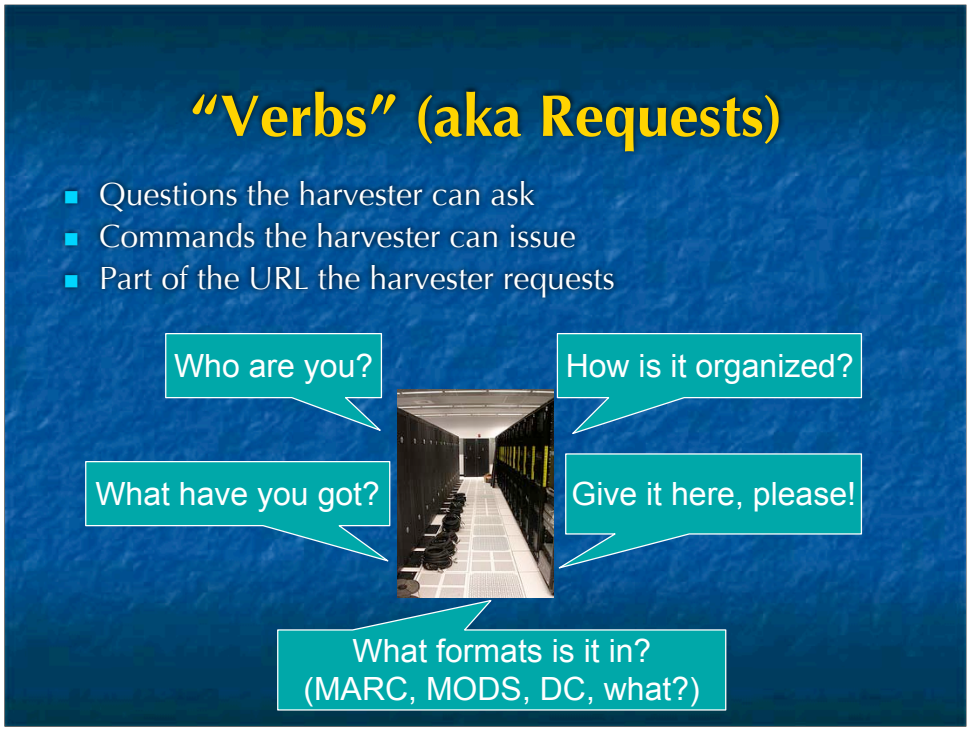



What does OAI-PMH stand for? It's the Open Archives Initiative Protocol for Metadata Harvesting. Taking that a bit at a time, the Open Archives Initiative is a techie think tank funded by grants and organizations like the Coalition for Networked Information and the Digital Library Federation to **help digital archives and repositories work better together.** Their first product was OAI-PMH.

Now, when normal people think of the word "**protocol**," they think of **etiquette rules to help people get along in a given situation, like Robert's Rules of Order** for meetings, or the protocol for meeting the Queen. In **techie-talk, a "protocol" is a set of rules that allow computers to pass information back and forth.** The most famous computer protocol is probably IP -- the Internet Protocol that allows everything from email to chat to web pages to flit around between servers and workstations and PCs. **OAI-PMH is another computer protocol, designed to allow computers to pass *metadata* back and forth.**



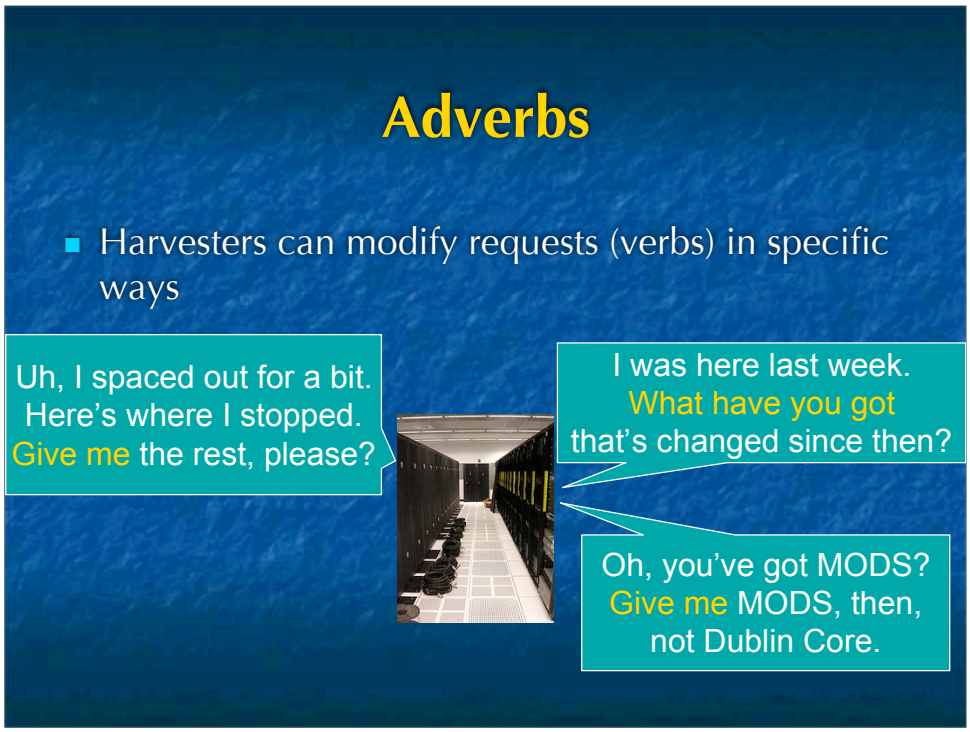

OAI-PMH distinguishes two roles that a computer can play: a "repository," which is a digital archive on the Web, and a "harvester," which acts a little bit like a search-engine crawler, picking up metadata from a whole lot of repositories. Maybe the harvester wants to create a portal (as in our case), maybe it wants to create a search index, whatever. In a basic OAI-PMH transaction, the harvester goes to a special URL on the repository server and asks the repository what metadata it's got. The repository tells it, and the harvester can then ask for the metadata it wants, which the repository hands over, or make available for harvesting. Note that the items themselves aren't changing hands. If you have a photo in your repository, the harvester can't ask for it; it only asks for your *description* of it.




