Collection access: Describing, cataloging, and processing with the future in mind

Strategies for Advancing Hidden Collections Webinar 4
Presented by Beth Knazook

Live Webinar: Wednesday February 1, 2017

Joy Introductions & platform information (Joy) (2-3 minutes)

Hello! My name is Joy Banks, and I am the Project Coordinator for the CLIR Strategies for Advancing Hidden Collections six-part webinar series. Welcome to our fourth webinar, Collection access: Describing, cataloging, and processing with the future in mind. This series is offered through the generous support of The Andrew W. Mellon Foundation.

Please review the items in the housekeeping box for technical information. As a reminder, if you have any technical issues during the event, please send a private message to Louise Gruenberg. If you are viewing this as a group, please send a private message to me with your name, email, and group count. Please keep in mind that the webinar is being recorded, including the audio, slides, and chat. Recordings will be sent to the registered participants as soon as they are available.

[Joy will move everyone to Classroom] [Slide 1]

It is my pleasure to introduce our speaker for today, Beth Knazook. Beth trained as a photographic preservation specialist at Ryerson University/George Eastman House Museum of Photography, and has worked as the Curatorial Specialist for Ryerson University Library Special Collections and as the Photo Archivist for the Stratford Festival of Canada. She has been involved in a number of digitization and cataloging projects spanning different collections and descriptive standards, and teaches courses on the digitization and description of digital image collections for Library Juice Academy.

She is interested in how varied descriptive practices shape the interactions that researchers and community users have with photographic collections. She is currently working towards her PhD at Queen's University where her research focuses on the introduction of photographic illustration into Canadian book publishing in the mid-nineteenth century, and the complex relationship that developed between early photography and nineteenth-century print culture. Please welcome Beth.

[Slide 2]

Outline (3 min)

Introduction

1. Why do we catalog?
2. Choosing and using standards
3. Dealing with legacy data
4. Sharing records

Questions, Answers, Feedback

Thanks Joy and welcome everyone to the fourth webinar in the series. By now, you've covered a lot of ground in these webinars: you've talked about planning and managing the scope of your projects, how to budget successfully, and acquire and manage the people needed to carry out your goals (whether paid or volunteer) – Now we find ourselves at the stage where we need to talk about how are you actually going to do the work of processing and describing your resources so that you can share information about your collections. We'll be discussing good practices for structuring the form and content of your records mainly to optimize them for sharing on the web, but also to give your work longevity. We do have a **future-focus** in this webinar: employing descriptive standards allows us to create robust catalog records and descriptions that will be useable and understandable for a long time to come. I should mention at this point that, like with the other webinars in this series, the topic is a lot bigger than we have time for here so if there's anything that I don't explain in a lot of detail, please use the Q&A forum to ask questions that we can talk about at the end, and please make sure to have a look at the Resource Library after the webinar for a lot of detailed instructional material on specific standards and tools that I will be mentioning.

I've broken this class down into four main topics of conversation. I want to start with some broad thinking on cataloging and description to try and find some common ground to launch from. I'm sure that there are many different levels of expertise and experience in the room, and I think that we probably all have something to say on the topic. Next, we will discuss some practical information about established standards used in cataloging (and describing) resources across the GLAMS. Third, we'll talk about how to deal with those lingering, old catalog records in order to bring them up to these standards, and we'll finish by talking about how to share those records so that we achieve the visibility we want for our hidden collections. And that will be primarily focused on visibility on the web.

[Slide 3] Discussion (3-5 min): *What is cataloging? What makes it useful?*
I'd like to begin with a discussion about why we do what we do. What is it we are trying to accomplish by creating records about our collections?
I should note that although I am using the term 'cataloging', I mean this to encompass descriptive practices across GLAM organizations.

[Slide 4]
<u>Introduction to Metadata and Planning (15 min)</u>
Thank you all for a fantastic discussion. I hope that what we can take away from that discussion is that cataloging is a lot more than just data entry! Cataloging and descriptive practices generate the necessary information to [quote] "identify, authenticate, describe, locate and manage resources in a precise and consistent way that meets business, accountability, and archival requirements." It gives us the information that we need to be good stewards of our collections and to connect our communities to those collections. Now for those of you who are not as familiar with the term 'metadata', this refers to all the data that we generate in the course of our work. It is often referred to as 'data about data', or as we see it defined here by

the International Standards Organization, “data describing the context, content and structure of records and their management through time.”

[Slide 5]
Building from that definition, we can extrapolate some of the major benefits of cataloging:
Describes our collections to end users
It is an access tool, and often an interpretive tool. Increasingly, online metadata is replacing in-person interactions with our collections.
Supports our daily activities
It is also an administrative tool – not just description & access! It helps us to organize our materials, monitor them, and generally take better care of them. The more information we have about our collections, the more easily we can identify issues, seek conservation treatments, or adapt procedures. We are not truly accountable for our collections if we don’t keep records.
Connects our content to that of other institutions
It makes us relevant. It establishes where our usefulness and our authority lies, and thus also provide us with a means of reaching and impressing donors. If we don’t have information about our collections, how can we explain to anyone what we are doing?

[Slide 6] Discussion (3-5 min Joy starts timer): *What are some of the perceived barriers to cataloging? Is there anything about it that you struggle with?*
I’d like to take another short discussion break to talk about why we don’t always reap all these benefits. What are some of the perceived barriers to cataloging? (We may have had a few concerns slip through in the first session, cataloging doesn’t always get the recognition it deserves.)

