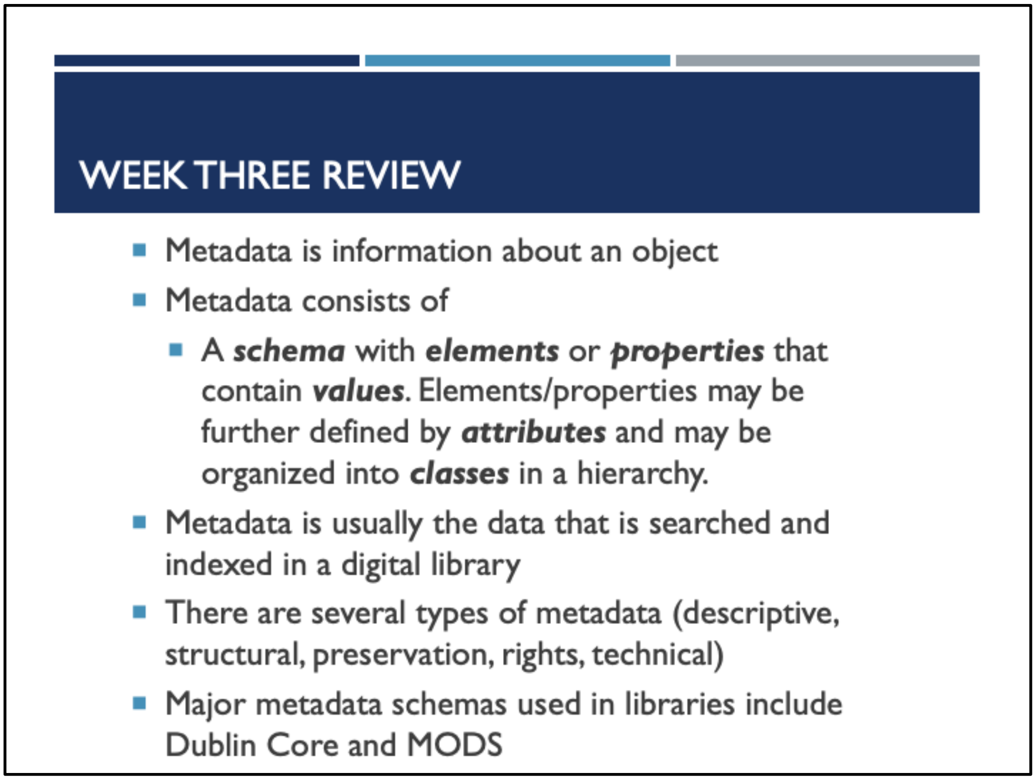
# WEEK FIVE
# METADATA: SHARING

PART TWO: SHARING AND APPLYING

We’re going to expand our understanding of metadata this week to talk about how it can be shared and applied in even more interesting ways.



## WEEK THREE REVIEW

- Metadata is information about an object
- Metadata consists of
  - A ***schema*** with ***elements*** or ***properties*** that contain ***values***. Elements/properties may be further defined by ***attributes*** and may be organized into ***classes*** in a hierarchy.
- Metadata is usually the data that is searched and indexed in a digital library
- There are several types of metadata (descriptive, structural, preservation, rights, technical)
- Major metadata schemas used in libraries include Dublin Core and MODS

Let’s just quickly review what we talked about last week:

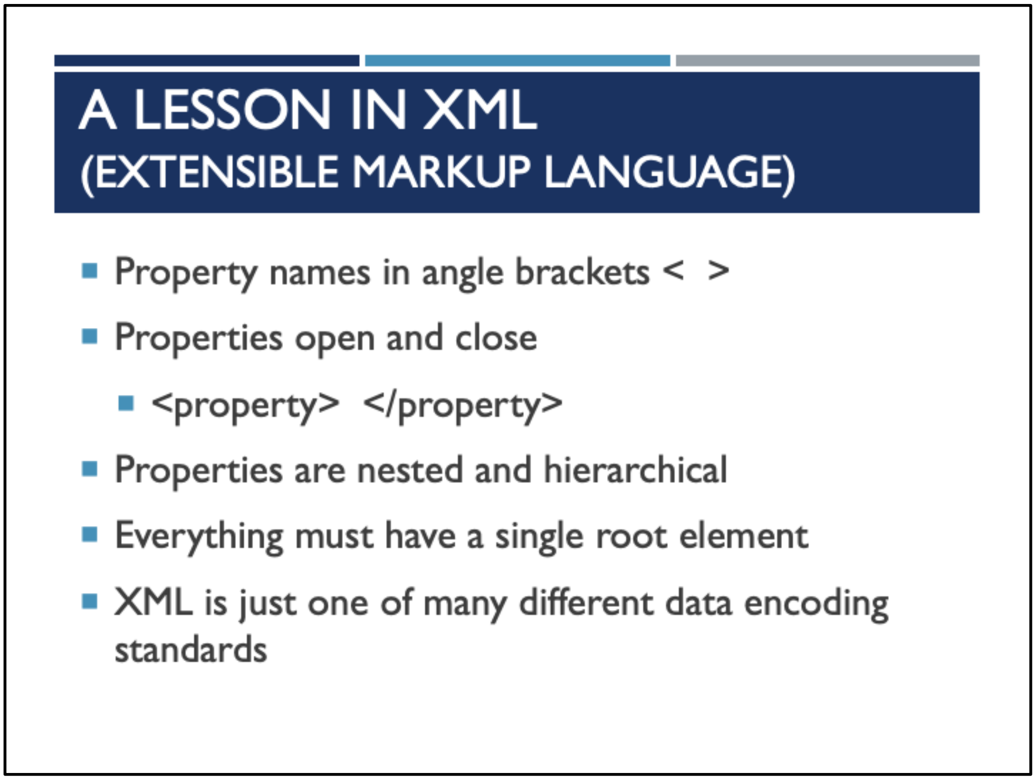
Metadata is information about an object

Metadata is organized into **schemas**, which are sets of **elements** or **properties** that are used to describe characteristics of an object. Similar elements can be differentiated by **attributes**, **hierarchy**, or membership of a **class**

Metadata is what is searched in a digital library

There are several types of metadata including **descriptive, structural,** and **administrative types: preservation, rights, technical**

Some major metadata schema used by libraries include **Dublin Core**, and **MODS**,



## A LESSON IN XML (EXTENSIBLE MARKUP LANGUAGE)

- Property names in angle brackets < >
- Properties open and close
  - <property> </property>
- Properties are nested and hierarchical
- Everything must have a single root element
- XML is just one of many different data encoding standards

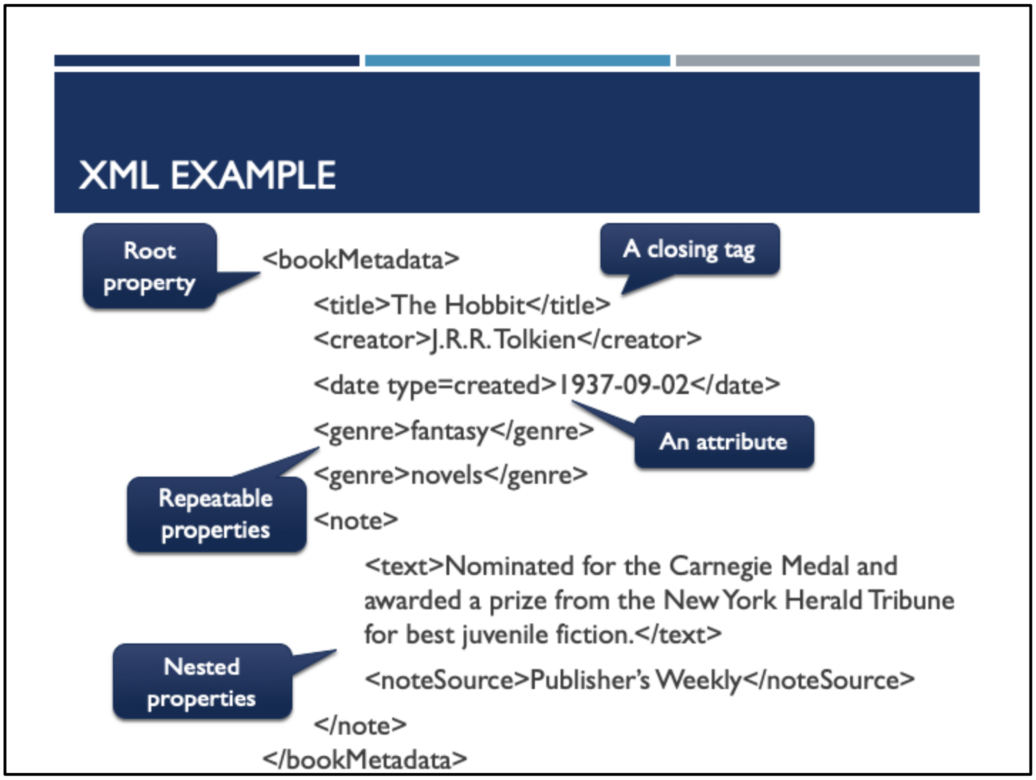
Before we move on I want to take a little side-track and talk a little bit about **XML**, which stands for eXtensible Markup Language.

XML is basically a protocol, or set of rules, about how you can format text in a way that certain software can understand. The reason we are learning about it is that a lot of metadata in the library community is shared in an XML format.

The rules of XML have you put property names inside angle brackets. The property name is in an angle bracket right before the text of the value. When the text of the value is finished, the property name is repeated again, but this time with a slash. So the properties wrap around the text of the value.

In XML, properties can be nested and hierarchical, so you can have that idea of hierarchy or classes, like we've talked about. But every XML document must have an initial, or parent, **root element**, that will open and close the file.

XML is just one of many different data encoding standards. There are other sets of rules that might dictate how to format metadata slightly differently for different types of programs to read (kind of like how there are different citation standards for different disciplines that all dictate how to format the same type of information in a citation). XML is a very common format used for metadata in libraries.



## XML EXAMPLE

```
<bookMetadata>
    <title>The Hobbit</title>
    <creator>J.R.R. Tolkien</creator>
    <date type=created>1937-09-02</date>
    <genre>fantasy</genre>
    <genre>novels</genre>
    <note>
        <text>Nominated for the Carnegie Medal and
        awarded a prize from the New York Herald Tribune
        for best juvenile fiction.</text>
        <noteSource>Publisher's Weekly</noteSource>
    </note>
</bookMetadata>
```

Root property
A closing tag
An attribute
Repeatable properties
Nested properties

So here’s a look at some XML.

Book metadata is the root property. It is enclosed in angle brackets. The property inside it’s angle bracket is often referred to in XML as a **tag**.

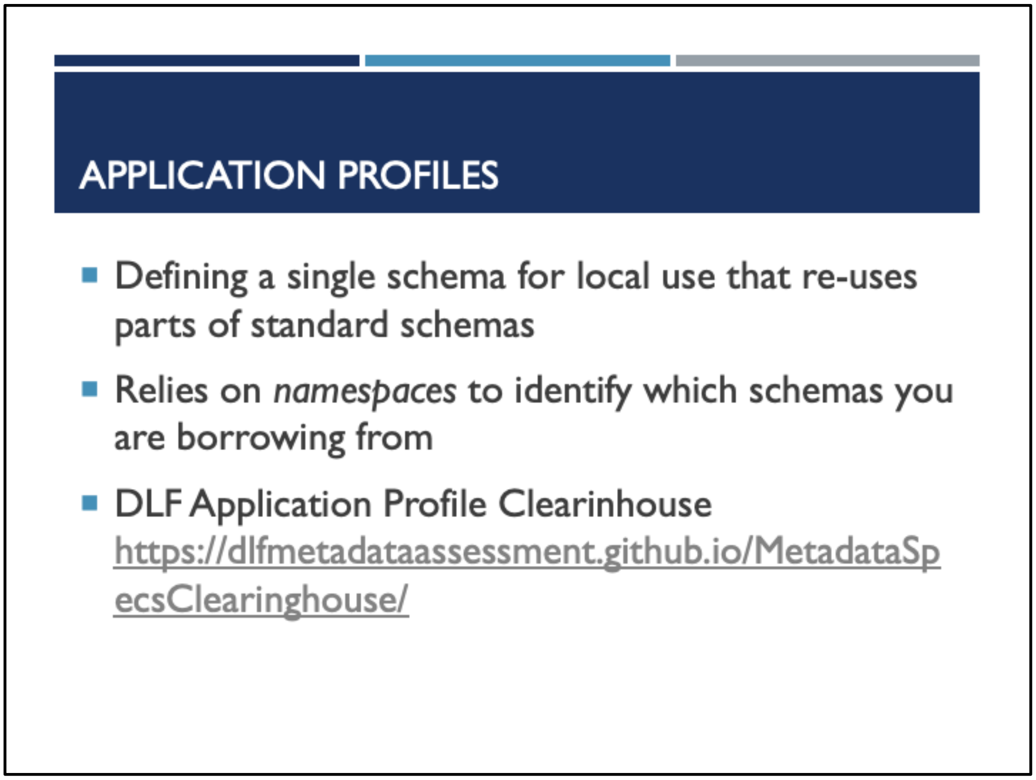
We can see how opening and closing tags contain all of the text values within them.