According to OAI-PMH, a harvester can issue a limited set of requests to a repository, which the repository can understand and respond to. Basically, these are the questions and commands the harvester can use to get what it wants from the repository. There are six verbs in OAI-PMH; you only see five on the screen here because there's a short version and a long version of "What have you got?"

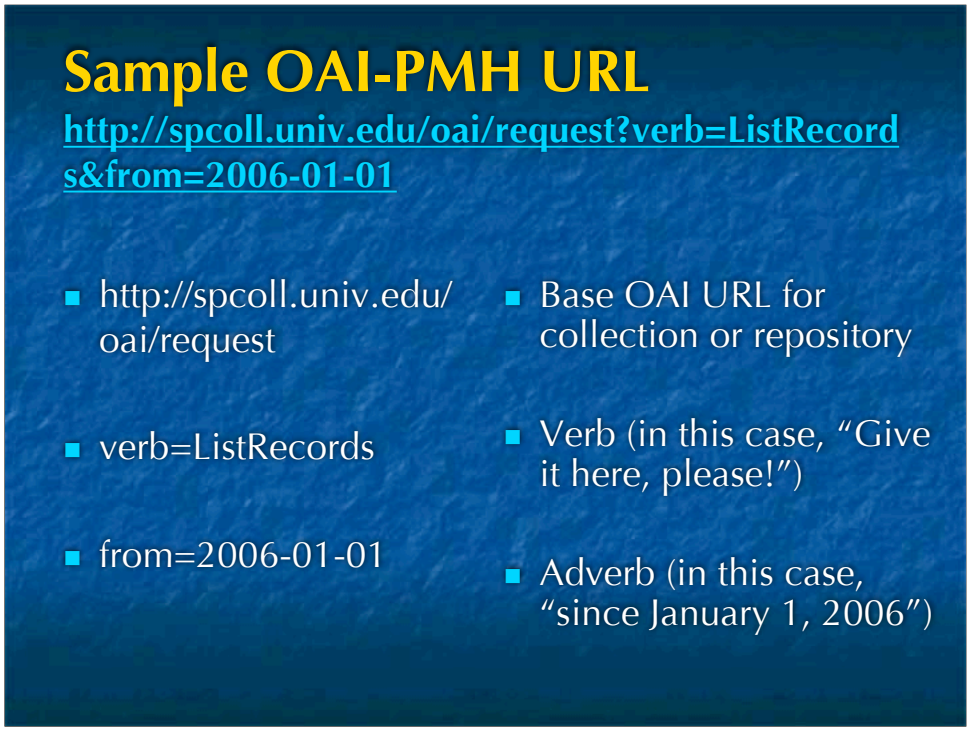
With regard to "How is it organized?" A repository can organize its content into "sets" if it likes, and a harvester can decide to harvest only certain sets. This can be quite useful. For example, a harvester that only wants metadata for ETDs can be told which set in a given repository contains them, and can harvest just that set, leaving everything else alone. Much politer than a Google spider!

Metadata formats: OAI-PMH is mostly format-agnostic. It allows computers to exchange metadata in any format that the repository has and the harvester understands. All repositories *must* have Dublin Core metadata; that is the agreed-upon base level. However, a repository can have MARC or MODS or METS or EAD or TEI headers or whatever in addition, and that's just fine.

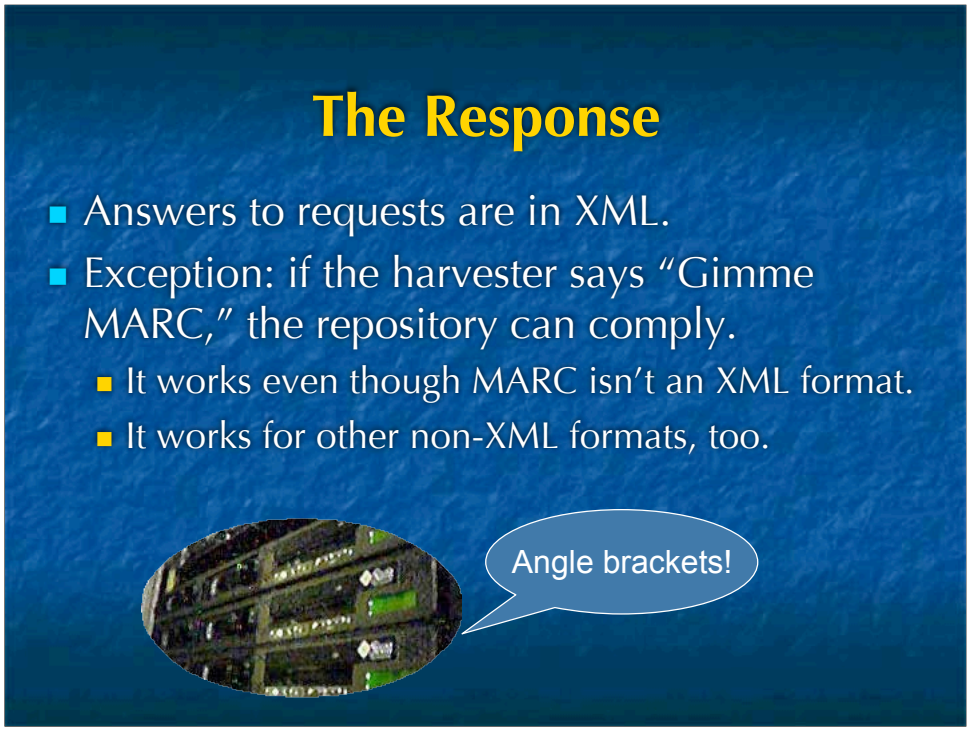



There are various ways that a harvester can modify the it requests to be more specific, or so that the response is more helpful. Dorothea likes to call these "adverbs" though that isn't OAI-PMH terminology. Generally these decrease the load on both harvester and repository, and allow problems caused by network issues or server flakiness or whatever to be fixed.

