#### University of Massachusetts Boston ScholarWorks at UMass Boston

Joseph P. Healey Library Publications

Joseph P. Healey Library

3-6-2014

### Working with Data in Archival Settings

Joanne M. Riley *University of Massachusetts - Boston*, joanne.riley@umb.edu

Follow this and additional works at: http://scholarworks.umb.edu/hlpubs

Part of the <u>History of Science, Technology, and Medicine Commons</u>, <u>Other Education Commons</u>, and the <u>Technology and Innovation Commons</u>

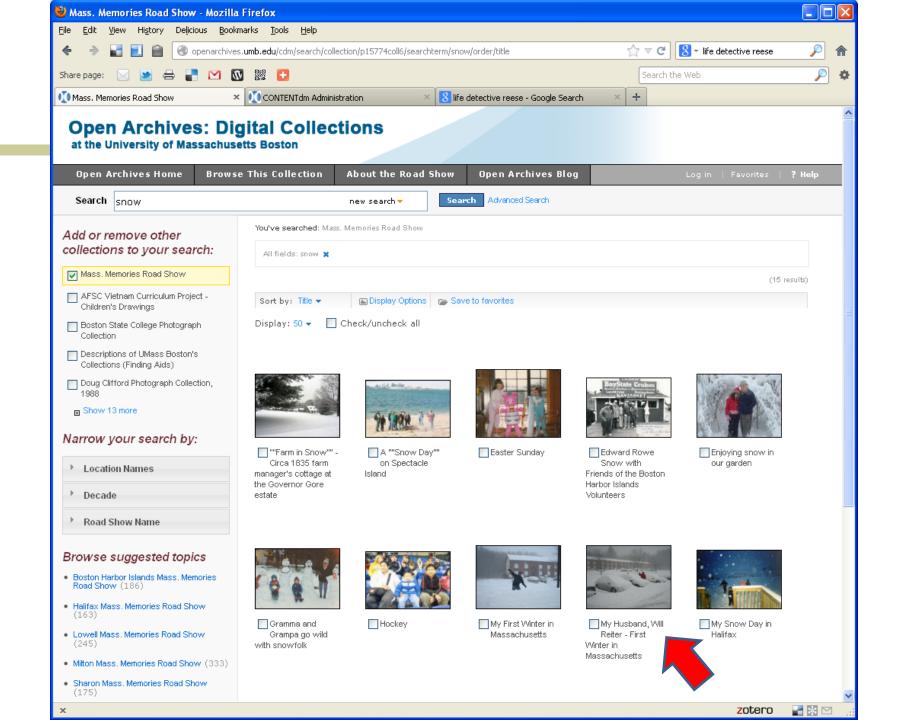
#### Recommended Citation

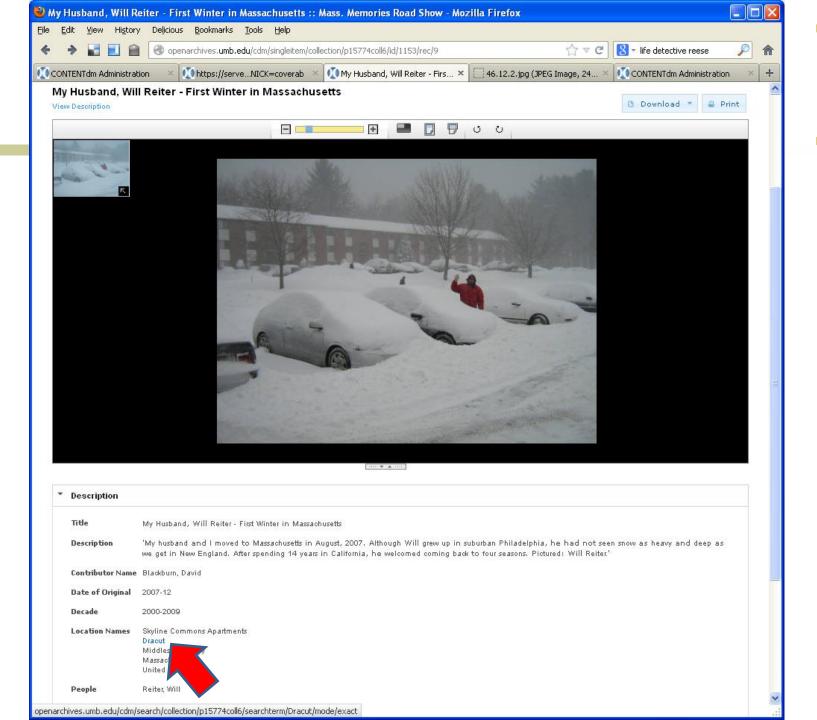
Riley, Joanne M., "Working with Data in Archival Settings" (2014). *Joseph P. Healey Library Publications*. Paper 26. http://scholarworks.umb.edu/hlpubs/26