[Slide 7]
Okay, so it appears we do have some valid concerns to work through. I think most of what we heard in the forum can be summed up with the following points:
Barriers to cataloging
Lack of trained/available staff to create records
Additionally, not enough time or resources available for remedying this.
Poor access to technology and technological support
We feel like we don’t have the right equipment to do the job, or afraid that technological investments will not have longevity. I think we’ve all been burned in the past over some piece of software or hardware that was discontinued or couldn’t be upgraded, and so we might be a bit wary of technology.
No connections or support from other institutions
I thought this would be the biggest one for this group. It’s very easy to feel invisible when you don’t have a network of colleagues and collaborators to bounce ideas off of. When we feel isolated, we get scared that we’re not doing it right, or that we’re falling behind the trends and that can be a very powerful inhibitor.

It’s understandable to be worried about these things, but as intimidating as these barriers

might seem, I want to assure you that these are not insurmountable.

[Slide 8]

What are some specific things we can we do to break these barriers down?

- **Meet staff at their skill level.** I think there's a common misconception that quality records are highly detailed. **Good metadata is, above all, consistent and reliable – not necessarily complicated.** Last week's webinar gave you some good ideas about how you can make the most of your people. If you have staff and volunteers without a lot of training, you can focus on training them to fill out a few fields well. Have examples available and give them the time and reassurance needed to boost their confidence. If staff time is a real issue, you might focus on crowdsourcing and use your staff to check and correct records rather than create them all from scratch.
- **Seek training opportunities and budget for them.**
  If there isn't enough money to support additional training so that you can get your catalog records to the desired level of completeness, can you budget for it for the future, whether by looking for grants or re-allocating funding from donations/fees/ etc. any of the sources discussed in the second webinar?
- **Start with the technology you have**
  (Do you have access to Microsoft Excel? Can you create tab-delimited files in a text editor? It is not as crucial as we might imagine it is to have the latest technology. Part of the reason why we are all trying to follow best practices and use metadata standards is so that we can manipulate, transform, and re-use our metadata in the future. Whatever technology you use now to create your catalog records will not be the technology that it lives in forever. As Angela mentioned in the first webinar, try to use programs that support data export in open formats like .csv or .xml. (or programs that create them in this format in the first place). Your database programs should allow you to export anything you've entered, and your community partners should also allow you to access and use the metadata you've created for collegial projects. I want you to remember than records and databases grow old and outdated, but good metadata is ageless. [And I want to note that I was in this boat, where I had several years of acquisitions and descriptive records in excel spreadsheets while working at Ryerson University, and I was able to import them into a new database quite easily.]
- **Seek community partnerships and be willing to join projects.** (Don't be afraid to try new things! Sometimes we shy away from opportunities because we are afraid that we don't have the expertise or that the project isn't going to be a good long-term investment, and that can prevent us from building helpful relationships and generating interesting metadata that we can then use for other things. Which brings me to my next point…)
- **Use social media and incorporate social tools.**
  Take advantage of the crowd! By allowing others to comment on your records and share your resources through the use of social media, you will help to make

your data more visible without requiring a lot of complicated in-house technology.

- **Create linked, open data.**
  Linking is one of the most powerful things that we can do on the web. Linking allows us to pull in information from other systems, and it establishes connections between our metadata and that provided by others. I can't stress this enough. Lateral connections build authority, visibility, and relevance, and they add a richness of information to your records that you may not have time to create in-house. And you might be surprised to find that you're already kind of been doing this if you're using controlled vocabularies or incorporating HTML links into your records, but we'll talk more about linked open data towards the end of the webinar.

  We are not all going to be working at the same level with our cataloging, but we can all do *something* to break through these barriers, and I hope that by the end of this webinar you're going to feel confident enough to make some choices and move forward with those choices. You don't need to possess significant technical skills to create high quality data, but you do need a plan.

[Slide 9]

Developing a plan for the future

Identifying the metadata that you need to capture is not as simple as looking to existing cataloging standards and just filling out all the fields. Institutional cataloging or metadata plans must encompass all the information that you want to share as well as the information you need to effectively manage your collections. As such, the guidelines that are created to govern cataloging practices are living documents that respond to an organization's changing needs, competencies, and outlook. These plans are enhanced through collaboration, both within your organization and with colleagues in the broader information and cultural heritage communities.

A good cataloging plan will provide guidelines that address these broad questions:

- **What do you need to know to manage and share your resources?**
- **What do your user communities need to know?**
- **How will you preserve your work?**

[Slide 10]

- You need three types of metadata about your collections to manage them effectively:
  - **Descriptive information**
  - **Administrative information**
  - **Technical/preservation information**

  I'm only touching on this briefly because in the next section we're going to cover the metadata standards that govern how we record this information.
- So moving on to the next question, When thinking about how you record your information, consider how your users will understand it:
  - **What is this resource?**

Describe your collections so that non-experts can understand the material quickly and easily. Explain what is important or interesting about them if you have that information available. Don't bury extra information in note fields or on another part of your website if it is relevant to the description. Remember that you are acting as an interpreter, particularly in those circumstances where there is no digital record to accompany the catalog record.