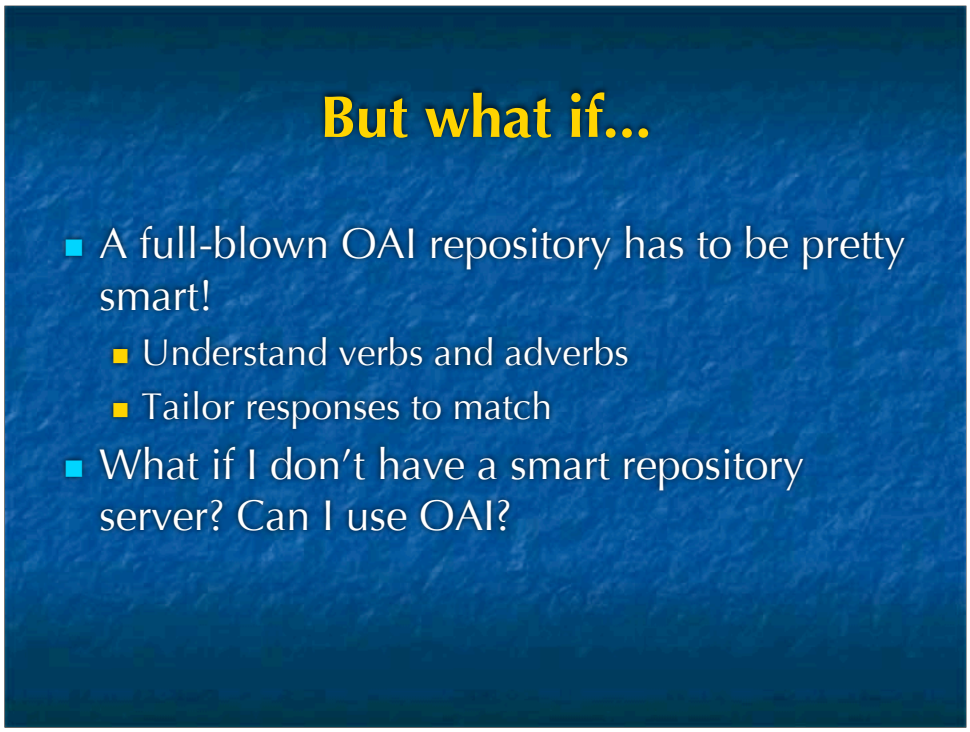
How does the harvester actually express the verbs and adverbs to the repository? It sticks them on the URL it asks for. The repository looks at the URL, and constructs a response based on the verbs and adverbs it finds. If it doesn't find a verb, it pitches a small fit.




Here's what one of those special URLs that the harvester uses might look like. GMU's Special Collections Department has a repository, and it accespts OAI requests at the URL http://dspace.univ.edu/oai/request. The harvester used the verb ListRecords, which is the "What have you got?" question, and it added an adverb "from," meaning "since the following date."




Usually, when a repository receives a request from a harvester, it will answer back with a little XML document. The specifics of what the XML looks like are laid out clearly in the OAI-PMH standard, but we don't need to know more about this complexity. The exception to this, however, is if the harvester asks for metadata in a non-XML format that the repository has, such as MARC. The repository can just hand over the MARC, no problem.

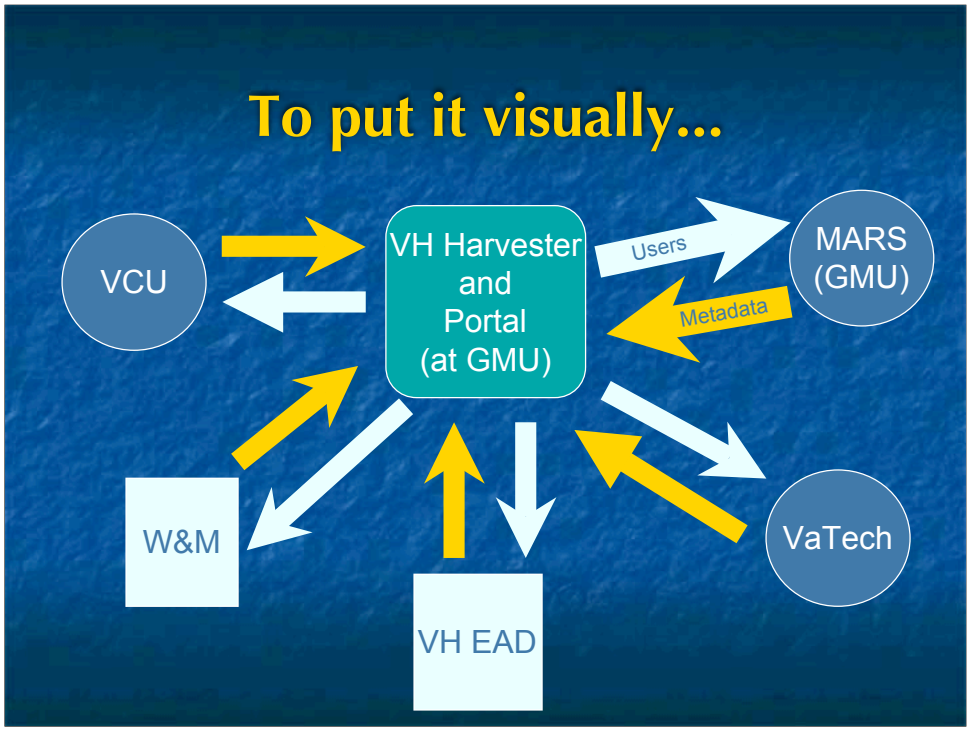



You don't have to have a fancy repository server or software to be an OAI repository.