An attribute is expressed in XML by being within the opening tag and using an equal sign (type=created)

We can see that XML properties are repeatable.

And also hierarchical, meaning tags can wrap around not only values, but also other tags… like the way a parenthetical statement is inside a sentence,

This is a faked metadata schema…if you wanted to make an XML record using the Dublin Core schema, you could go to that reference and find out what the properties are for that standard and then write it up as XML. Same for MODS, or PREMIS, or VRA or any metadata schema. If you remember, back in our week on text digitization we also saw some XML. That time it was using a schema called TEI that is used to mark up text values. If you are familiar with HTML that is also a metadata schema that is often employed using XML.



## APPLICATION PROFILES

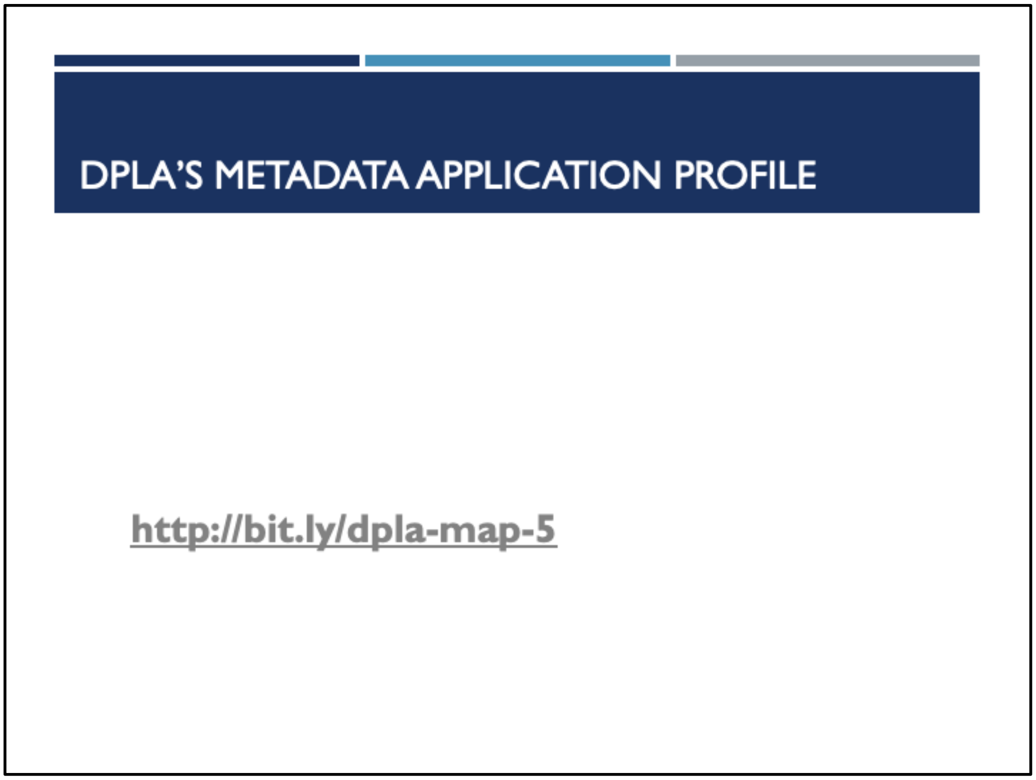
- Defining a single schema for local use that re-uses parts of standard schemas
- Relies on *namespaces* to identify which schemas you are borrowing from
- DLF Application Profile Clearinhouse https://dlfmetadataassessment.github.io/MetadataSpecsClearinghouse/

One other basic concept I want to cover is the Application Profile.

In a typical scenario, if you have a digital library that has a wide variety of material, you might find that no one schema is going to do the job for you really well. An application profile then is a local schema that incorporates elements from different standard schemas into one profile. It's like picking menu items from an ala carte menu, sort of.

This website, created by a subgroup of the Digital Library Federation, is a clearinghouse for metadata documentation including application profiles. It's a new thing and small now, but hopefully will grow over time. Let's look at the application profile from Rice University in the clearinghouse.

All it really is, is a list of elements taken from several schema that this university wants to use to describe Electronic Thesis and Dissertation. The rest of the docs are just some info on the definitions of these elements. Rice then takes these directions and creates metadata records for its Electronic Theses and dissertations that follow these rules. It's as simple as that really.



## DPLA'S METADATA APPLICATION PROFILE

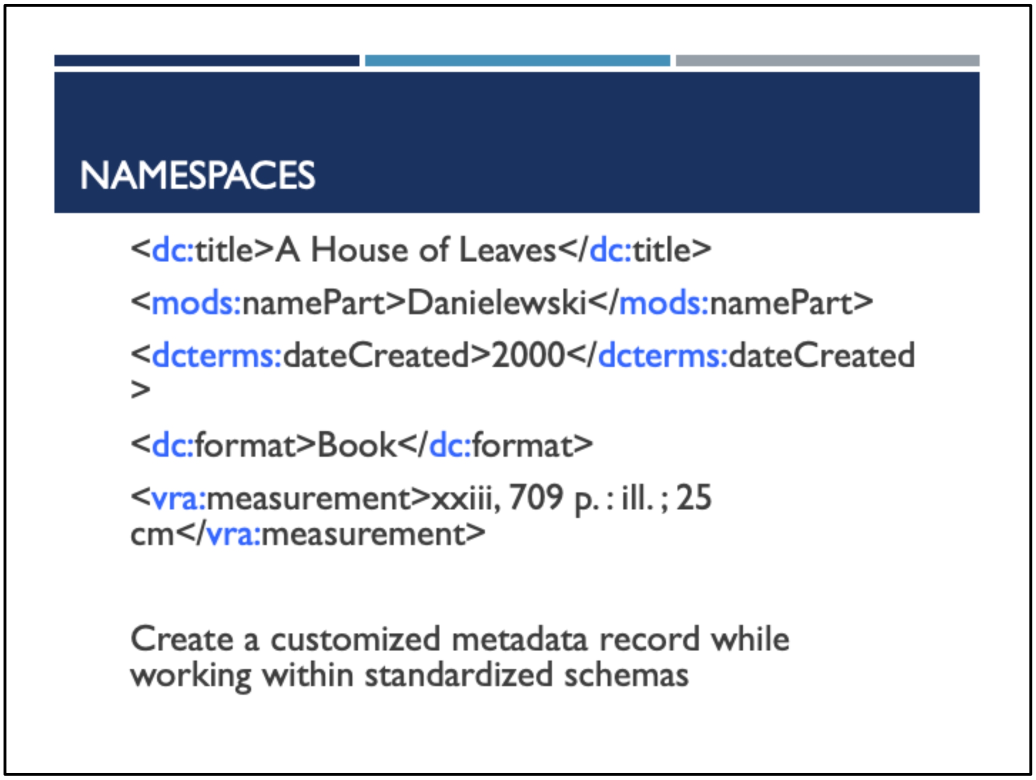
http://bit.ly/dpla-map-5

Let's look at another example, the application profile that DPLA uses.

This is a document that lays out the application profile used by DPLA to ensure that all of the records they share are the same schema. On page 5 you can see a declaration of the different schema used in the profile.

Looking at the table of properties used, you can see that a lot of them are from the DC and DCTERMS namespace.

But there are a few that DPLA has created for themselves because they couldn't find a schema with these specific things. By using an application profile, they only had to create two new properties, rather than declare an entirely new schema just for their purposes.



## NAMESPACES

```
<dc:title>A House of Leaves</dc:title>
<mods:namePart>Danielewski</mods:namePart>
<dcterms:dateCreated>2000</dcterms:dateCreated
>
<dc:format>Book</dc:format>
<vra:measurement>xxiii, 709 p. : ill. ; 25
cm</vra:measurement>
```

Create a customized metadata record while working within standardized schemas

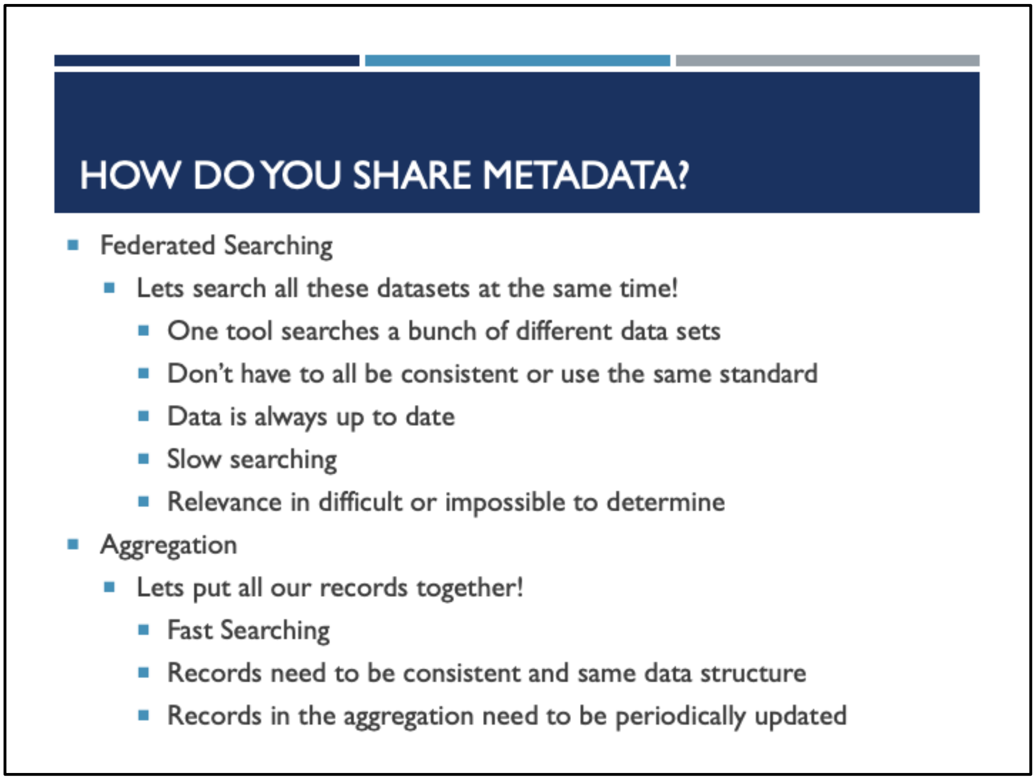
The way you distinguish where the different elements come from in an application profile is through the use of **namespaces**.

This takes the XML we just learned about and adds a new feature to the tag or the property, a namespace declaration: the blue parts there are the namespace: just an abbreviation identifying the schema the property comes from.

By adding that, we are saying that we are using title and format in the way they were described according to simple Dublin Core, we are using namepart as per MODS, we are using dateCreated as per dcterms (that‘s a schema that uses the qualified version of Dublin Core), and we are using measurement from VRA.

By mixing and matching elements from different schemas we can create one that is customized for our specific needs, while still working with standardized schemas.

Application profiles are now very common.



## HOW DO YOU SHARE METADATA?

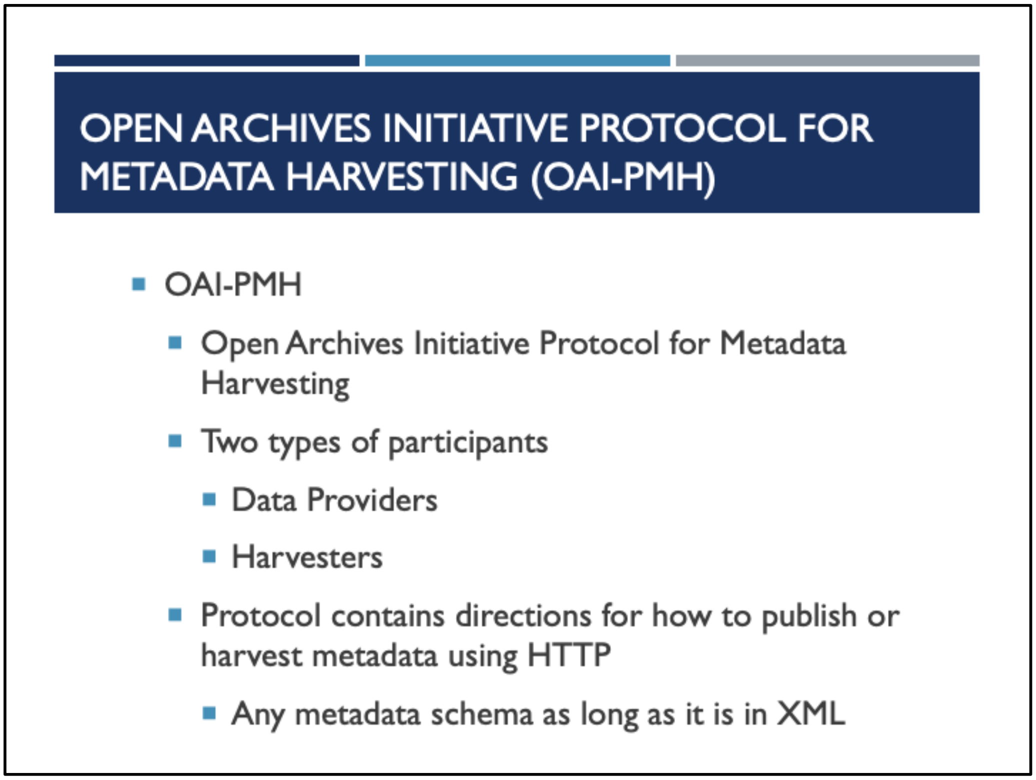
- Federated Searching
  - Lets search all these datasets at the same time!
    - One tool searches a bunch of different data sets
    - Don't have to all be consistent or use the same standard
    - Data is always up to date
    - Slow searching
    - Relevance in difficult or impossible to determine
- Aggregation
  - Lets put all our records together!
    - Fast Searching
    - Records need to be consistent and same data structure
    - Records in the aggregation need to be periodically updated

Now that we understand application profiles, which was really the last of the very basic concepts, let's move on to talking about how metadata interacts with the rest of the networked world.