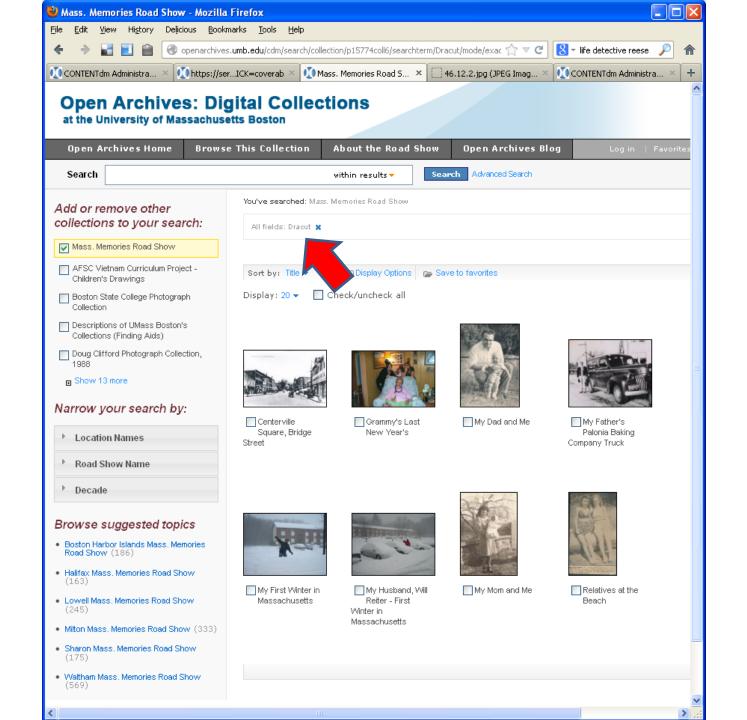
This Presentation is brought to you for free and open access by the Joseph P. Healey Library at ScholarWorks at UMass Boston. It has been accepted for inclusion in Joseph P. Healey Library Publications by an authorized administrator of ScholarWorks at UMass Boston. For more information, please contact library.uasc@umb.edu.

# Working with Data in Archival Settings

Joanne Riley
University Archivist
Joseph P. Healey Library
University of Massachusetts Boston
617-287-5927
Joanne.riley@umb.edu







## What is Metadata

### Standard definitions:

- "'Data About Data"
- "Information About Information"

But what, really, does that MEAN?!