Instead, you can create what's called a "static repository." This is a single XML file available via a single URL that contains all the Dublin Core metadata records for an entire digital archive. There are certain limitations to how a harvester can interact with a static repository; the repository can't organize its metadata into sets, and the harvester can't use any adverbs -- meaning it has no way to tell what's changed in a repository except by sucking down the entire set of records and comparing them! Still, this is a good option for relatively small and self-contained collections, and we expect a number of VIVA institutions to use it, including Virginia Tech which doesn't have an IR.

# OAI-PMH and the IMLS Grant

- VH Harvester
  - Will create a search and browse portal
  - Will be hosted at GMU
- Collections and Repositories
  - Created and hosted by participating institutions
  - Can grow over time!



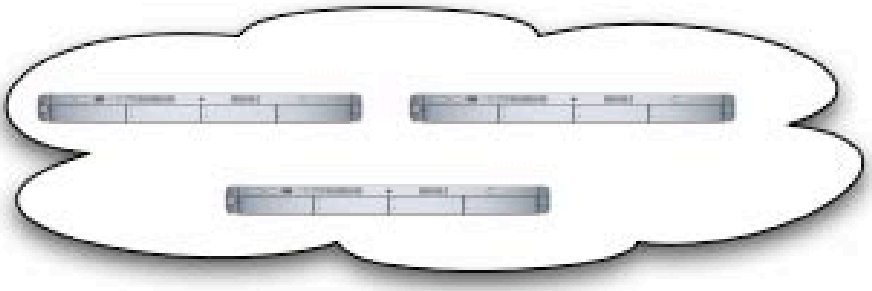

To put it visually, each participating institution will have its own collection or collections of digital objects related to the theme. Maybe the objects are in a DSpace repository, like MARS at GMU, or maybe they're in some other kind of repository, or maybe Radford has a web-accessible digital image collection. The specifics of storage are up to each institution. The harvester hosted at GMU will periodically ask each repository for its metadata. It will then put that metadata to work in a search and browse portal. When a user search from the to-be-developed VH portal turns up an item of interest, the portal sends the user to the site where the item lives in order to look at it.

The beauty of this arrangement is that institutions with broad collections can add to them all the time, and the portal gets updated automatically whenever the harvester comes back around; it's a very sustainable system. Institutions that want to have their own pages or search arrangements for their items can do so. Institutions that want to include these items in other portals can do so. Institutions that want to put items unrelated to VH in their repositories can do so; they just have to make sure to keep them in a separate collection from the one that our harvester is looking at.

# Any questions?

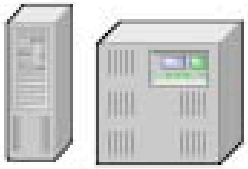
dsalo@gmu.edu
gailmac@vt.edu

Thanks!



IRs and other OAI-compliant Digital Collections

VHP (EAD Database)



OAI/EAD spider / index / database of appropriate EAD & IR (Dublin Core) metadata

User queries union index and follows links to appropriate item

1. Local institution creates and maintains metadata and hosts web-accessible digital objects.
2. Grant builds system to harvest and index database of OAI-compliant metadata and EAD Finding Aid metadata. Links in metadata enable access to digital content at local institution.
3. We leverage local work on IR/Digital Collections and VHP EAD database and present user with a single site providing access to content of both systems and preserve ability to search through local IR

Virginia Heritage: "Struggles for Freedom"