One of the first things the information world started to get excited about when they starting putting data online was the opportunity to share it. This could take the form of their metadata records simply being indexed by a search engine like Google, or it could involve in some library-based aggregation project:  For example, if two libraries in the same town had their data online in a metadata format, theoretically you could build a tool that would search both of those data sets together.

There are two primary ways to share metadata for combined searching. One is **federated searching**, which creates a tool that is smart enough to go out and search a bunch of sources simultaneously. In this type of sharing the issues of consistency and structure aren't as important because you can customize the search to the source, and the data is always up to date. However, the searching can be slow. It is at the mercy of the speed of the local institution. And there can be problems with determining relevance for ranking results of queries between the sources, because they might have different metadata schema that aren't compare-able.

Another method is **aggregation**. We are going to talk a lot more about aggregation this week. This is when a service goes out and collects records from a bunch of different sources, puts them all into one index that is then searched. This means that the single index can be searched really quickly. However, that index will have to be periodically updated to reflect additions and changes at the various sources.



## OPEN ARCHIVES INITIATIVE PROTOCOL FOR METADATA HARVESTING (OAI-PMH)

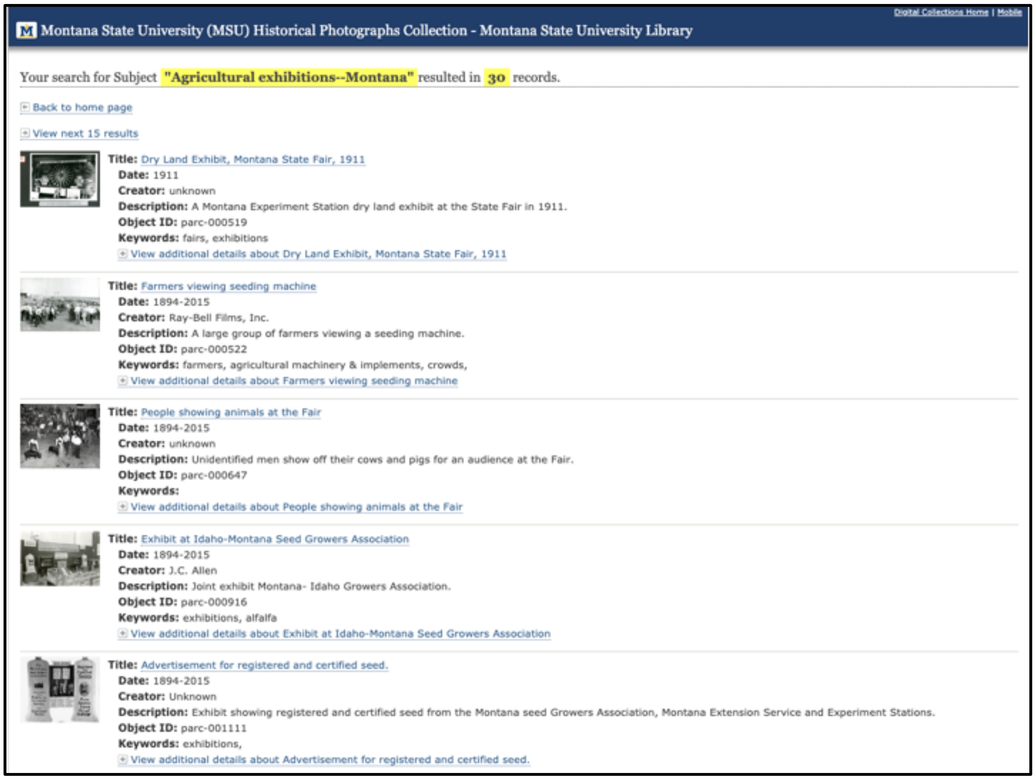
- OAI-PMH
  - Open Archives Initiative Protocol for Metadata Harvesting
  - Two types of participants
    - Data Providers
    - Harvesters
  - Protocol contains directions for how to publish or harvest metadata using HTTP
    - Any metadata schema as long as it is in XML

The Open Archives Initiative Protocol for Metadata Harvesting (OAI PMH (usually just referred to as OAI) ) is a technology that makes metadata aggregation possible.

OAI is based on two different roles. The first is the **data provider**, this is the organization that has records they would like to share. The second is the **harvester**, this is the organization that is gathering the records

OAI specifies a protocol with directions for how to either publish or harvest metadata using HTTP – another internet protocol (or set of directions for how to do something). The idea of a **protocol** is that there are tons of different ways to do this, but if we all do it the same way  -- in other words, follow the same protocol that OAI has set up -- then it's easier to cooperate without a lot of figuring out the wheel each time.

Besides the protocol there really aren't a lot of requirements, so you can share metadata in any format as long as it is expressed in XML



Montana State University (MSU) Historical Photographs Collection - Montana State University Library

Your search for Subject **"Agricultural exhibitions--Montana"** resulted in **30** records.

**Title:** Dry Land Exhibit, Montana State Fair, 1911
**Date:** 1911
**Creator:** unknown
**Description:** A Montana Experiment Station dry land exhibit at the State Fair in 1911.
**Object ID:** parc-000519
**Keywords:** fairs, exhibitions
View additional details about Dry Land Exhibit, Montana State Fair, 1911

**Title:** Farmers viewing seeding machine
**Date:** 1894-2015
**Creator:** Ray-Bell Films, Inc.
**Description:** A large group of farmers viewing a seeding machine.
**Object ID:** parc-000522
**Keywords:** farmers, agricultural machinery & implements, crowds,
View additional details about Farmers viewing seeding machine

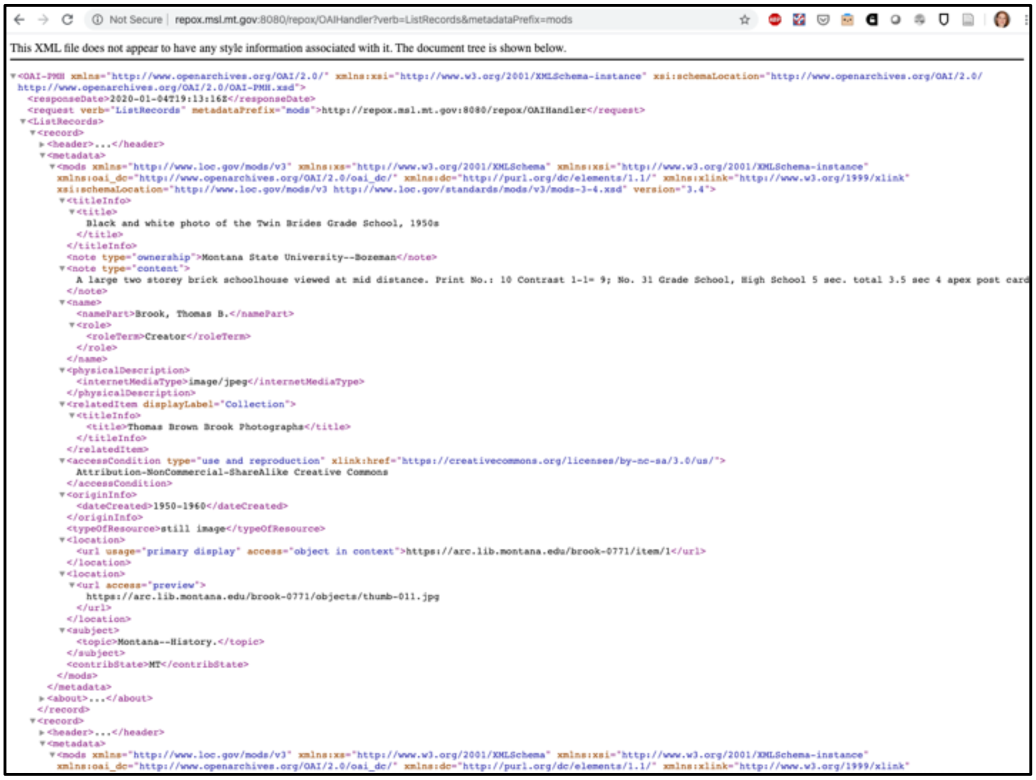
**Title:** People showing animals at the Fair
**Date:** 1894-2015
**Creator:** unknown
**Description:** Unidentified men show off their cows and pigs for an audience at the Fair.
**Object ID:** parc-000647
**Keywords:**
View additional details about People showing animals at the Fair

**Title:** Exhibit at Idaho-Montana Seed Growers Association
**Date:** 1894-2015
**Creator:** J.C. Allen
**Description:** Joint exhibit Montana- Idaho Growers Association.
**Object ID:** parc-000916
**Keywords:** exhibitions, alfalfa
View additional details about Exhibit at Idaho-Montana Seed Growers Association

**Title:** Advertisement for registered and certified seed.
**Date:** 1894-2015
**Creator:** Unknown
**Description:** Exhibit showing registered and certified seed from the Montana seed Growers Association, Montana Extension Service and Experiment Stations.
**Object ID:** parc-001111
**Keywords:** exhibitions,
View additional details about Advertisement for registered and certified seed.

Let’s look at this in action

This is a digital collection at the Montana State University. You can go to their website and see these items in their collection….



Not Secure | repox.msl.mt.gov:8080/repox/OAIHandler?verb=ListRecords&metadataPrefix=mods