### CDs store <u>audio</u> <u>information</u>



### Information about information



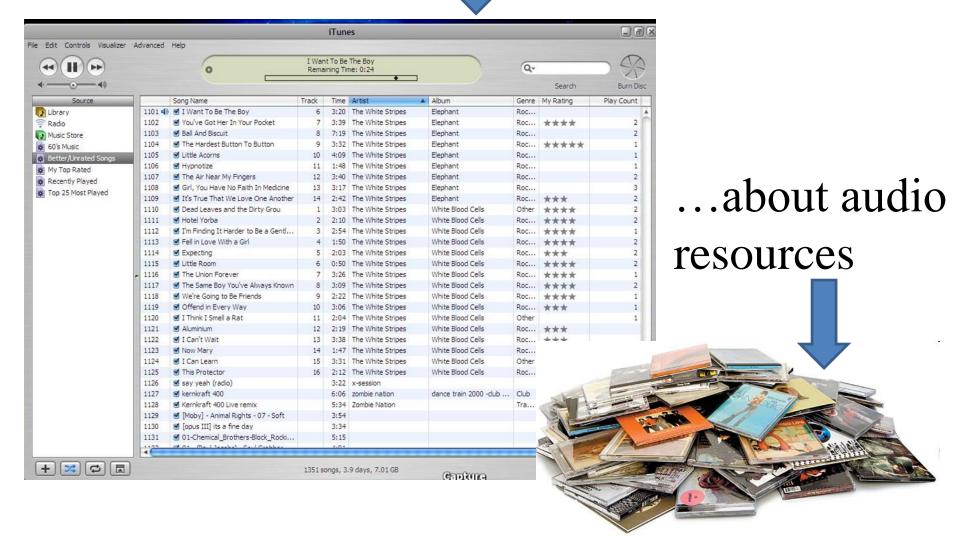
Tchaikovsky composed this piece

This song is 1'30" long

This is my favorite song

Organize this information in a structured way and it becomes "metadata"

### iTunes stores Metadata...



## What is Metadata?

"Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource."

## Benefits of Metadata

- Allows resources to be found, identified, described, annotated, collocated, shared and disambiguated
- "Describing a resource with metadata allows it to be understood by both humans and machines in ways that promote interoperability."
- Offers "ability to search resources seamlessly across a network"

### Archives are chock full of resources in need of metadata

Visual Textual Information



## Semantic information on the web can be expressed in three different ways:

- 1. Inferred
- 2. Free-form tagging
  - 3. Structured data

These go in order from the **consumer** of data doing most of the work, to the **producer** of data doing most of the work.

Source: "A brief introduction to the Semantic Web"

Yaron Koren

National Library of Israel

August 11, 2011

http://wikiworks.com/nli-seminar/intro-to-semantic-web.pdf

## Semantic information on the web can be expressed in three different ways:

- 1. Inferred
- Free-form tagging
  - 3. Structured data

These go in order from the **consumer** of data doing most of the work, to the **producer** of data doing most of the work.

Source: "A brief introduction to the Semantic Web"

Yaron Koren

National Library of Israel

August 11, 2011

http://wikiworks.com/nli-seminar/intro-to-semantic-web.pdf

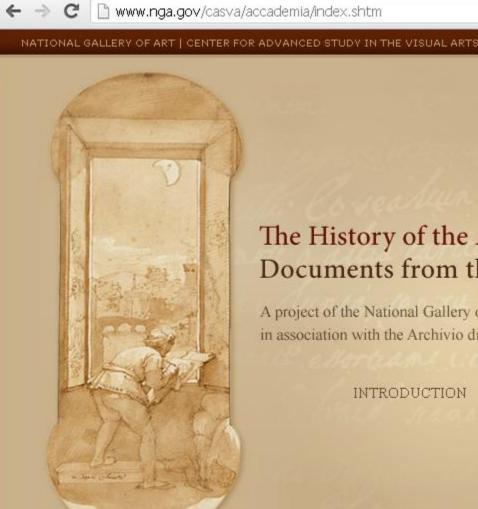
### Three Case Studies

- 1. <u>Medici Archive Project: Documentary Sources for the Arts</u> <u>and Humanities 1537 – 1743</u> (Relational database)
- 2. The History of the Accademia di San Luca, c. 1590 1635:

  Documents from the Archivio di Stato di Roma

  (Text markup TEI)
- 3. Healey Library's <u>OpenArchives</u> (login <u>server15774.contentdm.oclc.org</u>) (Dublin Core Schema in a proprietary data system)