- **Who created this resource? Can I trust it?**
  Don't take for granted that your users will know whether you have a copy of something or hold the original, or that they will know that the resource belongs to you.
  Copyright information is one of the most important pieces of information that your users will look for on the web, and it is surprising how often this is overlooked or buried in websites. If you are digitizing content that it is in the public domain and you want people to use it, make sure they know that. If it's not fine for them to use, make sure they know that too. Don't hide this information somewhere obscure.

- **How will you preserve your work?**
  Like the real estate mantra, "Location, Location, Location" you should by now be getting a sense of the GLAM mantra, "Document, Document, Document."
  - **Documentation and workflows**
    Create guidelines describing, in detail, how all aspects of cataloging are to be carried out. You should outline all the required fields and content choices, who is in charge of adding information, approving records, what resources or controlled vocabularies are necessary to complete the work, etc.
  - **Use the standards and practices shared by GLAM communities**
    Don't re-invent the wheel unless it's absolutely necessary.

[Slide 11]
Your final plan can be developed in a similar three-part structure.

- **Outline how much information is available.**
  Be realistic about what it is possible to know about your materials, and how much you can rely on your catalogers to know, and reconcile this with what you know you need to manage your collections and your project.
- **Identify your user communities and determine their needs.**
  Conduct a needs assessment, survey your users, invite comments on your website – find out what your users want to see and what they need to know.
- **Determine desired output.**
  Select an established metadata standard that can capture all this metadata, and which supports the complexity of relationships demanded by your records (i.e. archives require a hierarchical arrangement).

Metadata Standards (20 min)
[Slide 12]

I've mentioned 'metadata standards' a few times now in this webinar so I think it's time we took a closer look at them. Metadata standards govern how we organize and present information about collections, and we use these standards primarily because standardized information promotes the exchange of information. This exchange is necessary because descriptive practices in GLAM organizations vary considerably across the fields (and even sometimes within individual institutions). Although we might record a lot of similar information, like title or date, we put the emphasis on different areas of knowledge. There's an importance placed on describing the provenance of materials in archives, while in museums there is also a lot of weight given to explaining what things *are*, as well as a need to track their intellectual history within the institution, like what exhibitions they've been a part of. There's a much greater emphasis on subject as an entry point in libraries, which is something we've only recently seen in archives and museums brought about by the need to provide subject access to digitized image collections. In more recent years, we've seen widespread adoption of standards designed to try and formalize the way that we communicate.

I'm going to start with a very simple example of why metadata standards exist for any participants who have never cataloged or encountered standards before [although according to the lobby poll it looks like _____]

[Slide 13]

Anne of Gr. Gables l.m. Montgomery 1908

This is information about a book in a library that I have communicated to you. Now, although I have succeeded in the task of conveying the correct information about this book, the way I've presented it renders it pretty much useless to most audiences. Unless you're the one who wrote this down, how can you know what this information describes?

A lot of what makes information useful is the way we *structure* and *display* it.

Title: Anne of Green Gables
Author: L.M. Montgomery
Date: 1908

Here we see that same information represented in a structured format that clearly shows us which pieces of information describe which aspects of the thing I described. The information has also been formatted so that the words are spelled out ("green" instead of "gr."), and the author's initials have been capitalized so it is clear that those letters represent a proper name. The use of field labels for structure, and proper spelling and capitalization for the content, has turned that useless string of data into understandable information.

The quality of our information also depends, to a certain degree, on transparency. In a lot of cases, we understand certain things about the objects we have in our hands that would not be clear to someone who is not also able to hold that object in their hands. If I were to add another piece of information to my catalog record here – something that cannot be transcribed from the object itself, but results from my understanding of the object - you'd know even more about what this collection item is:

Format: Book

(Sometimes we take the most obvious information for granted, and that might be okay if the record remains within our own databases, but when we exchange these records with other

systems that sort of really straightforward intellectual information will be the first thing that gets lost.)

Structured, controlled, and well-formed metadata makes people, concepts, and things distinct in our minds, but, it does much more than that too. It makes information about our collections useful for computers. Hollywood would like us to believe that computers are two steps away from taking over the planet, but they're really not that clever when it comes to understanding information. We have to give our databases and systems a lot of rules to yield productive results. Think about that first string of information I presented you with, and understand that as much as the human reader may have struggled to make sense of it, the computer has no idea what any of that means. Your computer might be able to return records based on keyword searches, because it can identify a string of numbers and letters that matches whatever you typed in, but you need to be give your records some context in order for databases to carry out sophisticated processing like sorting and aggregating:

[Slide 14]

Sort by title across holdings:

- **Anne of Green Gables**
- **Alice's Adventures in Wonderland**
- **James and the Giant Peach**
- **The Little Prince**

(It can sort by title because you've consistently put title information in the same field, and you asked it to search only that field.)

The computer can also find relationships between distributed information: It understood that the author name in each of these records referred to the same person. (Keep this in mind for later because the more precise your terminology within a field, the better a computer will be at finding relationships).

Find other books by the same author:

- **Anne of the Island**
- **Emily of New Moon**
- **The Story Girl**
- **A Tangled Web**

In short, the more consistent you are with the way you express something and where you put your information, the easier it is for *both* human and computer to make sense of what you've cataloged.