# IMLS Grant Briefing: Scanning Overview

Gary M. Worley
Director, Digital Imaging, Virginia Tech

VIVA Virtual Library of Virginia

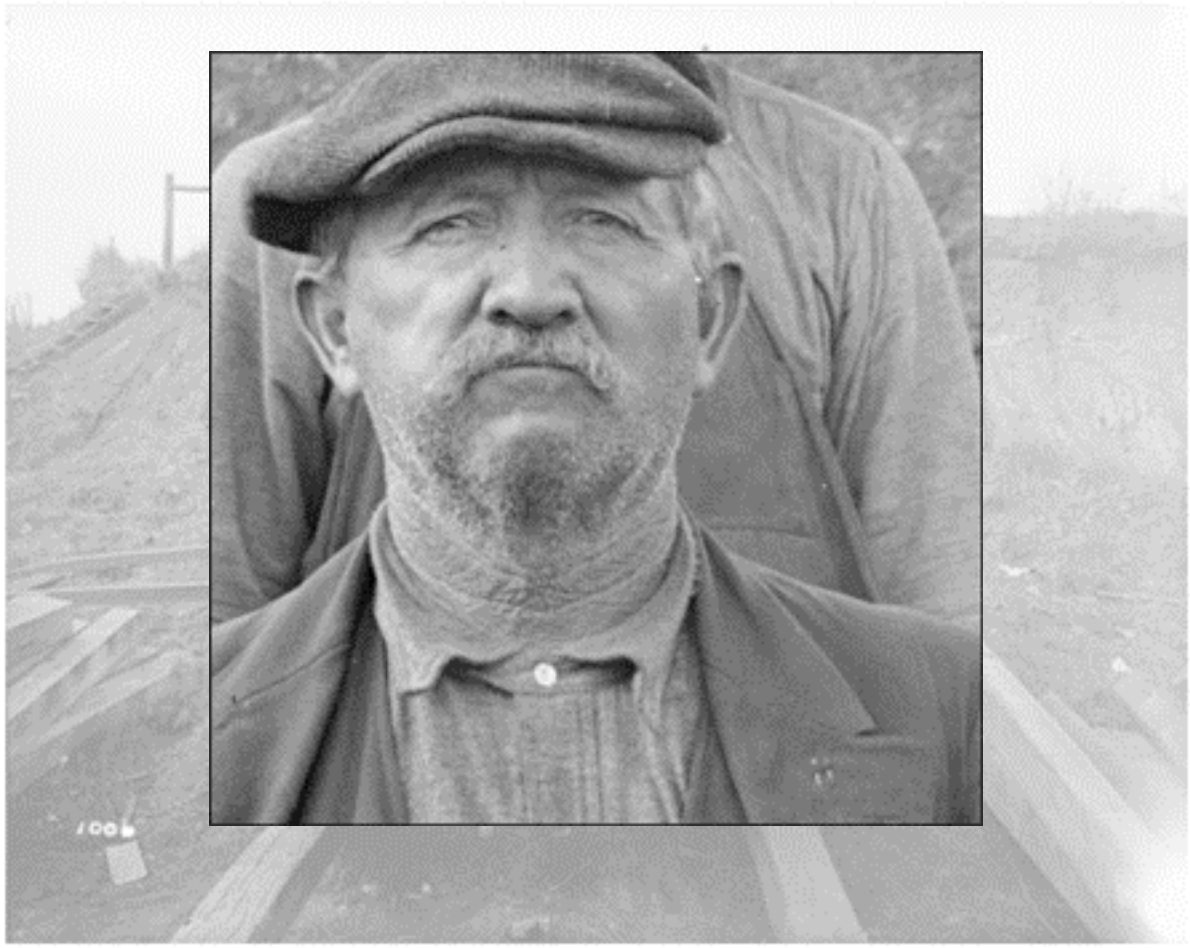
Virginia Heritage
Guides to Manuscript & Archival Collections in Virginia

# Scanning Overview

- Resolution and images

# Scanning Overview

# Scanning Overview



# Scanning Overview

# Scanning Overview

# Scanning Overview

Capture Standards for **Archival Masters**

| | |
|---|---|
| Resolution: | 300 ppi, minimum required* |
| Scale: | 1:1 for linear dimensions of the material |
| Bit Depth: | 24-bit color, 8-bit grayscale |
| File Format: | TIFF, tagged image file format |
| Color Space: | RGB color, grayscale |
| Compression: | None |

* 600 ppi allows for closer inspection of image detail. It is recommended that each collection of materials be evaluated prior to establishing a capture resolution setting.