### The History of the Accademia di San Luca, c. 1590–1635: Documents from the Archivio di Stato di Roma



www.nga.gov/casva/accademia/index.shtm





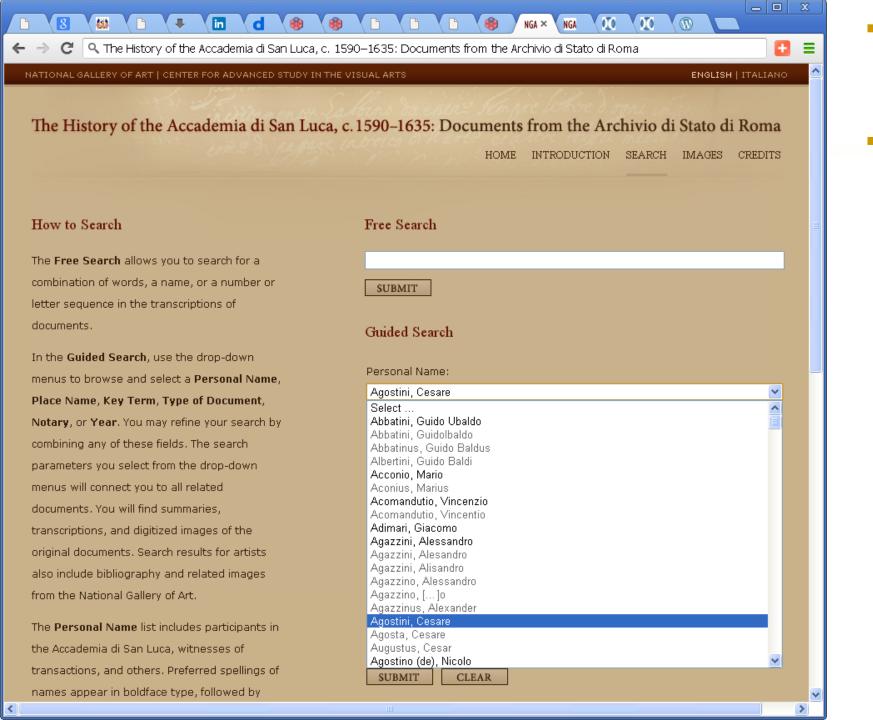


ENGLISH | ITALIANO

The History of the Accademia di San Luca, c. 1590-1635: Documents from the Archivio di Stato di Roma

A project of the National Gallery of Art, Center for Advanced Study in the Visual Arts. in association with the Archivio di Stato di Roma and the Accademia di San Luca

> SEARCH INTRODUCTION IMAGES CREDITS





#### **Document Transcription**

[fol. 210r]

Congregatio

Die 25 Mensis septembris 1611

IN Mei etc. Constituti personaliter, et Congregati pro secunda ut moris est in

loco infrascritto omnes infrascritti Videlicet

Ill.res DD. Cherubinus Albertus Princeps

Gaspar Cellius Consiliarius

<u>Durantes Albertus</u>

<u>Horatius Borgiannus</u>

Io'es Bap'ta de Novaria

Carolus Oldradus

Antivedutus Gramatica

<u>Hieronimus Masseus</u>

Marcellus Provenzanus, et

Hironimus Nanni

omnes de Congregatione Academica Pictorum, et sculptorum

^in <u>Ecclesia sancti Lucae</u>^

asserentes se esse numerum sufficientem ad infrascritta et

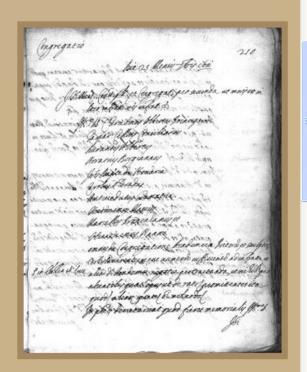
alia d*ict*ae Academiae negotia pertractanda, et nihil*ominus* pro

absentibus quatenus opus sit de rato etc. promittentes ita

quod etc. alias etc. sponte etc. omni meliori modo etc.

In primis decretarunt quod fieret memorialis Ill.mo D.

Io'i ipimage.sourceforge.net



ASR, TNC, uff. 15, 1611, pt. III, vol. 52, fol. 210r

```
<body>
  <pb xml:id="leaf01r" type="recto"/>
  <ld><ld type="poem">
     <head rend="underline" type="main-authorial">Af
       <del type="overstrike" seg="1">an</del>
       <add place="supralinear" type="insertion" se
       type="overstrike">unsolv'd</del> </add>
     </subst> argument</head>
     <1>
       <seg>
          <del type="overstrike">The</del>
          <add place="supralinear" type="insertion">
            <del type="overstrike">Coming in,</del>
            <subst>
               <del type="overwrite" seq="1">a</del>
               <add place="over" type="overwrite" seq="2">A </add>
            </subst>
          group of </add>
       little children, and their</seg>
       <seg>ways and chatter, flow
          <add place="inline" type="unmarked">in, </add>
          <del type="overstrike">
            <add place="supralinear" type="unmarked">upon me</add>
          </del>
       </seq>
     </1>
     <1>
       <seq>Like <add place="supralinear" type="insertion">welcome </add>
       rippling water o'er my </seq>
       <seg>heated <add place="supralinear" type="insertion">nerves and
       </add> flesh.</seq>
     </1>
     <closer>
       <signed>Walt Whitman</signed>
     </closer>
  </1q>
</body>
```

## Accademia di San Luca

The transcriptions are encoded in Extensible Markup Language (XML) following the guidelines of the Text-Encoding Initiative (TEI) for digital encoding of literary and linguistic texts (http://www.tei-c.org).