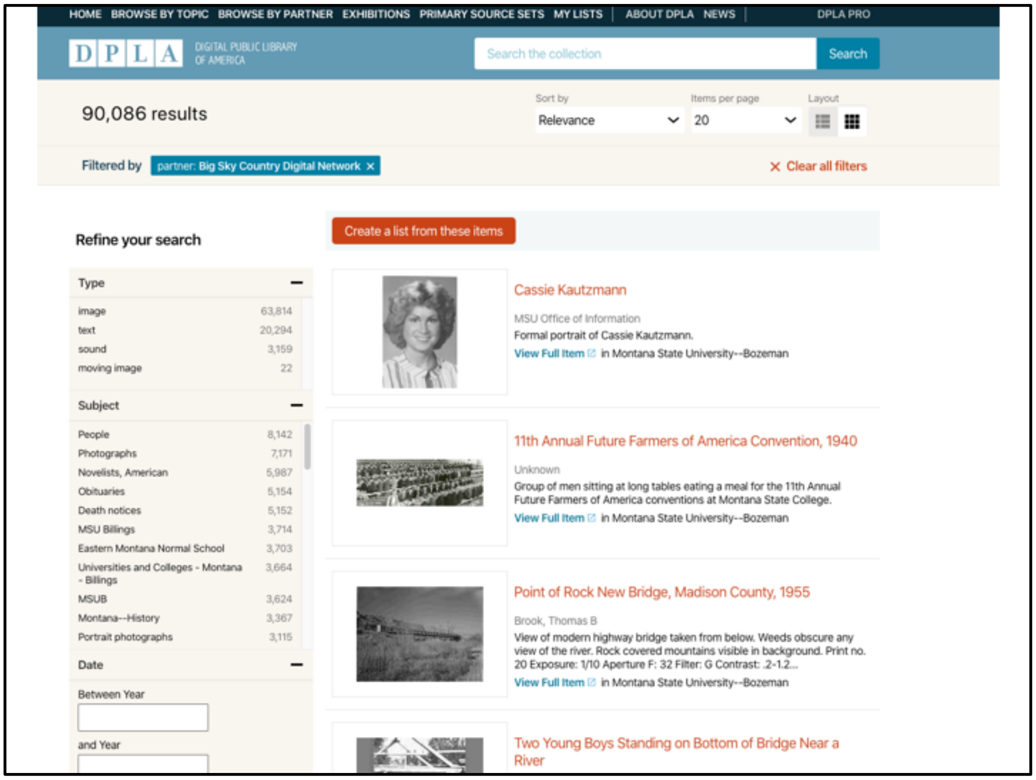
This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
▼<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
 http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
   <responseDate>2020-01-04T19:13:16Z</responseDate>
   <request verb="ListRecords" metadataPrefix="mods">http://repox.msl.mt.gov:8080/repox/OAIHandler</request>
  ▼<ListRecords>
    ▼<record>
      ▶<header>...</header>
      ▼<metadata>
        ▼<mods xmlns="http://www.loc.gov/mods/v3" xmlns:xs="http://www.w3.org/2001/XMLSchema" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
          xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:xlink="http://www.w3.org/1999/xlink"
          xsi:schemaLocation="http://www.loc.gov/mods/v3 http://www.loc.gov/standards/mods/v3/mods-3-4.xsd" version="3.4">
          ▼<titleInfo>
            ▼<title>
               Black and white photo of the Twin Brides Grade School, 1950s
             </title>
           </titleInfo>
           <note type="ownership">Montana State University--Bozeman</note>
          ▼<note type="content">
             A large two storey brick schoolhouse viewed at mid distance. Print No.: 10 Contrast 1-1= 9; No. 31 Grade School, High School 5 sec. total 3.5 sec 4 apex post card
           </note>
          ▼<name>
             <namePart>Brook, Thomas B.</namePart>
            ▼<role>
               <roleTerm>Creator</roleTerm>
             </role>
           </name>
          ▼<physicalDescription>
             <internetMediaType>image/jpeg</internetMediaType>
           </physicalDescription>
          ▼<relatedItem displayLabel="Collection">
            ▼<titleInfo>
               <title>Thomas Brown Brook Photographs</title>
             </titleInfo>
           </relatedItem>
          ▼<accessCondition type="use and reproduction" xlink:href="https://creativecommons.org/licenses/by-nc-sa/3.0/us/">
             Attribution-NonCommercial-ShareAlike Creative Commons
           </accessCondition>
          ▼<originInfo>
             <dateCreated>1950-1960</dateCreated>
           </originInfo>
           <typeOfResource>still image</typeOfResource>
          ▼<location>
             <url usage="primary display" access="object in context">https://arc.lib.montana.edu/brook-0771/item/1</url>
           </location>
          ▼<location>
            ▼<url access="preview">
               https://arc.lib.montana.edu/brook-0771/objects/thumb-011.jpg
             </url>
           </location>
          ▼<subject>
             <topic>Montana--History.</topic>
           </subject>
           <contribState>MT</contribState>
         </mods>
       </metadata>
      ▶<about>...</about>
     </record>
    ▼<record>
      ▶<header>...</header>
      ▼<metadata>
        ▼<mods xmlns="http://www.loc.gov/mods/v3" xmlns:xs="http://www.w3.org/2001/XMLSchema" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
          xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:xlink="http://www.w3.org/1999/xlink"
```

Or you can go to their OAI feed url…this one was set up for them through the Big Sky consortium in Montana. They set up an OAI server at this address that will output these records according to the OAI protocol

- To start the query I type "?verb="
- Then I include whatever OAI commands I want to use… Let's try this in real time

- Now this is how I as a single person who wants to look at the records in these feeds would do this. In a situation with an aggregation service, they would build some sort of software that would use the OAI protocol for harvesting to build a tool that would do this for the thousands of records in this feed and others and then save them in a new database. They then could make them all available through a new user-based website.



90,086 results

Sort by: Relevance

Items per page: 20

Filtered by partner: Big Sky Country Digital Network

Refine your search

| Type | |
|---|---|
| image | 63,814 |
| text | 20,294 |
| sound | 3,159 |
| moving image | 22 |

| Subject | |
|---|---|
| People | 8,142 |
| Photographs | 7,171 |
| Novelists, American | 5,987 |
| Obituaries | 5,154 |
| Death notices | 5,152 |
| MSU Billings | 3,714 |
| Eastern Montana Normal School | 3,703 |
| Universities and Colleges - Montana - Billings | 3,664 |
| MSUB | 3,624 |
| Montana--History | 3,367 |
| Portrait photographs | 3,115 |

Date

Between Year

and Year

Cassie Kautzmann

MSU Office of Information

Formal portrait of Cassie Kautzmann.

View Full Item in Montana State University--Bozeman

11th Annual Future Farmers of America Convention, 1940

Unknown

Group of men sitting at long tables eating a meal for the 11th Annual Future Farmers of America conventions at Montana State College.

View Full Item in Montana State University--Bozeman

Point of Rock New Bridge, Madison County, 1955

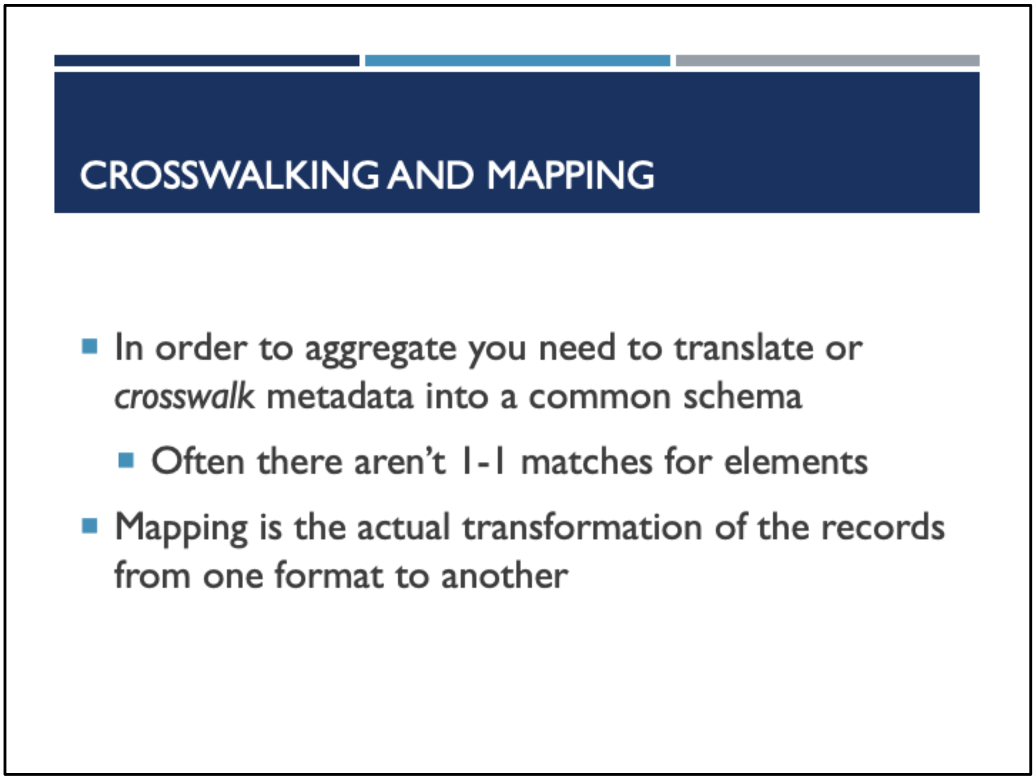
Brook, Thomas B

View of modern highway bridge taken from below. Weeds obscure any view of the river. Rock covered mountains visible in background. Print no. 20 Exposure: 1/10 Aperture F: 32 Filter: G Contrast: .2-1.2...

View Full Item in Montana State University--Bozeman

Two Young Boys Standing on Bottom of Bridge Near a River

Like this one at DPLA. They have an OAI harvester that will run and maybe capture 300,000 records in a couple of hours. They put those records into their own repository and then display them on their website. They may still end up with records in lots of different schemas though because some partners might use MODS, some use Dublin Core, some use MARC… so the next thing to do to get the records to be searchable together would be to convert them all to the same schema.



## CROSSWALKING AND MAPPING

- In order to aggregate you need to translate or *crosswalk* metadata into a common schema
    - Often there aren't 1-1 matches for elements
- Mapping is the actual transformation of the records from one format to another

The first step in being able to transform, or map, that metadata from one schema to another is to create a **crosswalk**.

The crosswalk is just a plan for what properties in the source schema we will convert to which properties in the schema we want to end up with.

Often there are not 1-1 matches between schemas, so things can get complicated. But it's basically like translating the values from one schema to another.

That crosswalk can then get transferred to some code in a scripting language like Python, or PERL, or Ruby, to create brand new records in the schema we want to end up with, based on the ones that we started with.

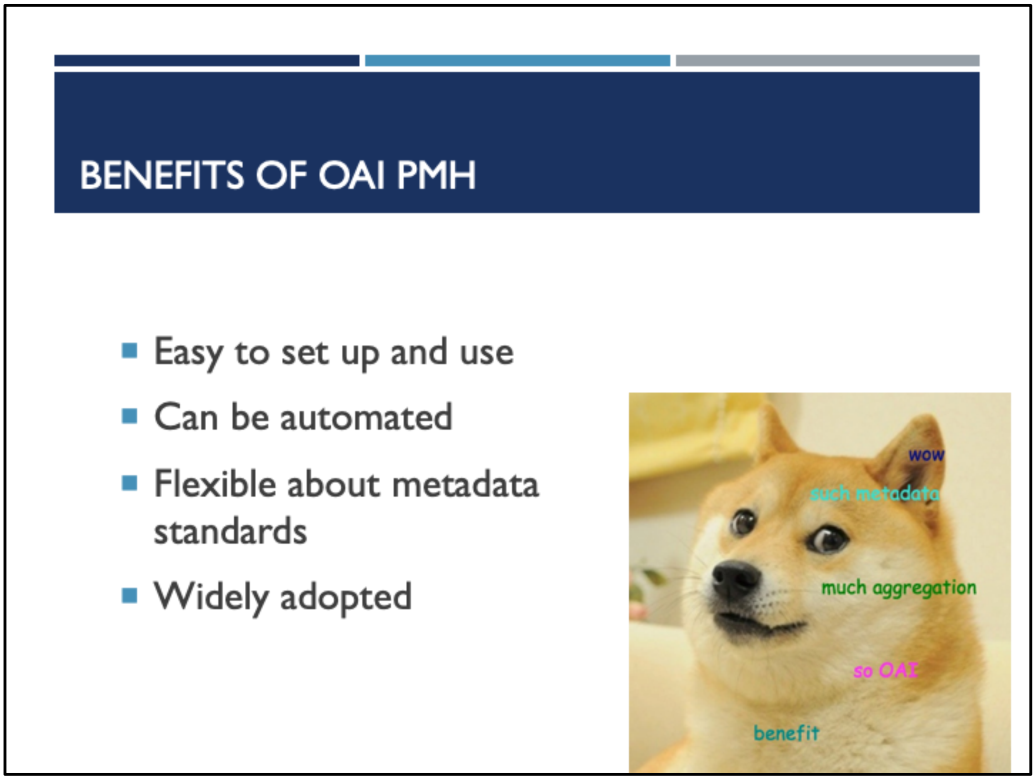
The way I think of them, **mapping** is the term usually used for actually doing that transformation and the crosswalk is the directions, but the two terms are often used interchangeably in the profession.

Let’s look at DPLA’s crosswalks:

http://bit.ly/dpla-MAP4-crosswalk

Let’s take a quick look at the crosswalk set up by DPLA.