The need for structure and consistency applies to all types of records you might encounter as a cataloger. Some of you will be working in institutions where cataloging is an entirely separate task from acquisitions and processing, while other institutions may lump this kind of stuff together with cataloging (particularly if you're working in a one-person operation). To recap, you need three key types of information to manage your resources: Descriptive, Administrative and Technical (or Preservation). Descriptive information is meant to be seen by the public,

while administrative and technical metadata is mostly behind-the-scenes information. I want to address the behind-the-scenes records briefly first.

[Slide 15]
Example of administrative information:
Acquired: October 5, 2016
Appraisal value: $20
Barcode: 43954067560959685
There are many different manuals that have been produced over the years that can help you to decide what administrative information you will require to manage your collections materials, but it's mostly up to you how to structure this information and where to keep it. *This is generally not shared data, so as long as it is standardized internally, it can be used effectively in-house.*
Examples: *The Small Museums Cataloging Manual* by Museums Australia has guidelines on registration, naming conventions for image files; the *Standards for Archival Description* Handbook has a chapter on labeling and filing. These are examples of the kind of information that is generated for local use.

[Slide 16]
Example of technical (or preservation) information:
Digitized by: Beth Knazook
File format: JPEG
File size: 2 MB
Date: November 2, 2016
Checksum: 345GJODH5745607469$5697
Technical metadata is relatively new to the cataloging scene. This type of information pertains to digital objects, and it is recorded to ensure that digitized collections and born digital files retain their integrity and usability over time. ('Born digital' refers to those files that were created digitally and have no real-world counterpart – this powerpoint presentation is an example of a 'born digital' file.) The content on the left of the slide is an example of technical metadata recorded for a digitized photograph. As you can see, most of this information is pretty straightforward and can be found simply by clicking on the file on your desktop.

Technical metadata has a lot of overlap with preservation metadata, which is used to verify provenance and authenticity in digital objects. If you are planning on creating digital collections, you're going to want to take a closer look at some of the guidelines that have been developed to deal with this information, such as PREMIS data dictionary. The checksum field at the bottom left is an example of preservation metadata. A checksum must be generated by a computer program designed to create checksums, and it is used to monitor digital files for fixity (basically, to ensure the files haven't changed). You utilize checksums by generating them periodically over the life of a file. If the number you get from a subsequent checksum doesn't match the original number, that means the file has changed. It might be something as small as a single pixel that has been corrupted, but that's the whole point – the checksum will notice problems before you will and allow you to take action, either by copying the file to another device,

migrating it to a new format, etc.
Now, I will warn you that this is an area of cataloging that can get very technical very quickly, primarily when it comes to storing born digital collections, which tend to be our most fragile collections, BUT before there's any panic, I want to reassure you that no matter your level of expertise, you can record quality preservation metadata about your collections in adherence with best practices without relying too heavily on technical tools. (If you are interested in the technical tools, please check out the Resource Library for programs like BitCurator and Fixity.)

[Slide 17]
The purpose of preservation metadata is simply to answer the following questions:

- ***What is it?***
- ***Who created it?***
- ***Where did the information about it come from?***
- ***What can users do with this information?***
- ***How do I know that the information has not been altered?***

A lot of the technical tools that have been developed around digital preservation are designed to prevent objects from becoming altered, corrupted, or unusable, but that's just a tiny piece of the preservation pie. You can answer all these questions without resorting to technical tools at all. And in fact, if you have digital objects in your institutions, you might have been recording preservation metadata without knowing that was what you were doing.

Just like we want to know all we can about how an item wound up in our collections, we want to know all we can about how something digital was created. For instance, on the left is a brief overview of some of the metadata required by the PREMIS data dictionary published by the Library of Congress:

- *File Types, Dimensions, and Size* were also captured by technical metadata.
- *Inhibitors*: If the file is password protected or encrypted for whatever reason, they want you to describe this in your records.
- *Provenance:* Literally, where did the digital object come from? Did it come in on a CD, USB key, or the camera make and model that it was downloaded from. If the digital object has moved repositories or has been altered in any way, describe this. Particularly if this is a copy of the original file that has been made smaller.
- *Significant properties*: Are there any features of the document that are important to maintain, but might not be apparent without opening it? (e.g. this is an animated GIF and it should *move*)
- *Rights:* This is a standard field in most descriptive records, but PREMIS considers it to be of vital importance to the usability of the digital image. (Please, record clear rights statements now and your future selves will thank you.)

[Slide 18]
Like with administrative information, it is largely up to you how to record preservation information. The PREMIS data dictionary is a set of guidelines, not a rulebook. There are some descriptive structural schemas that have fields for preservation information, but there are

others that don't and so you get to decide where you're going to store this extra information. There are two popular options:

- **Metadata is stored within the digital file**
- **Metadata is stored in a database and linked to the file**

Let's look at the first option. When a digital file is created, a lot of technical details are stored in the file automatically anyway, and you can use any image editing program to embed copyright statements and even descriptive information.

[Slide 19]
This is an example of the file info dialogue box in Photoshop. You can add a surprising amount of information to this area, including copyright. Storing information in the digital file itself has some perks – the information travels with it, particularly if it is posted online! And if you don't have a database yet, this can help to manage your digital collections as keywords you enter will be searchable from your computer desktop file manager program. Just be aware that the more places you put catalog information, the more time you will spend cataloging. I would say that most institutions store this information elsewhere, or they use a combination of information in a database and minimal information in the digital file. And just a warning here: if you have a problem with digital file and you never copy any of this information elsewhere, you're going to lose both your file and your information about it.