The transcriptions may be searched by personal name (under all known variant spellings), place name, key term, document type, notary name, and year. The documents included are not only newly available to students and scholars of early modern Italy, but are also accessible in a way that both promotes their use and allows for the identification of additional archival material.

Implementation description: We used TEI tagsets specific to manuscript material, personal names, and place names. Our programmer implemented Perl script to batch process the documents from MSWord to XHTML to XML to HTML. We tagged the XML files in TEI according to the standards of the Text Encoding Initiative (TEI) P5 Guidelines (November 2007).

Source: http://www.tei-c.org/Activities/Projects/hi01.xml

### Organizing Metadata

### **STRUCTURE**

- Which fields
- What they're called
- Whether they're required or not
- Best to base the structure on a standardized schema

### FIELD CONTENTS

- Guidelines
   ("in our project we enter names as "Last, First MI")
- Controlled
   vocabularies
   ("in our project we use
   LOCSH for subject terms
   and TGN for Places")

### Widely used schemas

■ MARC 91% *Used by librarians for library collections* 

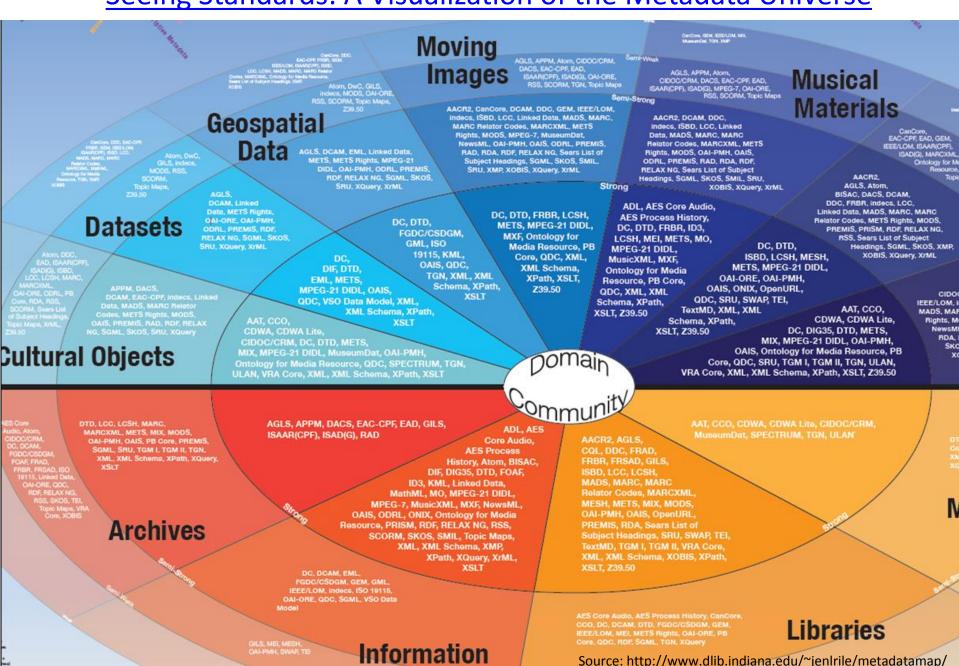
■ EAD 84% *Used by archivists for finding aids* 

■ Dublin Core 78% Used by cultural heritage practitioners for items

## Choose a schema (set of fields)

- □ A Metadata Schema is a "set of metadata elements designed for a specific purpose, e.g. describing a particular type of info resource for a specific community of practice."
- Many possible schemas!
  - □ Always best to adopt a schema already in use by many others
  - □ But you can also make up your own... if you must...
- □ **Choose** which to employ based on:
  - 1. your resource types
  - 2. your audience
  - 3. your community of practice