I’m going to go into this in more depth in another video this week, because our first assignment will be to create a crosswalk like this for a sample record. But to give you a quick overview, this is a crosswalk set up DPLA. It basically is like a translation of how metadata in other formats would be expressed using DPLA’s application profile. So, if we look at the column for one of the DPLA partners, the Empire State Digital Network, who were the ones who published the record we just looked at, we see that it is like a series of equivalences… ESDN expresses creator in this way, and that would correspond to how DPLA expresses creator as dc:creator.



## BENEFITS OF OAI PMH

- Easy to set up and use
- Can be automated
- Flexible about metadata standards
- Widely adopted

So that’s essentially what OAI does…it allows you to gather records from all over and put them into one database. The crosswalking and mapping step is another optional way that you can make them all the same for better searching though. The point of OAI is just to gather the records

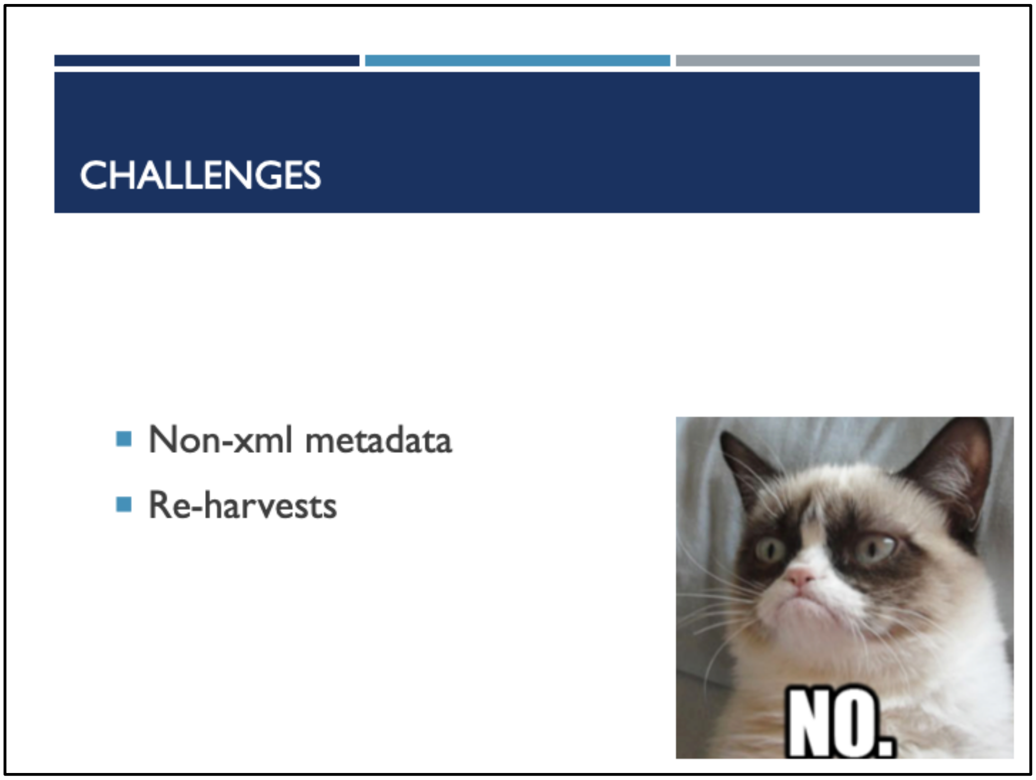
There are some pretty clear benefits to using OAI

First of all, it’s pretty simple. We’ve been looking at it in a very simple way, single HTTP requests through a browser, but there are tools out there to help set up a feed and harvest feeds and they are pretty easy to understand once you know some of the basic concepts.

The other thing that we saw is that the protocol can be used to create tools to do harvests in an automated fashion. So you can configure a bit of software to do these requests and gather these records, and it will go and do it on its own, you don’t have to deal with downloading say an export of all your files and manually sending them to an aggregator. And as an aggregator, I don’t have to be involved in every single request.

As we saw the OAI standard is flexible about metadata. It has it’s own very simple wrapper standard, but after that any metadata that is in XML can be posted.

Finally, OAI was developed in the early 2000s, and at this point it is widely adopted. Many repository services have OAI publishing as a standard part of their operation. Even our Omeka.net accounts we will set up for the final project  allow us to harvest from OAI feeds. (it's one of the plugins, check it out if you are interested).



## CHALLENGES

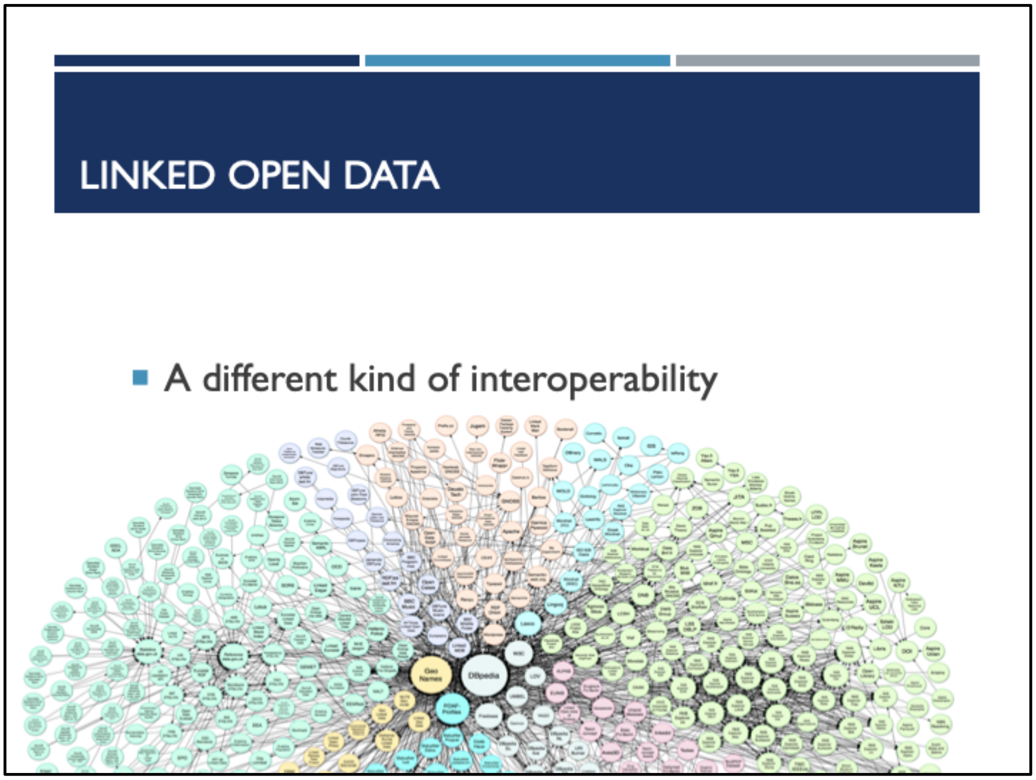
- Non-xml metadata
- Re-harvests

That’s not to say that there aren’t drawbacks however.

One that isn’t too big of an issue, but which sometimes becomes  a problem is the fact that we can only share metadata in xml formats. We can’t, for example, send files of csv encoded data using this protocol.

Another issue is with the time it takes to reharvest. OAI send a request over the web for records, typically in groups of no more than 500 at a time. While each request can only take milliseconds, as your provider grows, your harvests grow as well, taking up more and more time and resources to complete. If you amass collections of hundreds of thousands of records it could take more than 24 hours to complete depending on the speed of the server at your source, which you have no control over.

As I mentioned earlier , there is also the problem of the data that we have becoming out of date. When I worked at DPLA we reharvested records from our partners typically four times a year. Any changes they make between harvests aren’t reflected in our data until a few months later when we harvest again. Luckily, it’s mostly a question of just not having new records added. The records themselves don’t often change much, but if an item is removed from the source collection it could still have a record in the aggregation for several months..



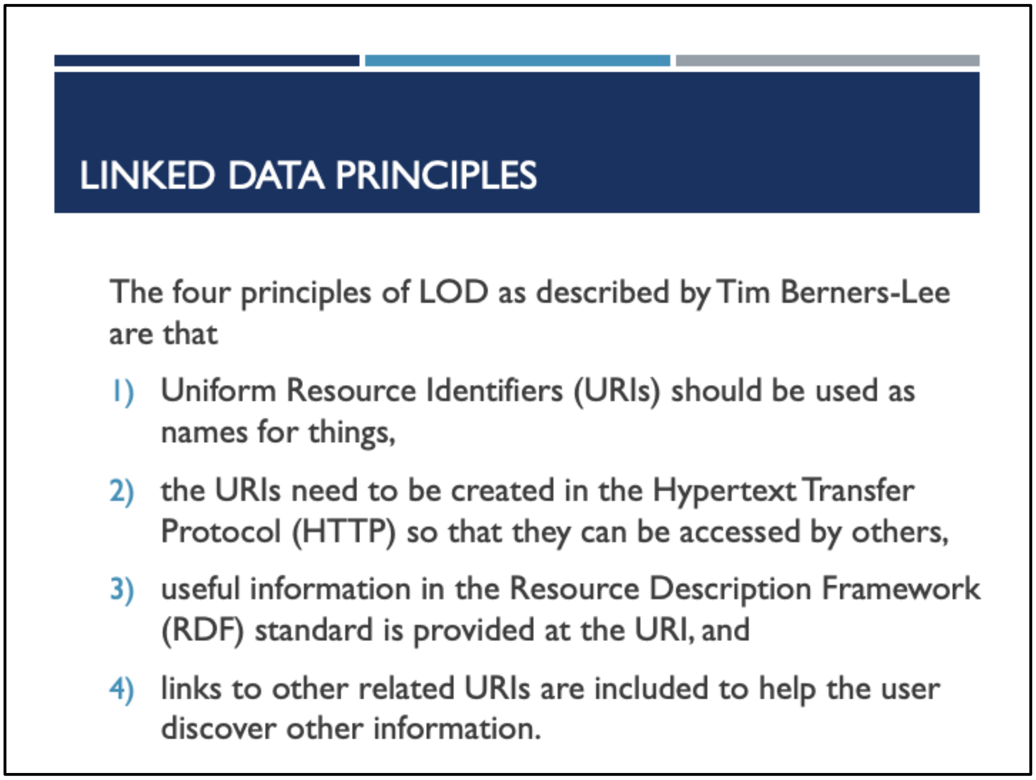
## LINKED OPEN DATA

- A different kind of interoperability

So OAI was one way to share metadata, and it involved sharing entire collections of records.

Another way of sharing data that is a more modern technology is call Linked Open Data or LOD. It is another kind of interoperability and it really involves sharing discrete portions of data, rather than entire collections or sets.

At this point, if you haven't yet watched the video about linked data that was posted in the readings, you should stop this lecture and go and do that now. I'm not going to cover linked data in as much depth, because the video does a pretty good job.



## LINKED DATA PRINCIPLES

The four principles of LOD as described by Tim Berners-Lee are that

1) Uniform Resource Identifiers (URIs) should be used as names for things,
2) the URIs need to be created in the Hypertext Transfer Protocol (HTTP) so that they can be accessed by others,
3) useful information in the Resource Description Framework (RDF) standard is provided at the URI, and
4) links to other related URIs are included to help the user discover other information.

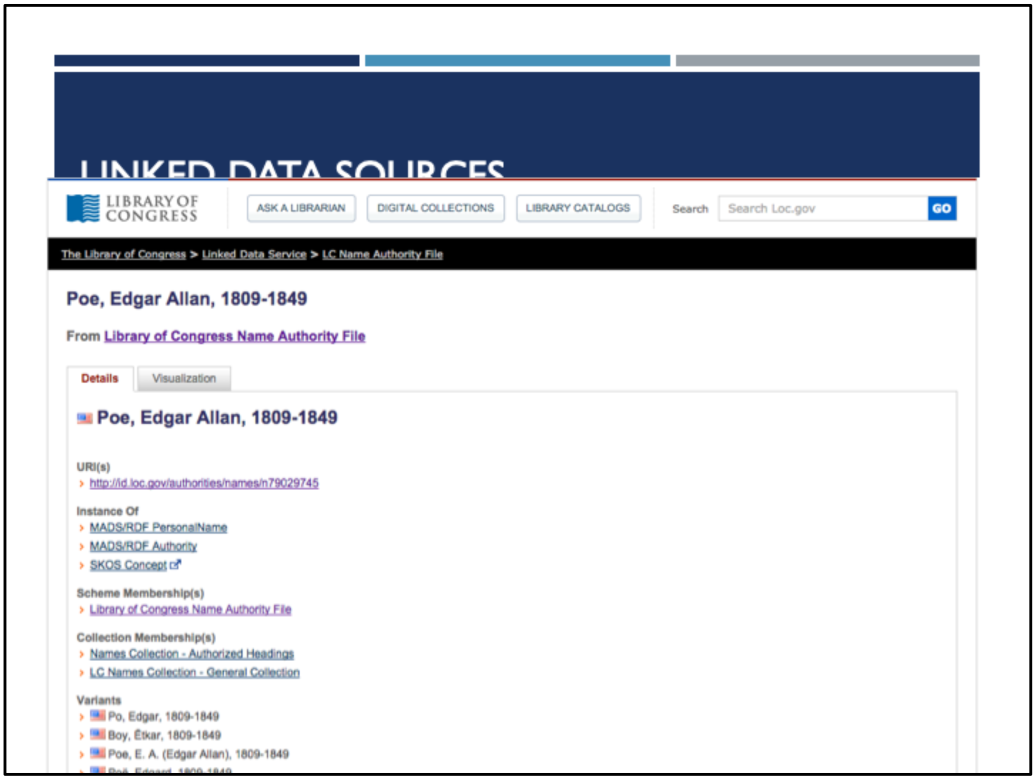
Assuming you've seen the video, I'll just reinforce a few concepts.

The premise of linked data hinges on the idea that everything has an URI. So concepts like an LCSH controlled subject heading, or the entry for an author's name in an authority file…these would all have a identifying URI (same thing as a URL really with a few important differences, but those aren't crucial for us to understand at this point).

When you wanted to say something about that concept or thing, you would use the URI as a reference point. Think of URIs as just like an identifying number for a concept.

There are rules about the technology used around URIs (that they are created according to the HTTP protocol, and that they are provided in a data encoding standard called RDF (not XML for example)).

Then, you can create resources that combine URIs from different sources and anyone who wants to find something out about a concept can look for that URI in different sources.



## LINKED DATA SOURCES

Poe, Edgar Allan, 1809-1849

From Library of Congress Name Authority File

Poe, Edgar Allan, 1809-1849

URI(s)
- http://id.loc.gov/authorities/names/n79029745

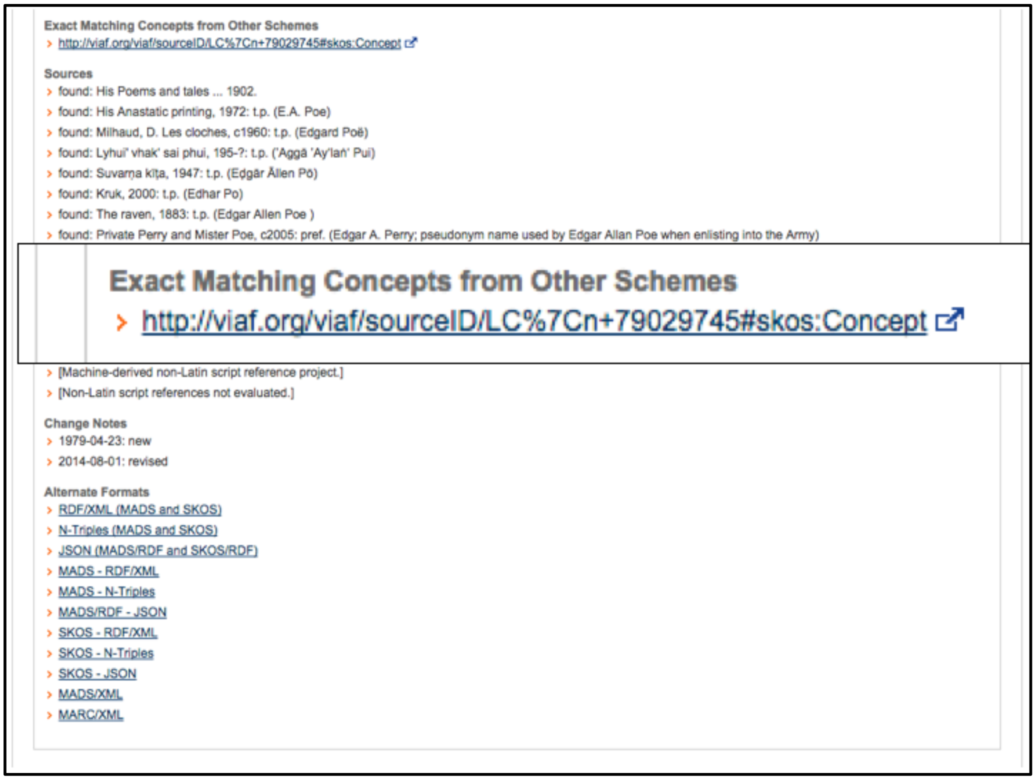
Instance Of
- MADS/RDF PersonalName
- MADS/RDF Authority
- SKOS Concept

Scheme Membership(s)
- Library of Congress Name Authority File

Collection Membership(s)
- Names Collection - Authorized Headings
- LC Names Collection - General Collection

Variants
- Po, Edgar, 1809-1849
- Boy, Ėtkar, 1809-1849
- Poe, E. A. (Edgar Allan), 1809-1849
- Poë, Edouard, 1809-1849

One thing I thought the video didn't necessarily cover explicitly is where those URIs come from. It's not like some universal council sits down and decides what the URI for cat is. Instead what you have are lots of different data services that have vocabularies of terms or definitions. One example here is the Library of Congress's Name Authority File. LC has created a URI for each name in it's file. So that's great, but then what if another organization has another URI for that name?