[Slide 20]
This brings us, finally, to the kind of metadata that is governed by established standards: descriptive information. Producing good catalog records requires adherence to both structural standards and content standards. Structural standards (also called encoding standards) tell you what pieces of information you need to record. They determine the fields you will use in your database, and dictate what *type* of content belongs there. They also define the relationships and complexities of information. In EAD and CDWA, people or agents are described separately from their materials, while in MARC people or authors are added to a controlled index, but not described outside of the catalog record. They have different functionality.

Structural Metadata

- **Title**
- **Creator**
- **Date**
- **Description**
- **Rights**
- **Source**

Examples:

- **Libraries – MARC, MODS, METS**
- **Archives – EAD**
- **Museums/Galleries – CDWA**
- **Visual Resource libraries – VRA Core 4.0**

[Slide 21]
Content standards define *how* you should describe the information in a given field. For

instance, do you transcribe the title with all capitals? Do you use square brackets to indicate uncertainty? Do you record the date as month/day/year or write it out in words? Content standards create uniformity in the way that data is expressed, and they tend to be the standards that your catalogers will constantly refer back to when developing a descriptive record.

Most descriptive content standards were developed in conjunction with structural standards, which means that they pair well together. The structural standard CDWA works really well with CCO. MARC works well with RDA. It is possible to mix and match them, though you will need to watch out for conflicts between what the structure standard defines as the field information, and what the descriptive standard suggests you put in that field. *If you use a field for a purpose other than that for which it was designed, you may have problems exchanging data with others, or you may not be able to use some of the interesting tools developed to work with these standards. You may even have trouble migrating your data to a new database in the future. It is generally advisable to use the structural and descriptive standards that were designed to work together, and if you do make any local customizations, write them down so that those customizations are understood later on.*

Descriptive Content Metadata

- **Silver-plated kettle**
- **[ca. 1900]**
- **unknown**
- **Decorative hot water kettle with flowers, foliage and scrolls with shaped carrying handle.**
- **Public Domain**
- **Private collection**

Examples:

- **Libraries – AACR2, RDA**
- **Archives – DACS, RAD**
- **Museums/Galleries – CCO**
- **Visual Resource libraries – CCO**

[Slide 22]

How do you choose the right standards?

- **Schema appropriate for material type**
- **Schema appropriate for institution type**

Picking one of these pairs of standards is like joining a club. It ties you to a group of like-minded collections managers and it establishes a style of presentation that builds familiarity among users. The thinking is that someone who has learned how to use one archive will be better equipped to navigate another archive. But you don't *have* to use the schema that everyone else is using if another would be better suited to your collection needs. If you work in a museum, but you think an archival standard would be good for your collections of manuscripts and documents, then you might want to use EAD and DACS. Fit is key here though. Our end users expect to see materials explained in a certain way because of the way our institutions have positioned themselves, and we don't want to do something so radically different that our digital

collections don't reflect what we do, or adopt a schema in which we have to shoehorn information into fields that really weren't designed for our purposes.

[Slide 23]

- **Schema fits the needs of my collections**
- **Schema is open and extensible (can adapt to new uses)**

That said, a good metadata standard will allow you a certain level of customization while ensuring that your records can be migrated and exchanged with other systems. The demands of metadata are constantly changing, and whatever schema you pick should be able to adapt. All the schemas mentioned are syntactically interoperable – ie., they are all based on an XML framework (MARC has MARCXML). This means that if you decide one day that you really need to do something completely different, you can move your data from one schema to another. And that is why we feel confident that collections described according to standards will be sustainable, because these standards are not technology specific and you can manipulate and exchange them with other standards.

Dealing with Legacy Data (10 min)

[Slide 24]

This is a good point to move on to the topic of legacy data. Everything we've talked about so far in terms of standards and guidelines gives us an idea of where we want to be with our metadata – it projects our future direction. But what about the past? Many of us are dealing with leftover legacy data that makes us unsure if or how we can use older catalog records that may not have been created according to one of these standard schemas.

[Slide 25] Discussion (3-5min, Joy sets timer)

Louise reveal poll from lobby

From our poll at the start of the class I can see that [some][many] of you are dealing with old records so I'd like to take a few minutes so that we can talk about any concerns you might have on this topic. *Do you have concerns about legacy data/ past catalog records? Where do you currently store data (think of all types – do you have processing information on handwritten sheets in an office filing cabinet? If you have digital information, where is it all stored and in what format?*

[Slide 26]

Steps for evaluating and transforming legacy data

- Locate your data
- Create a metadata crosswalk
- Clean-up your data
- [Import it into a new database]

Okay, so I can see we have some concerns:

Interoperability, quality, reconciling different purposes for the metadata.

We can deal with these concerns by following the three (or four) steps outlined here. First,

identify all the available metadata. Then, build something called a crosswalk or data map, which essentially means that we evaluate field by field what metadata elements from our old catalog may or may not work in the new standard. If you encounter a lot of problems in this process, you can either decide to perhaps try a different standard, or you can take steps to edit or 'clean' the data from your old catalog so that it fits better. Chances are that you will have to do some minor clean-up no matter how well your crosswalk works, but you can decide how much effort to put into this stage.
I've put the last step in square brackets because not all of you will proceed to step four. Crosswalks have broader applications and can be incredibly useful tools for preparing your records to contribute to union catalogs and shared repositories, identifying metadata for harvesting, or to allow for search interoperability across systems. I've heard from the feedback on this webinar series however that many of you are looking to move your records into a new system, so we'll talk about databases briefly to address this.