### Seeing Standards: A Visualization of the Metadata Universe



### Simple Dublin Core (metadata schema)

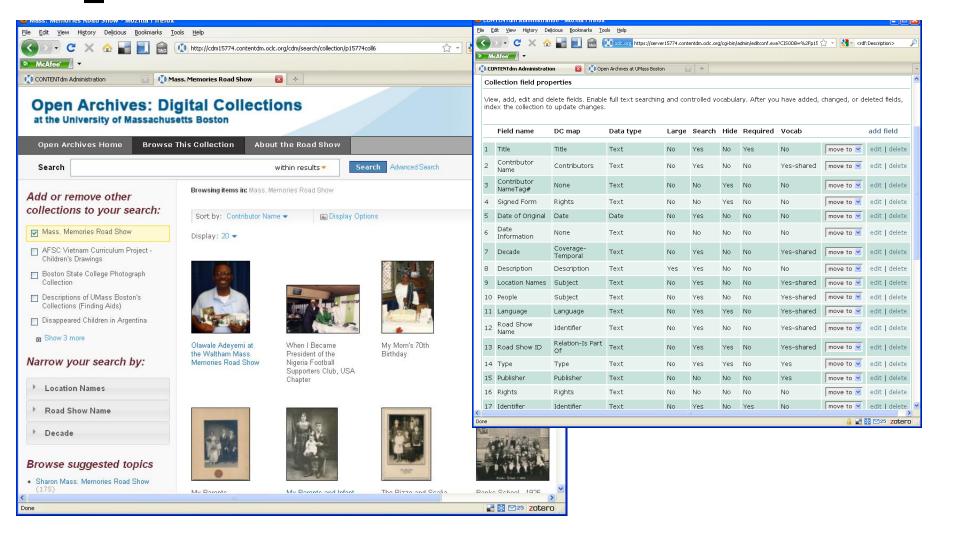
#### 15 metadata elements or fields:

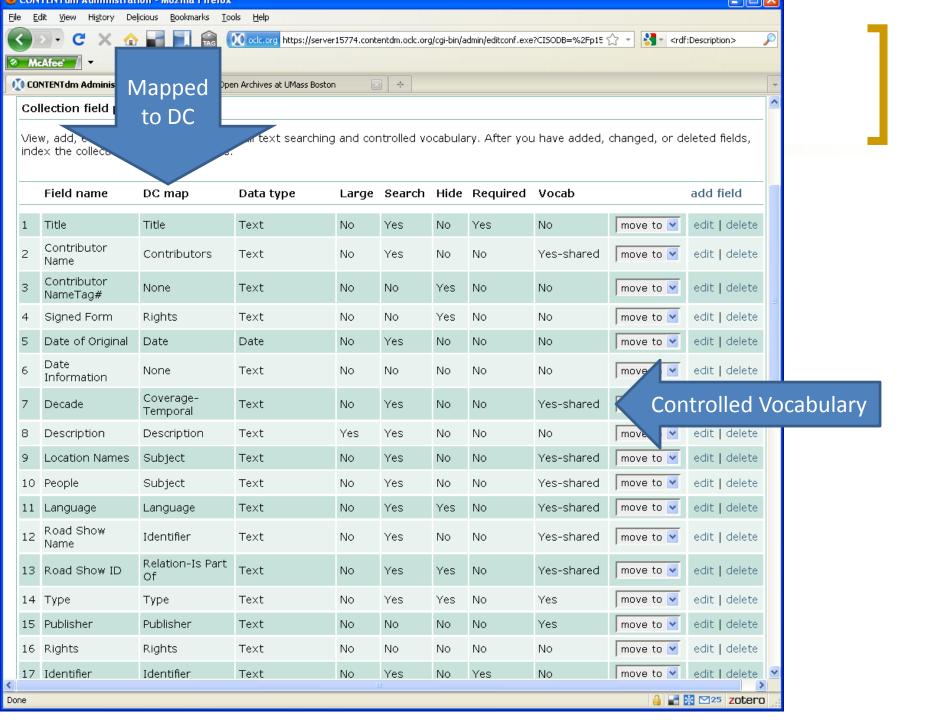
- 1. Title
- 2. Creator
- 3. Subject
- 4. Description
- 5. Publisher
- 6. Contributor
- 7. Date
- 8. Type
- 9. Format
- 10. Identifier
- 11. Source
- 12. Language
- 13. Relation
- 14. Coverage
- 15. Rights

←<u>Structural</u> Elements

But what controls the <u>content</u>?