**Exact Matching Concepts from Other Schemes**

- http://viaf.org/viaf/sourceID/LC%7Cn+79029745#skos:Concept

**Sources**

- found: His Poems and tales ... 1902.
- found: His Anastatic printing, 1972: t.p. (E.A. Poe)
- found: Milhaud, D. Les cloches, c1960: t.p. (Edgard Poë)
- found: Lyhuiʹ vhakʹ sai phui, 195-?: t.p. (ʹAggā ʹAyʹlaṅʹ Pui)
- found: Suvaṃa kiṭa, 1947: t.p. (Eḍgār Ālen Pō)
- found: Kruk, 2000: t.p. (Edhar Po)
- found: The raven, 1883: t.p. (Edgar Allen Poe )
- found: Private Perry and Mister Poe, c2005: pref. (Edgar A. Perry; pseudonym name used by Edgar Allan Poe when enlisting into the Army)

**Exact Matching Concepts from Other Schemes**

- http://viaf.org/viaf/sourceID/LC%7Cn+79029745#skos:Concept

- [Machine-derived non-Latin script reference project.]
- [Non-Latin script references not evaluated.]

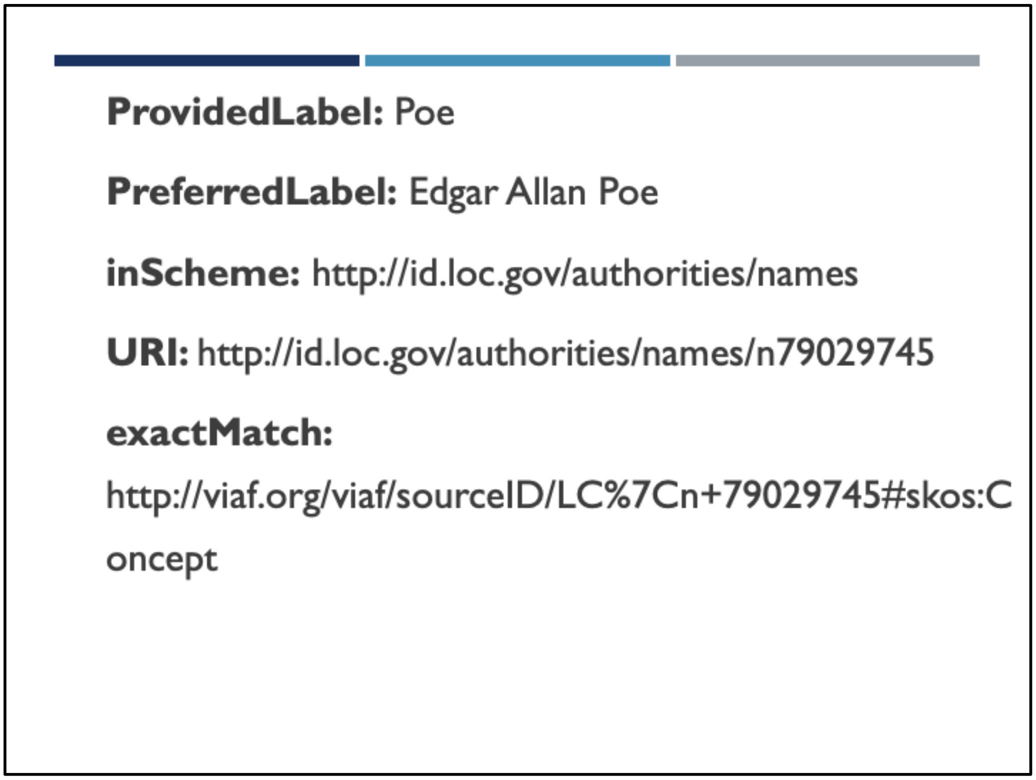
**Change Notes**

- 1979-04-23: new
- 2014-08-01: revised

**Alternate Formats**

- RDF/XML (MADS and SKOS)
- N-Triples (MADS and SKOS)
- JSON (MADS/RDF and SKOS/RDF)
- MADS - RDF/XML
- MADS - N-Triples
- MADS/RDF - JSON
- SKOS - RDF/XML
- SKOS - N-Triples
- SKOS - JSON
- MADS/XML
- MARC/XML

Well, if they say, like this LCNAF page does that their URI is the same thing as another source's URI, that's one way to link the two. Here at the top, we see that this LCNAF record states that their term for Edgar Allan Poe is exactly the same as this other term from VIAF (the Virtual International Authority File created by OCLC). So already we've joined the data here at the LC page, with the data over at VIAF.



**ProvidedLabel:** Poe

**PreferredLabel:** Edgar Allan Poe

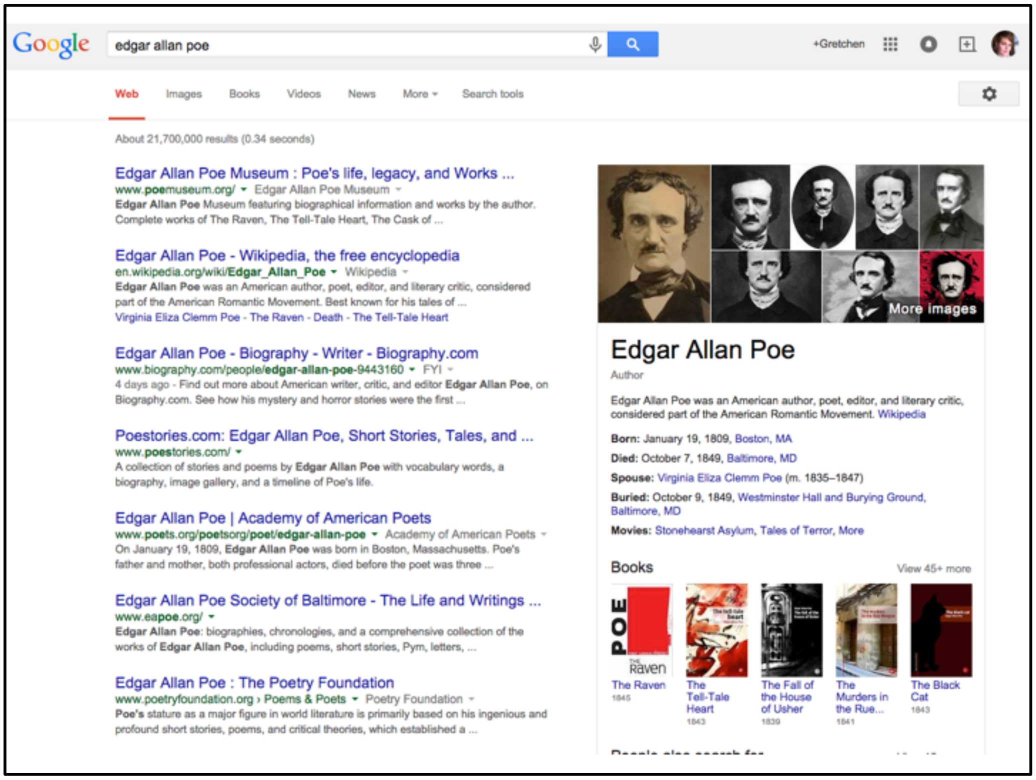
**inScheme:** http://id.loc.gov/authorities/names

**URI:** http://id.loc.gov/authorities/names/n79029745

**exactMatch:**
http://viaf.org/viaf/sourceID/LC%7Cn+79029745#skos:C
oncept

Another way that things get linked is when another party comes along and reuses those URIs.

This is some metadata from a theoretical authority record in a catalog. The original source record just said the name “Poe”. In our authority record, we are asserting that our author named Poe is the same as the Edgar Allan Poe named by LCSH, which is also the same as the entry for Edgar Allan Poe in VIAF. Now we have relationships between three data sources.



About 21,700,000 results (0.34 seconds)

Edgar Allan Poe Museum : Poe's life, legacy, and Works ...
www.poemuseum.org/ - Edgar Allan Poe Museum
**Edgar Allan Poe** Museum featuring biographical information and works by the author. Complete works of The Raven, The Tell-Tale Heart, The Cask of ...

Edgar Allan Poe - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Edgar_Allan_Poe** - Wikipedia
**Edgar Allan Poe** was an American author, poet, editor, and literary critic, considered part of the American Romantic Movement. Best known for his tales of ...
Virginia Eliza Clemm Poe - The Raven - Death - The Tell-Tale Heart

Edgar Allan Poe - Biography - Writer - Biography.com
www.biography.com/people/**edgar-allan-poe**-9443160 - FYI
4 days ago - Find out more about American writer, critic, and editor **Edgar Allan Poe**, on Biography.com. See how his mystery and horror stories were the first ...

Poestories.com: Edgar Allan Poe, Short Stories, Tales, and ...
www.**poes**tories.com/
A collection of stories and poems by **Edgar Allan Poe** with vocabulary words, a biography, image gallery, and a timeline of Poe's life.

Edgar Allan Poe | Academy of American Poets
www.**poets**.org/**poetsorg**/**poet**/**edgar-allan-poe** - Academy of American Poets
On January 19, 1809, **Edgar Allan Poe** was born in Boston, Massachusetts. Poe's father and mother, both professional actors, died before the poet was three ...

Edgar Allan Poe Society of Baltimore - The Life and Writings ...
www.ea**poe**.org/
**Edgar Allan Poe**: biographies, chronologies, and a comprehensive collection of the works of **Edgar Allan Poe**, including poems, short stories, Pym, letters, ...

Edgar Allan Poe : The Poetry Foundation
www.poetryfoundation.org › Poems & Poets - Poetry Foundation
**Poe's** stature as a major figure in world literature is primarily based on his ingenious and profound short stories, poems, and critical theories, which established a ...

Edgar Allan Poe
Author

Edgar Allan Poe was an American author, poet, editor, and literary critic, considered part of the American Romantic Movement. Wikipedia

**Born:** January 19, 1809, Boston, MA
**Died:** October 7, 1849, Baltimore, MD
**Spouse:** Virginia Eliza Clemm Poe (m. 1835–1847)
**Buried:** October 9, 1849, Westminster Hall and Burying Ground, Baltimore, MD
**Movies:** Stonehearst Asylum, Tales of Terror, More

The neat thing about Linked Data is seeing what you can do once you have those URI identifiers in sources. The video we watched about linked data used information about edgar allan poe in google search results. So let's talk about how this would really work.

First, google has recognized that my search term "Edgar Allan Poe" is something that has a URI reference in some different sources. Probably it has a list of such terms and kicks off this linked data tool when someone searches for them.

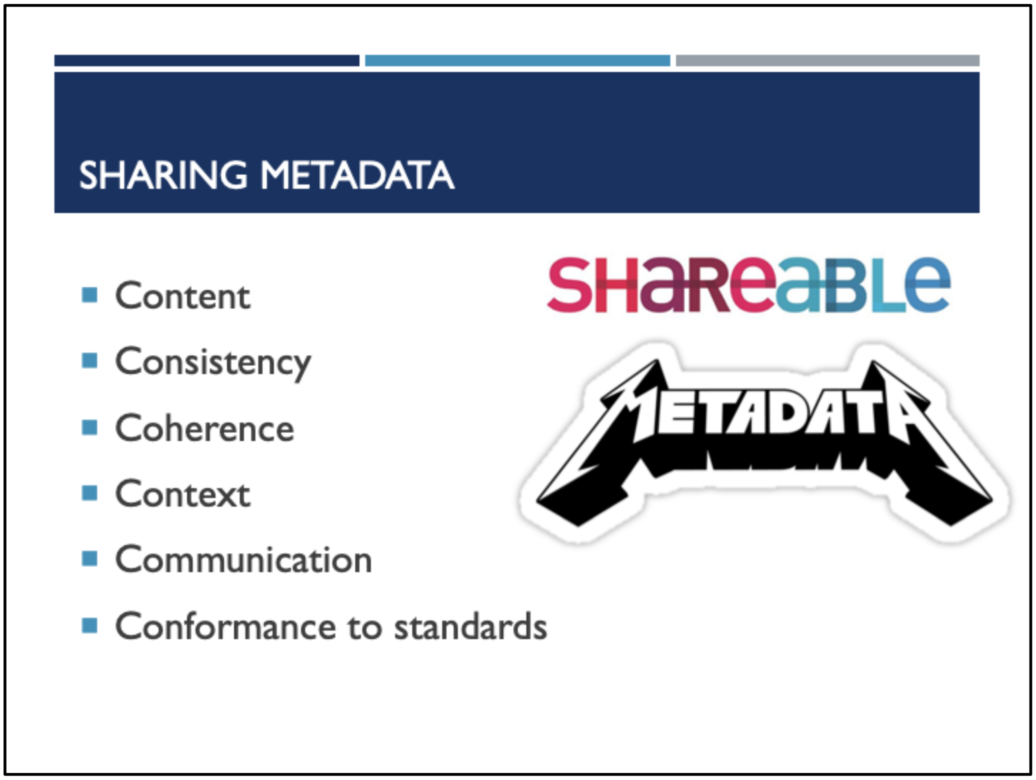
Next Google has gone to a couple of sources to look for matches to edgar allan poe. First it looks like it's gone to something like wikidata, which is a database of content from wikipedia put into linked data. So wikidata would have assigned a URI to edgar allan poe.

Google can look up that uri and then use it to grab wikidata's information about poe while it is bringing you the results of your search. It looks like it also went to a book source, probably it's own google books records.

Now, why is this any different from searching by Poe's name? Well, for one thing, using a single URI instead of a name gets around the problem of people writing the name in different ways. If one source said Poe, Edgar and another just said Poe, but they all used the same URI we can be sure that it really meant this Edgar Allan Poe.

It’s a way of dis-ambiguating or exerting authority.

The other thing that is a benefit here, is that linked data technology allows you to do this in real time if you want. So when you do the search, google can actually going out and pull that information dynamically and live from wikidata and elsewhere. So if you add another book, or update the information, it gets updated here as well, in real-time. This removes problems with synchronization and indexing. It might cause your application to be really slow though, so you might actually create a local cache of what’s at the source and update it periodically instead. But the point is you have the ability.



## SHARING METADATA

- Content
- Consistency
- Coherence
- Context
- Communication
- Conformance to standards

SHAREABLE
METADATA