[Slide 27]
Locate all your data
The first step before you can transform your records is to find all available records, and I'd like to emphasize this point because relevant and useful data about our collections can be found in a surprising number of places. It's easy to forget about or write off inventory lists you did 10 years ago that might contain a lot of good collections information. Even a Microsoft Word document can be transformed into a spreadsheet or .csv file and crosswalked to a database. I would locate and evaluate everything I can get my hands on before duplicating prior work in cataloging later.

[Slide 28]
Create a metadata crosswalk
Once you have all your metadata, you're going to begin creating a crosswalk or 'map' that will help you to see how your existing collections data fits (or doesn't fit) into your new standard. You can start this process with something as simple as a spreadsheet. List all the existing fields in your current database and consider how they have been used. Consult recent and past records to get a sense for whether or not they have been used consistently. Then take a look at the published guidelines available for the standard that you want to adopt. Now you're ready to map which field from the old catalog is equal to which field in the new standard. We have an example here showing which fields in the Dublin Core standard are equivalent to which fields in the VRA Core 4.0 standard.

Once you make your choices, be sure to document your reasons for these choices so that if you run into issues later, you'll be able to remember your reasoning and adjust as necessary. Documentation throughout this process is also incredibly useful to help you explain yourself later (whether you are explaining to fellow catalogers how to use the new system, or explaining to your administration or a grant organization how you plan to carry out the task of standardizing and updating your metadata).

Be prepared for the fact that some standards will map better than others.

[Slide 29]
If we continue mapping fields from Dublin Core to VRA Core, we're going to encounter some fields that don't map very well. GLAM organizations tend to use a lot of common pieces of information when describing collections, but they express that information differently and so the standards are constructed differently. Some fields might map to more than one field, while some might have nowhere obvious to map to.

Let's take a look at the first row. The Contributor field in Dublin Core is defined as "An entity responsible for making contributions to the resource." Agents in VRA Core are people, so it would make sense that Contributor would map to Agent, but the Location field in VRA can refer to institutions or corporate entities as well as geographic locations, therefore, it could be that Location is an appropriate field to map to as well. Determining which field is the most appropriate will all depend on how the Contributor field was used in Dublin Core. If it primarily contains information about the contributing institution, then it would make sense to ignore Agent as a possibility and map it to Location only.

[Slide 30] Louise Open Getty Crosswalk pop-up
I'm going to take you away from the Adobe Connect platform for a moment and over to a website containing a Metadata Standards Crosswalk. (You'll see something that looks like a really big table with some of the structural and descriptive standards we saw earlier at the head of the columns.) If you don't see this table, you can try copying the link in my powerpoint and pasting it into your browser manually.

This data crosswalk was produced by the J. Paul Getty Museum, to address the mapping of the CDWA standard to all the various other community standards used by GLAMS, both structural and descriptive: VRA Core, MARC, EAD, DACS, etc. This is a fantastic tool if you are already working with a metadata standard and you want to share your metadata with other institutions or contribute to shared repositories. It might also be helpful to consult if you are using different standards within your institution and you want to figure out if they can be amalgamated into a single database or queried by a single search engine. Finally, it might be useful if your old catalog records are kind of similar to an existing standard, and you want to see where there might be potential problems with mapping to a new standard.

[Slide 31]
Having looked at the structure, we need to figure out if the content is useful the way it was previously recorded.

A lot of editing of old catalog records can be automated through simple spreadsheet software. You can remove extra rows and duplicate records, do a find and replace on Unicode characters that didn't import properly (although be very careful with the find and replace option as it can find and replace things you weren't anticipating if you're not careful), and split cells with

multiple values.
Don't expect everything to automate perfectly. Make sure you do spot checks on records before you import or move on to the next step. Even if you are using a more robust data cleaning tool like Open Refine or Data Wrangler (these tools are linked in the Resource Library if you want to learn more about them), you'll have to be prepared to deal with data that just has to be cleaned by hand. In the last row we see a poorly composed description that has been greatly improved by re-writing. if you know you're going to have a lot of problems with certain fields from the old catalog, you can choose NOT to map these fields at all, but you will lose the information in the new records unless you copy them by hand.

[Slide 32]
Clean-up is not something that should be done on the fly. Before you undertake this process, review the catalog information with your colleagues and staff and clearly outline your expectations. Identify the staff members who will be responsible for certain steps, and clearly detail proposed actions and decisions.
(I'm showing you an example from one of the many spreadsheets that I used in my own past work in an instance where my department and another department were going to begin sharing a database and we had to come together to talk about different problems that might arise. As it turned out, one of the problems we encountered is that we were using duplicate accession numbers.)

[Slide 33]
Most of us will want to move our old data from existing databases to new databases, so let's spend a moment talking about purchasing a new database. How do you decide on the right system for storing all this information?