# DIGITAL IMAGING CONTRACT

**Bailey-Law Collections**

University Libraries

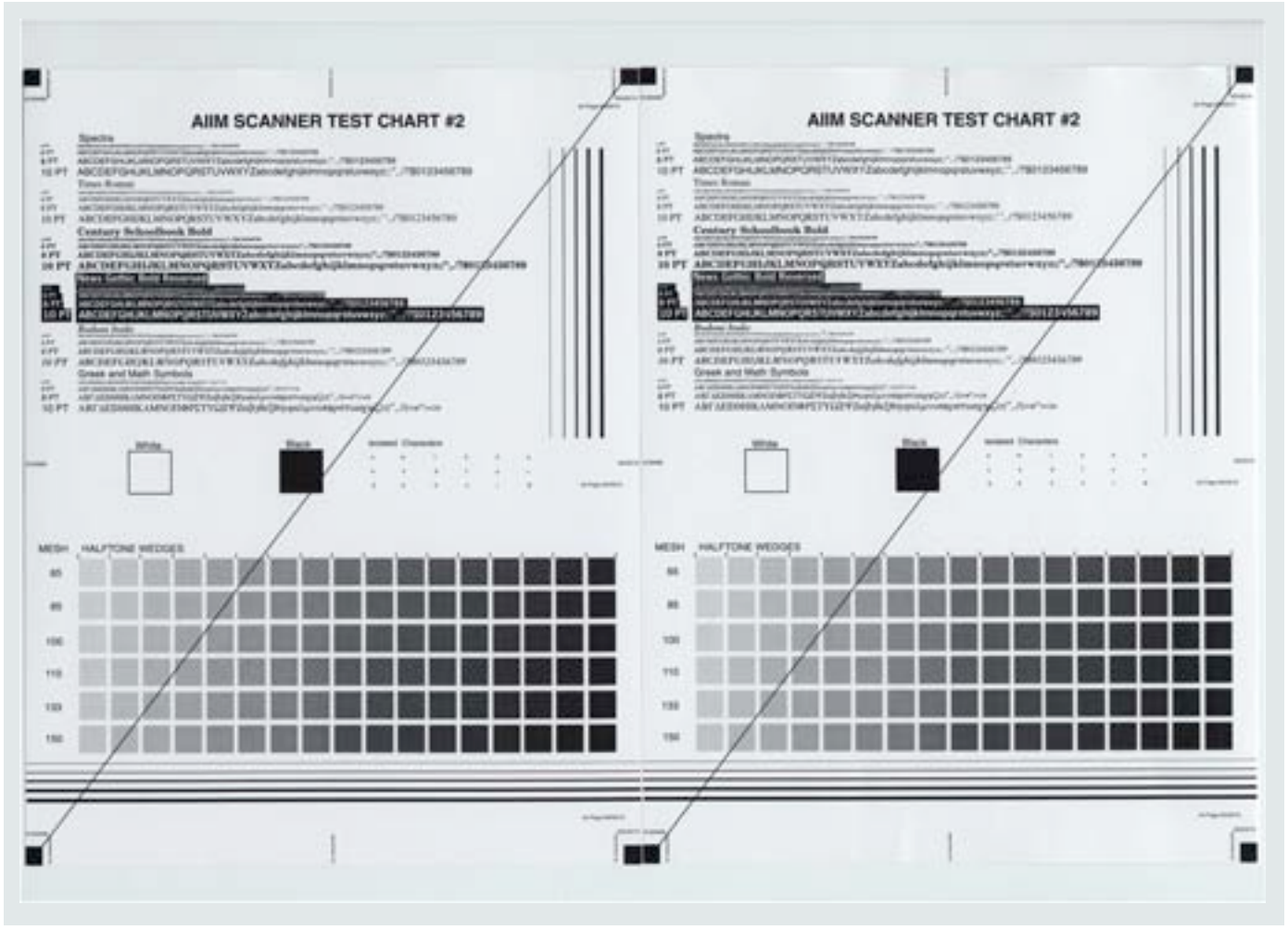
Active

| SOURCE | OUTPUT FORMAT |
|---|---|
| Paintings and Watercolors | TIFF Files; **300** ppi as specified for archival scanning from materials provided by the University Library. |
| | The objective for this project is to create a digital record of the original artwork representing the Bailey-Law Collections and to permanently archive the image data for future research involving this historic record. |
| | Our aim is 300 lines of descriptive points of resolution at a scale of 1:1 for the image data that accurately describes the linear information of the images while remaining true to the original size. The outside edge of the material should be preserved resulting in a digital image that extends slightly beyond the material image. And to create a file suitable for printing a duplicate image without the need to enlarge the dataset for the digitized materials. |
| Cropping | A small border should be left around the entire area of the map so the material edges are visible. The background material should be a bright white surface. |
| **HANDLING** | **All material handling for this collection requires the use of gloves.** |
| **FILE NAMING** | **SPECIFICATIONS** |
| | Example Name: **BLC**20070121102B.tif |
| | Name segment description |
| | Initials (as provided) — Date (yyyymmdd) |
| | Time (hhmm) — extension |
| **STORAGE** | **FOLDER ORGANIZATION** |
| | Primary = Collection — Example: Bailey-Law Collections |
| | Secondary = Specific — Example: E.L.Poole |
| | This collection consists mainly of large format Paintings and drawings. |
| | Scanned at **300 p**pi (one-to-one), saved as TIFF files |
| | Color or grayscale for polytonal pages using the Adobe 1998 color profile. |