Let's transition now to talk about quality.  Issues of quality are really important when we want to share metadata. We read an article this week that talked about "Shareable Metadata." That phrase was coined more than 10 years ago, and it that refers to thinking about your metadata in terms of not just the local community, but the global community. So when you are creating your metadata in your local context, you are also thinking about what it would look like if it was picked up and reused elsewhere.

We learned from the Shreeves article that when thinking about the creation of metadata with sharing in mind we should think about these six "c"s

**Content**: what information would it take to make this content understandable to anyone. Will someone from another country understand that a picture of "LBJ" is Lyndon Johnson or will they not recognize that acronym

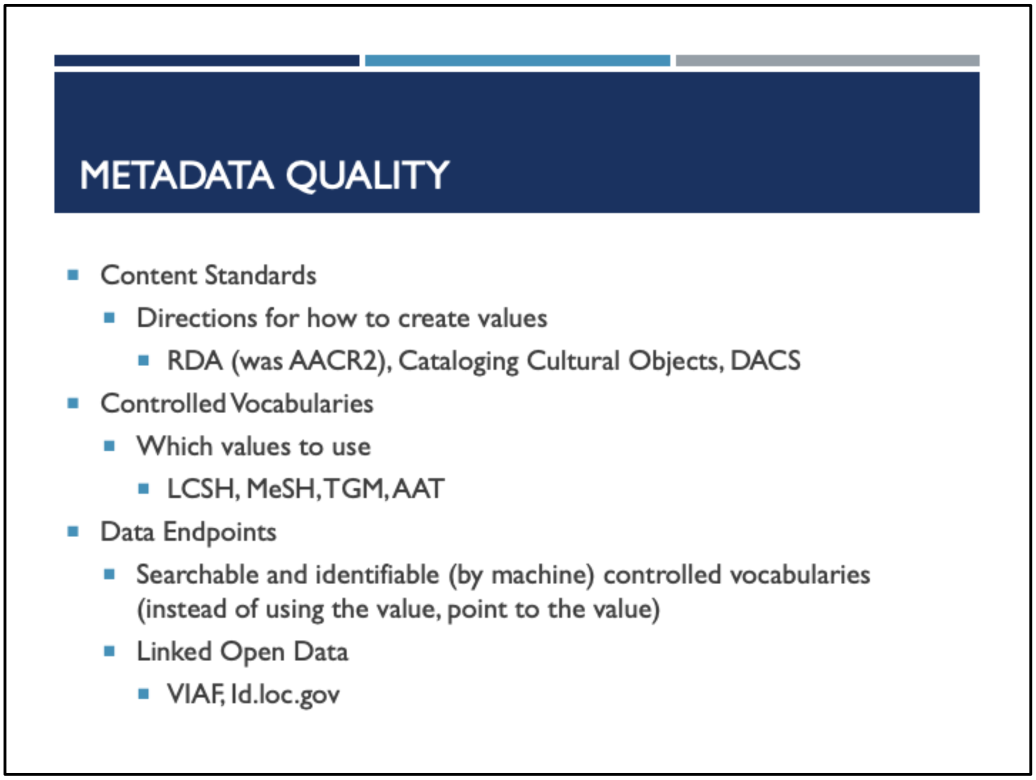
**Consistency**: refers to consistent use of similar forms of metadata values. For example, using the same controlled vocabulary throughout, or formatting dates or author names the same way throughout all records

**Coherence**: means that records are self-explanatory and complete. A record shouldn't mix descriptions of different views of the object. For example, a record for a digitized image shouldn't include both the date of the original objects creation and the date of the digitization in the metadata without distinguishing the different between them.

**Context**: means that any information about context that is needed to make the record understandable outside of it's original collection is explicitly included in the metadata. That could include information about a collection the item is a part of for example.

**Communication**: relates to the actual interaction between those who own and organize collection and another organization it is sharing data with. You won't always be able to control this element, but when you can you should include all the relevant information like schema and vocabularies used, when the data was last updated, etc.

Finally, **Conformance** to standards is key for sharing. Creating your own local standard may really suit your needs, but if no one else understands it, it won't be useful. As we mentioned application profiles can really help with this.



## METADATA QUALITY

- Content Standards
  - Directions for how to create values
    - RDA (was AACR2), Cataloging Cultural Objects, DACS
- Controlled Vocabularies
  - Which values to use
    - LCSH, MeSH, TGM, AAT
- Data Endpoints
  - Searchable and identifiable (by machine) controlled vocabularies (instead of using the value, point to the value)
  - Linked Open Data
    - VIAF, Id.loc.gov

In addition to those concepts, we can also rely on different standards to ensure quality.
**Content standards** give us directions for how to create values. Examples are how to format dates or author names, or when creating a title for an untitled item how to formulate that.

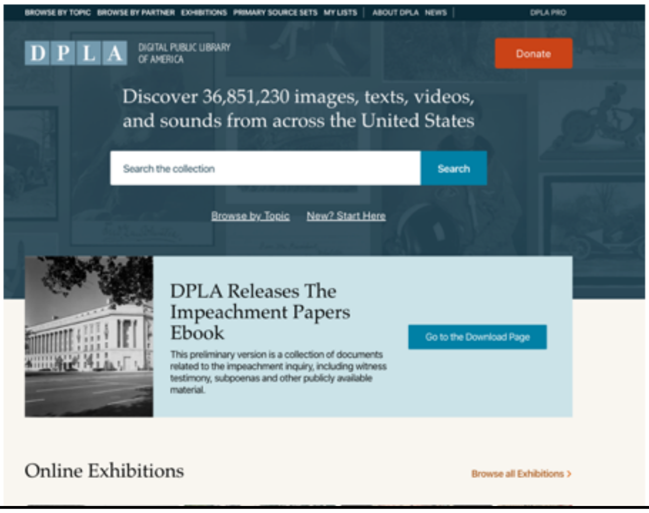
**Controlled vocabularies** are lists of defined terminology for specific purposes. A lot of them involve subject headings (like LCSH or MeSH) but other are related to terms for formats, or authorized versions of names.

In the linked data world we also have **Data Endpoints**, which are like linked data versions of controlled vocabularies You can put in an identifier for a term in the vocabulary and you can write code in your web interface that will go and look up the term. The benefit is that if the term is updated, what you display is always updated. Or if you want to have an option to display the interface in a different language, if the end point includes variations of the term in different languages, you can do that automatically. This is a core concept of Linked Open Data. I'll post a supplementary video this week with an example of using a Linked Data endpoint to add some data to a spreadsheet.

So some examples of endpoints are the virtual international authority file (for persons names), id.loc.gov which holds all of the library of congress authorities, and geonames is a particular geographical endpoint.

## SHARING IN THE REAL WORLD

- DPLA uses both OAI PMH and LOD



Discover 36,851,230 images, texts, videos, and sounds from across the United States

Search the collection

DPLA Releases The Impeachment Papers Ebook

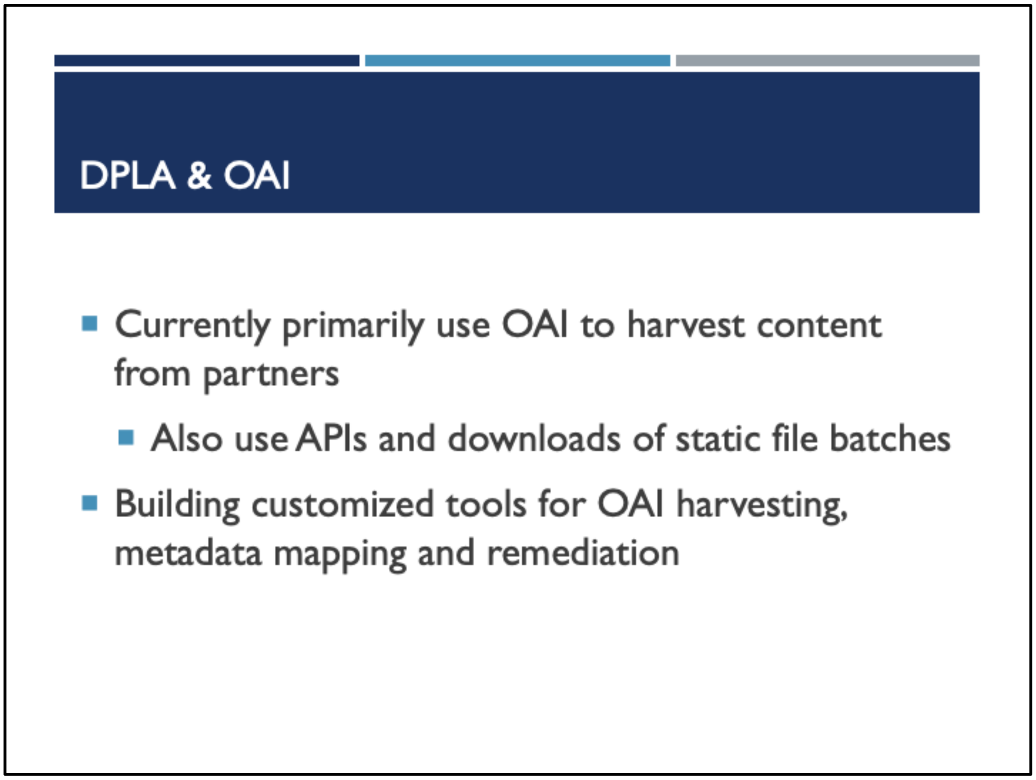
This preliminary version is a collection of documents related to the impeachment inquiry, including witness testimony, subpoenas and other publicly available material.

Online Exhibitions

So how about we look at an example of both of these technologies in the wild. My previous employer, DPLA, uses both OAI and Linked Data

--as an aside, you'll hear people say both Linked Open Data and Linked Data. When the word "open" is used it connotes linked data that is freely available to all. Not all linked data is open, some are subscription services. Just as a side point…

Anyway

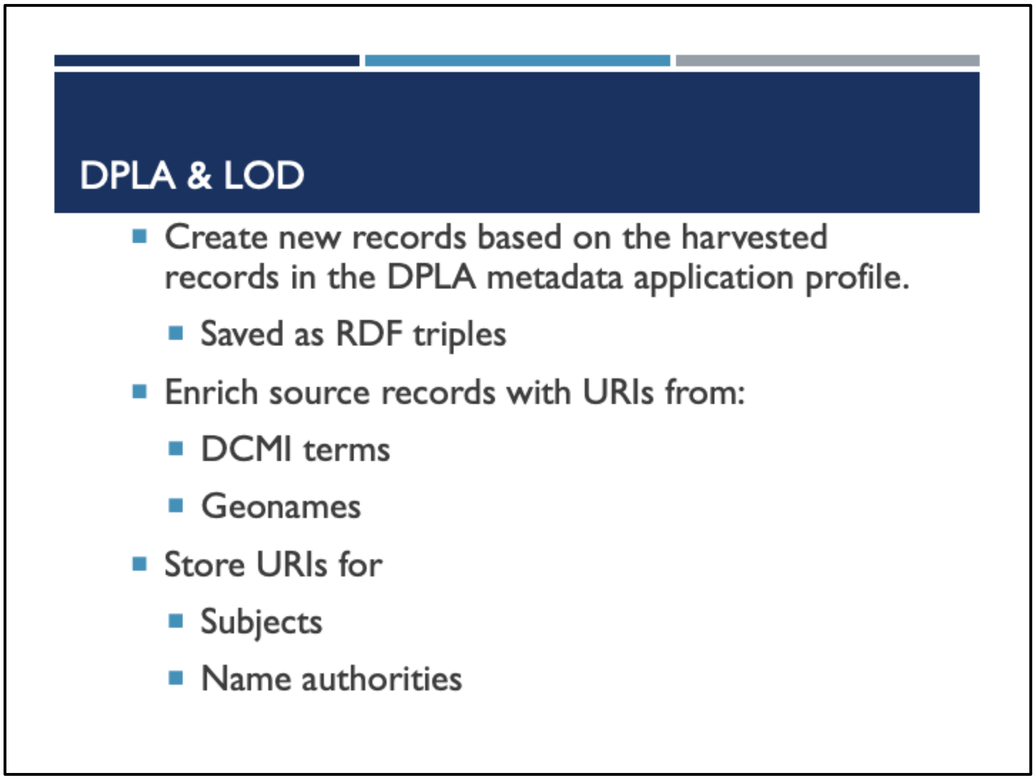


## DPLA & OAI

- Currently primarily use OAI to harvest content from partners
  - Also use APIs and downloads of static file batches
- Building customized tools for OAI harvesting, metadata mapping and remediation

At DPLA we aggregated metadata from more than 40 hubs, representing about 3,000 individual institutions and aggregated more than 30 million records. We received metadata in 9 different schemas, including major standards as well as product-and institution-specific profiles, and several are not in an XML format. We harvest records from OAI feeds, APIs, and through submitted data files, and many of these records have already been aggregated by one of our hubs.

We built our own customized tools for OAI harvesting, metadata mapping and metadata remediation because we need to work on it at such a large scale.



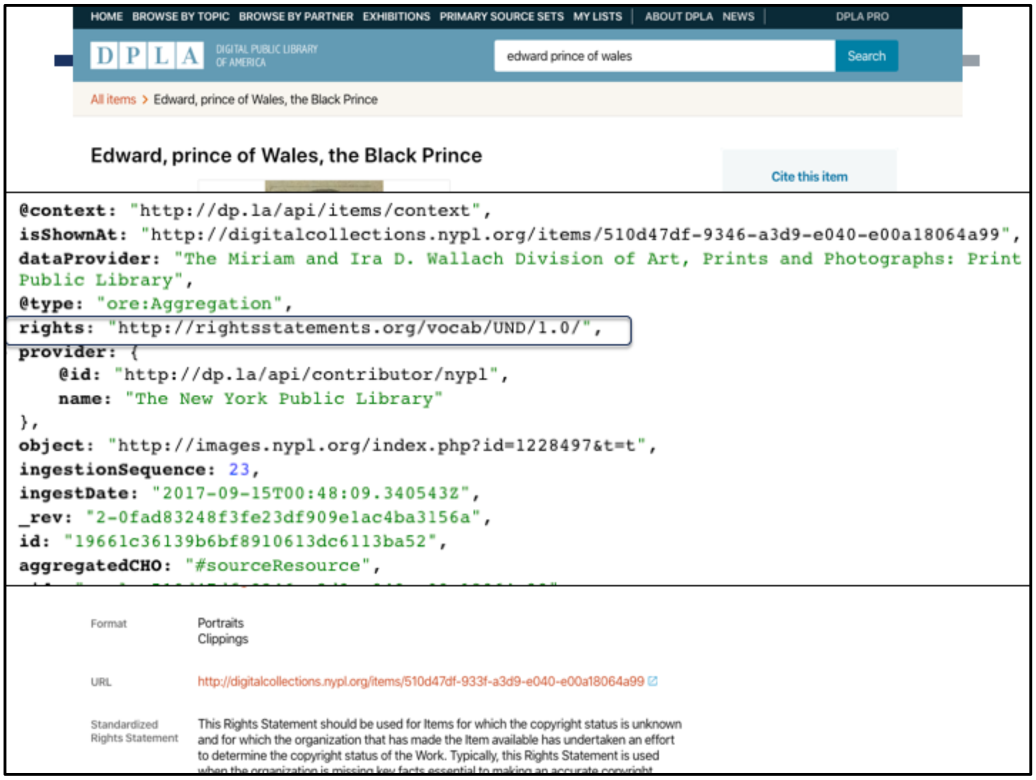
## DPLA & LOD

- Create new records based on the harvested records in the DPLA metadata application profile.
  - Saved as RDF triples
- Enrich source records with URIs from:
  - DCMI terms
  - Geonames
- Store URIs for
  - Subjects
  - Name authorities

Once we had the partners' source records we mapped that data to make a record that adheres to the DPLA metadata application profile according to our crosswalks (remember we looked at that back at the beginning of the lecture).

We also added to our records URIs for some of the terms in the records if we can make a match for them, particularly the types from the Dublin Core type list (things like image, text, video, etc.) as well as geographical names in the Geonames vocabularly

We didn't add URIs for things like subjects or name authorities yet, but if the source records do have those URIs we could store them in our record as well. We were working on plans to enrich records with those URIs in the future.


HOME BROWSE BY TOPIC BROWSE BY PARTNER EXHIBITIONS PRIMARY SOURCE SETS MY LISTS ABOUT DPLA NEWS
DPLA PRO
D P L A DIGITAL PUBLIC LIBRARY OF AMERICA
edward prince of wales
Search
All items > Edward, prince of Wales, the Black Prince
Cite this item

Edward, prince of Wales, the Black Prince