### Metadata in Action at UMB





### The Case for Content Control

### LOCATION:

- Massachusetts?
- MA?
- Mass?

#### **DECADE:**

- 1950s?
- Early 1950s?
- 1950-1959?

#### PEOPLE:

- Mary Smith?
- Mary C. Smith?
- Smith, Mary?

#### **FORMAT:**

- Sound?
- Audio?
- MP3?

### The Case for Content Control



```
Cechov, Anton P. 1860-1904
Čehov, Anton Pavlovič (1860-1904). 📑 📟 🛄
Čechov, Anton Pavlovič [forme avant 2007] 76
Chejov, Anton Pavlovich 1860-1904
Čechov, Anton Pavlovič, 1860-1904
Tjechov, Anton, 1860-1904
Čechov, Anton Pavlovič
Чехов, Антон Павлович, 1860-1904
צ'כוב, אנטון פבלוביץ', 1860-1904 🖭
Tchekhoff, Anton, 1860-1904
💳 تشيكوف، انطون، 1860-1904 م
Tchekhov, Anton Pavlovitch, 1860-1904 [1]
MAE ID OFOADEDE
```

### Widely Used Content Vocabularies:

- AAT Art & Architecture Thesaurus Getty Institute
- LCSH Library of Congress Subject Headings
- LC Name Authority File *Library of Congress Name Auth.*
- MeSH Medical Subject Headings
- TGM Thesaurus of Graphic Materials (LOC)
- TGN Thesaurus of Geographic Names Getty Institute)
- ULAN -Union List of Artists' Names Getty)

### Project Profile – which fields should contain controlled content?

#### **STRUCTURAL ELEMENTS:**

- 1. Title
- 2. Creator
- 3. Subject
- 4. Description
- 5. Publisher
- 6. Contributor
- 7. Date
- 8. Type
- 9. Format
- 10. Identifier
- 11. Source
- 12. Language
- 13. Relation
- 14. Coverage e.g. spatial
- 15. Rights

#### POSSIBLE CONTROLLED CONTENT:

- 1. Title
- 2. Creator
- 3. Subject
- 4. Description
- 5. Publisher
- 6. Contributor
- 7. Date
- 8. Type
- 9. Format
- 10. Identifier
- 11. Source
- 12. Language
- 13. Relation
- 14. Coverage
- 5. Rights

### "Decade" element: Controlled Vocabulary – pick one!

1700-1709
1710-1719
1720-1729
1730-1739
1740-1749
1750-1759
1760-1769
1770-1779
1780-1789
1790-1799
1800-1809
1810-1819
1820-1829
1830-1839
1840-1849
1850-1859
1860-1869
1870-1879

1880-1889 1890-1899 1900-1909 1910-1919 1920-1929 1930-1939 1940-1949 1950-1959 1960-1969 1970-1979 1980-1989 1990-1999 2000-2009 2010-2019 2020-2029 2030-2039 2040-2049 2050-2059



### Digital Repositories

- There are decisions to be made before creating any digital resources project
  - Which schema to use (fields)?
  - Which content standards to apply?
- Those decisions are never clear-cut
  - Think them through, commit to standards and move on
- Standards (schemas and content) allow for sharing, finding, interoperability

#### How DigitalNZ works



Much of New Zealand's rich content is hidden or buried. We're working on new ways to create, describe, licence, store, surface & share NZ digital content.

Source: https://natlib.govt.nz/about-us/open-data/digitalnz-api

## Some Useful Resources

- *Understanding Metadata*, NISO Press. 2004.
- *Metadata*. SPEC Kit #298: Executive Summary. 2007.
- A Practical Application of Dublin Core:
   Worthington Memory. 2002. Allen and Davis.
- This presentation: Metadata Lecture\_Riley

  <a href="https://docs.google.com/leaf?id=0B68-">https://docs.google.com/leaf?id=0B68-</a>

  DLZjnv5INmI2MzU2NDYtZjA3NC00YzE2LWJIZjYtOWZjMDAwOTk5ZGE3

  &hl=en&authkey=CJDIgfgK