- **Supports selected structural/descriptive standard**
  Not all database vendors are clear about what standard they have based their field design on, if any. Others are more upfront. You might have to test out the software to be sure that it will work for you.
- **Provides appropriate fields for administrative and preservation metadata**
  Because these types of metadata are not governed by standards, there's no easy question to ask to determine if the database will work for you. Ask questions, and use the PREMIS or NISO data dictionaries alongside some of the manuals for archival/museum/library administration to figure out what you need.
- **Supports the OAI Protocol for Metadata Harvesting**
  The OAI protocol is an exchange protocol that has historically allowed you to expose your data to the web.
- **You have the technical support to install and implement the database**
  Often collections databases require more than just clicking a download icon and waiting for it to install itself. Server-based databases usually require additional software licenses to mount and run, and may require an IT staff person to maintain. Some companies offer technical support packages for an additional cost, which you'll want to be able to budget for if you need it.
- **It fits your budget**

You've heard this before in the previous webinars and it applies here too: Free isn't free. Open source software will have costs associated with server maintenance and customization. Vendor software might require costly upgrades. Ask questions of the vendors and developers, and of other institutions who have used the database.

- **There are exit strategies**
  If the company that manages the database goes under, what happens to your data? Will you be able to export it out of the database into one of those open file formats? If you don't upgrade, will you still receive support? If the data is being stored 'in the cloud', where exactly are the servers located and what procedures are in place to protect and preserve your data?

  In terms of actually picking a database, the Resource Library is a really great resource if you want some specific software recommendations.

Controlled Vocabularies and Linked Open Data (15 minutes)

[Slide 34]

The metadata that we create using structural and descriptive schemas promotes interoperability, but connecting distributed pieces of information across the internet requires additional effort. Remember the example from the start of the webinar, showing a list of book titles by the same author. Controlled vocabularies and linked open data allow us to define relationships across records, so that we can pull a vast amount of information together.

[Slide 35]

From Data Silos to the Web of Data

Keywords derived from controlled vocabularies and established indexes provide a much higher level of interoperability than standardized descriptive content alone. They help to avoid ambiguity between similar terms (iris for the eye or iris to describe a flower), they can give an official term for something (funiculars describe outdoor elevators that operate on inclines), and they can describe relationships between concepts (for instance, a Siamese cat is a more specific term than cat; and grapes are related to wine-making).

[Slide 36]

How do we create and add controlled terms to our records?

- **Work from an existing vocabulary resource:** Controlled terms are most often copied from an existing vocabulary resource or index. Certain communities have developed extensive thesauri that serve their specific needs. The *Getty Art & Architecture Thesaurus*, for instance, provides detailed terms for describing artistic concepts, periods, genres, materials and techniques. Another type of controlled list is available through the Library of Congress, which provides an extensive list of published author names in their Name Authority index. You can access these lists online (for free or with a subscription) and in some cases you can purchase a database that has these controlled terms pre-loaded so you just have to select the appropriate one. Your descriptive standard may suggest certain vocabularies to use with certain fields, or you might have the option of

using several. If you are uncertain which vocabularies might be right for you, you can use the websites to clarify their purpose. Bear in mind that they are not all designed to be used for people and subject terms, they can describe location, physical description, time periods, etc.

- **Create a local vocabulary list**: Perhaps your collection material is really unique or esoteric, or there simply aren't words in the English language to describe some materials. Starting from scratch is an incredible amount of work, but may ultimately serve your purposes better than using an established vocabulary. A good compromise however, is to use an established vocabulary *alongside* local terms. Clearly identify your local terms as unique to your institution, and use them to add breadth to your resource.
- **Create terms as-you-go from guidelines**: We know that our descriptive metadata schemas define the types of information that go into certain fields, and sometimes they are so specific about the phrasing of that information it effectively creates a local controlled list. In DACS there are rules for composing person access points for records. An access point is an index term, or a controlled entry point into the records. Most archival repositories do not use controlled vocabularies for persons because it is far too difficult to anticipate what names will be needed.

[Slide 37]

Using a Controlled Vocabulary

Each of the published vocabulary resources have their own quirks, and I cannot instruct you on how to use all of them, but I can outline some general guidelines on how to use them:

- Read scope notes / observe the hierarchy to confirm that the term selected is the desired one
    - Sometimes vocabularies are just lists, but most have a scope note that describes what terms mean so that you can decide if the one you've selected is an accurate choice.
- Use the preferred form of the term
    - Just because you can look something up in the vocabulary doesn't mean that you should use that term. There is always a preferred term identified, and if you haven't landed on it in the first try, you'll see some note directing you to use the preferred term instead.
- Use the narrowest term applicable
    - If the vocabulary is arranged hierarchically with broader and narrower terms, use the most specific one.

[Slide 38]

This is a screenshot from the Getty Thesaurus of Geographic Names, for a town called Brazil in Tennessee. I can see from the hierarchy that this location is not the nation of Brazil in South America, so if I wanted that Brazil, I could back out and look for another term. I can also see that Brazil is the more specific term than Tennessee, and that Brazil is the preferred form of the name and should be used instead of Poplar Grove.

[Slide 39]

Linked open data is sort of the next generation of controlled vocabularies - it is a means of connecting anything to anything else – concepts, people, places.
This is accomplished through something called Resource Description Framework (RDF). RDF creates links between content on the web, using Uniform Resource Identifiers (URIs) and HTML, but unlike plain HTML, RDF describes the *context* of those links.