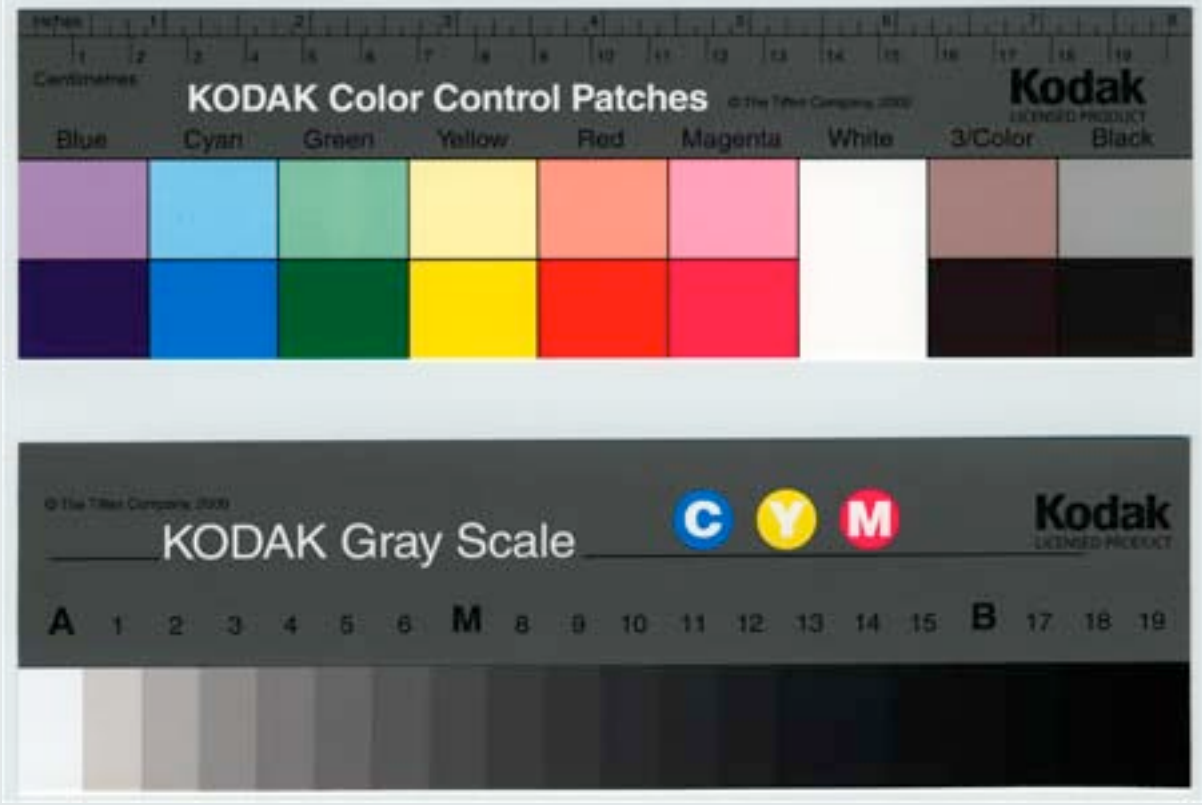
# Scanning Overview

- Resolution and images
- Benchmarks and quality control

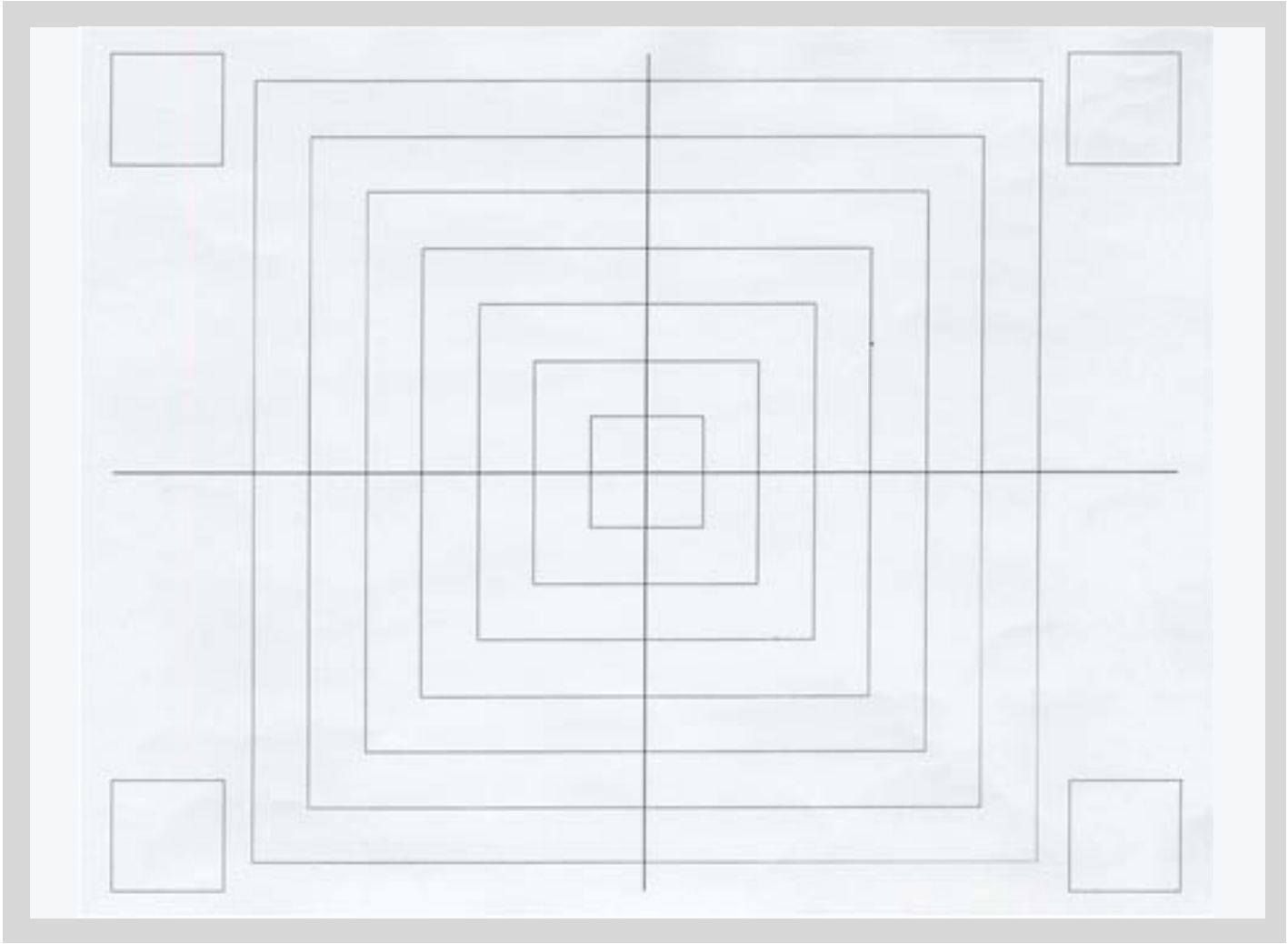
# Scanning Benchmarks



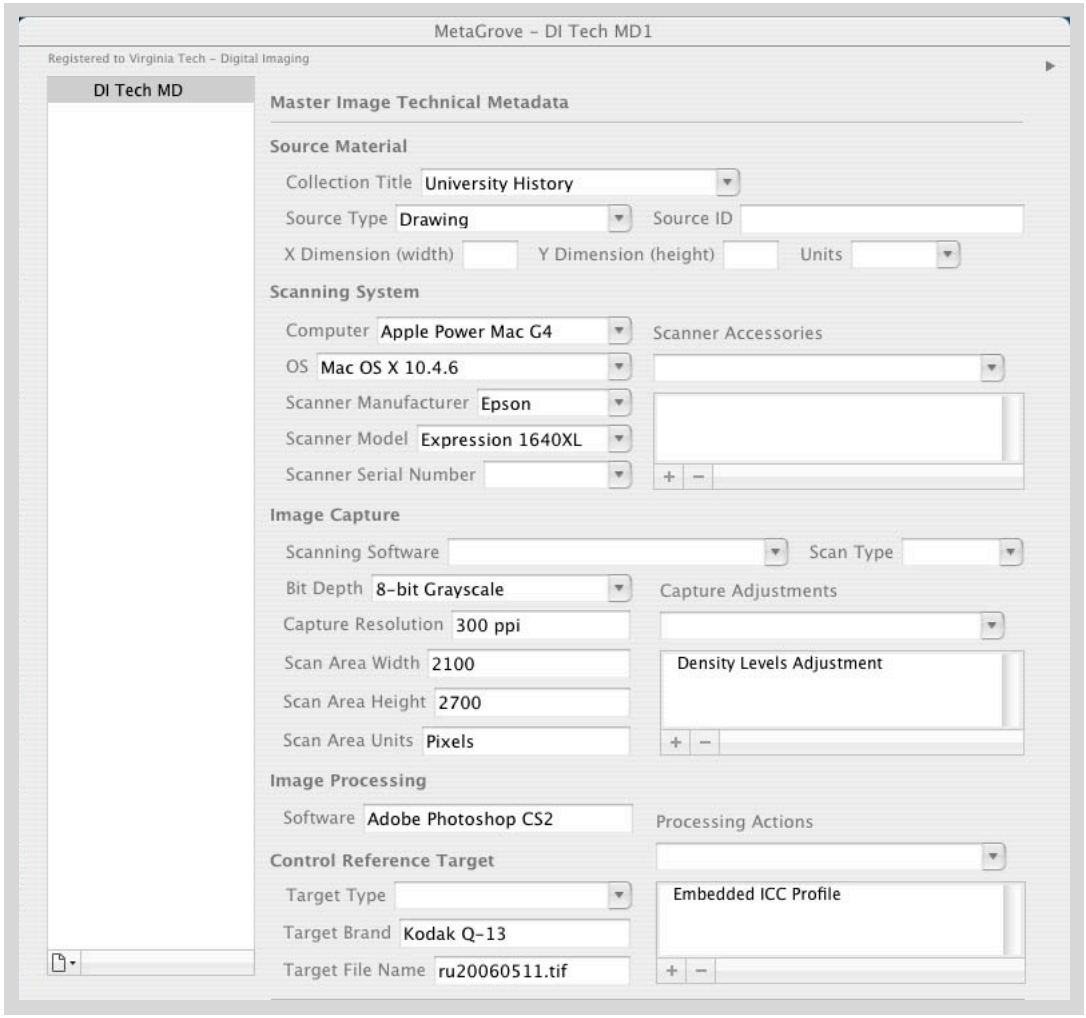
# Scanning Benchmarks



# Scanning Benchmarks



# Technical Metadata



# Training Activities

- Resolution and images
- Benchmarks and quality control
- Imaging exercises (Photoshop)
  - Equipment benchmarks
  - Scanning
  - Technical metadata entry
  - Optimizing images for web delivery

VIVA Special Collections Committee
GRANT MEETING January 26, 2007

# METADATA: The Who, What, Why, Where, and When

Bob Vay George Mason University

# Metadata:

Who?
What?
Why?
Where?
When?

# Who?
# Our “New Friend”

*and…*

# Our "Old Friend"

# What?

## Metadata Object Description Schema (MODS)

<title> <type> <genre> <origin> <language>

<physical description> <location> <access>

<preservation level> <object category> <file size> <format>  etc., etc…

*and…*

# Dublin Core Metadata Element Set

<contributor><coverage><creator><date>
<description><format><identifier><language>
<publisher><relation><rights><source>
<subject><title><type>

# Why?

The OAI protocol requires a Dublin Core record to be available with every item.

*but*

This *does not mean that one cannot use other metadata schemes* in addition to Dublin Core. OAI is designed to support records in multiple metadata formats for each item in a repository. An item can be exposed as a MODS, MARCXML, or Qualified Dublin Core record, as well as the required simple Dublin Core record.

# Where?

As part of your metadata in your item record. Most Digital Repository systems (ContentDM, Fedora, and D-Space to name a few) already have Dublin Core as its primary level metadata scheme, making them OAI-friendly. MODS metadata elements will be added to a baseline Dublin Core set.

# When?

- When you set up your metadata scheme in your repository system.
- When you create metadata in your item record.

# DLF MODS Implementation Guidelines Summary of Requirements

| Element | Required | Subelement(s) / Attribute required | Repeatable | Content Controlled |