```
@context: "http://dp.la/api/items/context",
isShownAt: "http://digitalcollections.nypl.org/items/510d47df-9346-a3d9-e040-e00a18064a99",
dataProvider: "The Miriam and Ira D. Wallach Division of Art, Prints and Photographs: Print
Public Library",
@type: "ore:Aggregation",
rights: "http://rightsstatements.org/vocab/UND/1.0/",
provider: {
    @id: "http://dp.la/api/contributor/nypl",
    name: "The New York Public Library"
},
object: "http://images.nypl.org/index.php?id=1228497&t=t",
ingestionSequence: 23,
ingestDate: "2017-09-15T00:48:09.340543Z",
_rev: "2-0fad83248f3fe23df909e1ac4ba3156a",
id: "19661c36139b6bf8910613dc6113ba52",
aggregatedCHO: "#sourceResource",
```

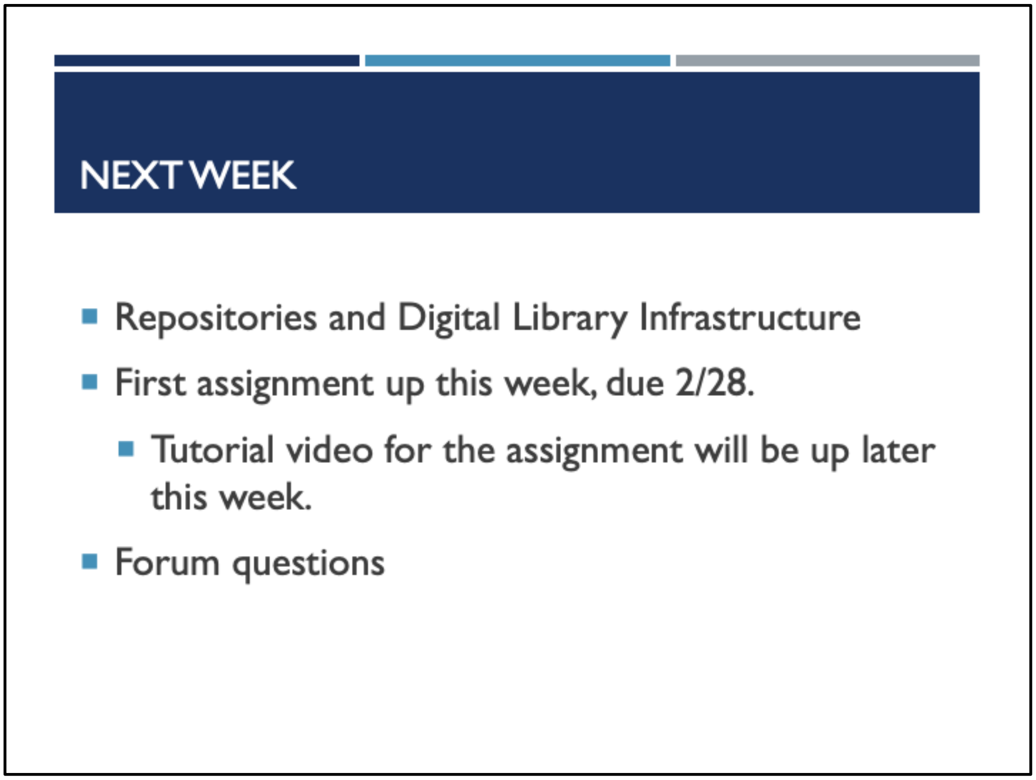
| | |
|---|---|
| Format | Portraits<br>Clippings |
| URL | http://digitalcollections.nypl.org/items/510d47df-933f-a3d9-e040-e00a18064a99 |
| Standardized Rights Statement | This Rights Statement should be used for Items for which the copyright status is unknown and for which the organization that has made the Item available has undertaken an effort to determine the copyright status of the Work. Typically, this Rights Statement is used when the organization is missing key facts essential to making an accurate copyright |

With one particular type of URI we were enriching records with data from an endpoint.

In this record, the text and the icon for rights status comes from a linked data endpoint. The original record only contains the URI for a rights statement, which is outlined in red here. That's what the underlying metadata record for this object (or a part of the record) looks like (this is a different format than XML, by the way, this format is called JSON).

But the way we've written the code for the web-displayed version shows the full statement label and description which comes from a website called rightsstatements.org (we'll be learning more about this later, by the way). Right now the statements are available in other languages, so we could create an alternate language version of the site that would use a specific language instead. It will also automatically update if the descriptions get modified or the icons change.

So that’s it for metadata part two. I hope this wasn’t too mind-blowing. I know the linked data stuff can be hard to grasp at first. Metadata is pretty cool and pretty powerful. I think the key thing to keep in mind is that metadata should be consistent and as understandable as possible to make it the most useful if it gets aggregated or re-used.



## NEXT WEEK

- Repositories and Digital Library Infrastructure
- First assignment up this week, due 2/28.
  - Tutorial video for the assignment will be up later this week.
- Forum questions

This week we have our first assignment. I’ll be posting a demo of me working on a crosswalk that will help you with that. It will be due on the 28th

Next week we’ll bring our objects and metadata together into a digital repository.

As usual, you also have a forum question available to you this week.