[Slide 40]
Here's one way to imagine how linked open data works:
This statuette of a seated lion was made in Sparta, Greece around 550 B.C.
https://en.wikipedia.org/wiki/Sparta
Here we have a little statuette of a lion from the Getty Museum, and there's a link that directs us to information that pertains to this lion. This is the sort of HTML link we might encounter on any website. We can read the wording of this link and infer what the relationship is, but sometimes links are not so self-explanatory. And without RDF, they're always a mystery to the computer. An HTML link does not establish the relationship between webpages. That's where RDF comes in.

[Slide 41]
If we imagine that the lion, the place it was created, and the concept of 'place created' are all things we can link to, we have the basic idea behind Linked Open Data. RDF triples define contextual relationships between content on the web, and they're the reason why the web is getting smarter.

Not everyone is going to be in a position to utilize Linked Open Data right now, but that does not mean that you cannot create the metadata that will be useful for these types of projects somewhere down the road. Controlled vocabularies are at the heart of LOD, and you can start using those much more easily. There are data tools out there that are being developed to process structured and controlled metadata and convert that metadata to RDF, such as Ontomaton https://github.com/ISA-tools/OntoMaton and OntoWiki http://ontowiki.net. (These were developed primarily for research data, but there are new tools being developed every day!)

[Slide 42]
Activity
"Tag, You're It!" (1 min in classroom) (Joy sets timer)
Having put such an emphasis on the value of control, I'm now going to switch it up and talk about uncontrolled keywording. This is a 1 minute exercise, so don't spend too much time thinking about your work. Use the chat box on the right to type whatever words come to mind when looking at this picture. Remember, a picture is worth 1,000 words!

As I hope you can see from this exercise, controlled vocabularies clearly have their place, but they also have their limitations. They are often not exhaustive, and despite the use of technical and specific terms, are never as narrowly specific as the keywords that might be dreamed up by a whole crowd of people.

User-added descriptors (like word tags) allow our users to tell us what we missed that might be important to them. We might shudder at the thought of purposely allowing user typos or inaccuracies into our descriptive records, but consider that not everyone will know how to articulate what they are looking for, particularly if they do not understand the catalog information same way the cataloguer does. Chances are your users have *some* words to express what they mean that are different than yours. Connecting those users to our records might be as simple as employing the crowd.

[Slide 43]
In the previous webinar, Sarah mentioned that crowdsourcing projects are a great way to engage a lot of volunteers. Tagging collections is an excellent, simple task to give this kind of diverse group.
Options for harnessing the crowd:

- **Purchase a database with tagging/social media tools built-in or develop your own.**
  A lot of databases come with this option built-in now, so it might simply be a matter of allowing it to happen. Ideally you'll vet the tags so that inappropriate content isn't posted to your website, so there is some consideration here for staff time.
- **Allow access to social bookmarking tools on your website such as del.icio.us, pinterest or LibraryThing**
  You can download the code for their buttons and paste them into your website - and then watch and see what happens!
- **Put an email address at the bottom of each record inviting questions.**
  This wouldn't be a good tool for a blitz cataloging project, but inviting email using encouraging language like "Have a comment you'd like to share? We'd love to hear it!" is another way of connecting with the crowd. This is the easiest solution, but may require the most staff time as users will expect a response.
- **Put your content on a third-party website that supports comments and tagging.**

[Slide 44]
A surprising number of well-respected institutions are using a variety of social media tools to reach audiences and solicit feedback. The Smithsonian is on Flickr, the Museum of Modern Art is on Instagram, The Library of Congress is on Facebook – and a lot of smaller institutions are too. These are more than just tools for announcing events and attracting visitors. They can also be used to understand how your users are engaging with the catalog information you put out there, and where there might be gaps in information. Bear in mind that social media is most useful when a staff member is paying attention to these accounts to see what people are saying/doing with the content.

[Slide 45]
A case in point:
I've compiled a couple of screenshots of an uploaded picture on the Smithsonian's Flickr page. There are a number of tags here that complement the Smithsonian's brief record, and make this image discoverable on Flickr. Now, not all the tags are accurate, and that's a risk you take

by opening it up to the crowd. However, in prepping this webinar it was pointed out to me that the cataloger information is not accurate in this record, and so in this case we see how social media can also be a great feedback mechanism for attracting expert opinions and corrections. Louise (you might want to say something or type/ whatever you prefer).

Based on this feedback, the cataloger could enhance this entry and provide the end user with the information needed to understand which of the tags are correct.

Conclusion (1 min)

[Slide 46]

What does the future hold?

Everyone here is at different stages and levels of expertise in their cataloging, but hopefully you are all beginning to see answers forming to those planning questions I posed at the start of the webinar. Whether you are only at the stage where you can create spreadsheets, you can begin to use standards and controlled vocabularies that will allow you to easily upload, use, and share that information later. If you have great records but aren't sure how to begin linking them, you can look at some of the RDF conversion tools and social media tools that will allow you to reach out broadly.

[Slide 47] Beth's contact information and Link pod with survey

Joy

Thank you Beth and everyone for a great session. Please be sure to complete the webinar evaluation while the content is fresh in your mind. You can access the evaluation using the link visible on the screen or wait to be redirected when the webinar is ended. The evaluation link will also be available on the website with the recordings. Live participants will also receive the link with their email containing the access link for next week's session. Please join us at the same time and place next week for the fifth webinar in our series: Overcoming project hurdles: Approaches to identifying and managing collection red flags. Have a great day!