|---|---|---|---|---|
| <titleInfo> | Yes | - One <title> subelement | Yes | No |
| <name> | No | N/A | Yes | No |
| <typeOfResource> | Yes | No | Yes | Yes (see guidelines) |
| <genre> | Yes | No | Yes | Recommended authority attribute limits content |
| <originInfo> | Yes | - At least one date subelement must have attribute keyDate="yes" | Yes | Recommended encoding attribute limits content |
| <language> | Yes, if language primary to resource | - Subelement <languageTerm><br>- type attribute required | Yes | Required attribute type="code" limits content |
| <physicalDescription> | Yes | - One subelement <digitalOrigin><br>- At least one subelement <internetMediaType> | No | Yes (see guidelines) |
| <abstract> | No | N/A | Yes | No |
| <tableOfContents> | No | N/A | Yes | No |
| <targetAudience> | No | N/A | Yes | Recommended authority attribute limits content |
| <note> | No | N/A | Yes | No |
| <subject> | Yes, if applicable | No | Yes | Recommended authority attribute limits content |
| <classification> | No | No | Yes | Recommended authority attribute limits content |
| <relatedItem> | No | No | Yes | In some cases (see guidelines) |

| | | | | |
|---|---|---|---|---|
| <targetAudience> | No | N/A | Yes | Recommended authority attribute limits content |
| <note> | No | N/A | Yes | No |
| <subject> | Yes, if applicable | No | Yes | Recommended authority attribute limits content |
| <classification> | No | No | Yes | Recommended authority attribute limits content |
| <relatedItem> | No | No | Yes | In some cases (see guidelines) |
| <identifier> | Yes | - type attribute required | Yes | Required type attribute limits content |
| <location> | Yes | - Subelement <url> is required for one and only one <location> element | <location> <url> is not repeatable | Yes |
| <accessCondition> | Yes | - Must use attribute type="use and reproduction" | No | No |
| <part> | No | No | Yes | No |
| <extension>* | No | N/A | N/A | N/A |
| <recordInfo> | Yes | - Subelement <languageOfCataloging> is required. | No | Required authority attribute limits content in some subelements |

*Not recommended for use


Questions? Comments?
General Grumbling?