Understanding Metadata

Agenda

- March 4: Assignment 1 due
- Last Week Tonight
- Robocall killer
- Metadata
- Presenters (recorded)
  - Wenting Shang
  - Zichao Wang
  - Tabetha Pombo

What is Metadata?

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

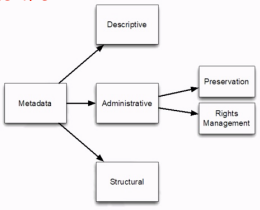
Understanding Metadata
National Information Standards Organization 2004

Music Example

Title: Unknown Georgian Folk Song
Duration: 03:03
Authors: Georgian Folk Singer
Audio channels: Stereo
Sample rate: 44.1 kHz
Album: USSR Bone Music Georgian Folk Song
Musical genre: Georgian Folk
Year recorded: 1950
Comment: Drum, accordian, and singer

Three main types of Metadata

- 1. *Descriptive Metadata* – Can be used to perform resource discovery and to identify data.
- 2. *Structural Metadata* – Can be used to define how data "fits together", e.g., what is the ordering of slides in a presentation, or what files are part of this package?
- 3. *Administrative Metadata** – Consists of management information about the data. For example when was it created, what type of file is it, who can access it, etc.



Answers

1. *Descriptive Metadata: It describes a specific cuneiform tablet*
2. *Structural metadata (what is the data about)*

Another example:

- Page numbers
- Sections
- Chapters
- Indexes
- Table of contents

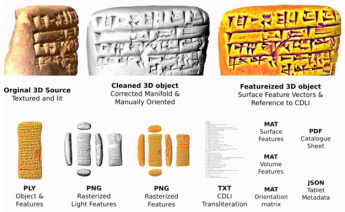
3. *Administrative Metadata: how to access data, what it is, etc.*

Agenda

- What is Metadata?
- Types of Metadata
- What does Metadata do?
- Specific Metadata Models and Tradeoffs
- Creating Metadata
- Quality Control for Metadata

Visual Example of descriptive metadata

Reading and understanding cuneiform tablets



What is Metadata (2)?

- "Metadata is data that provides information about other data. Two types of metadata exist: structural metadata and descriptive metadata. Structural metadata is data about the containers of data. Descriptive metadata uses individual instances of application data or the data content."-

Wikipedia https://en.wikipedia.org/wiki/Metadata

Questions

1. What type of metadata is the cuneiform description? descriptive ~, describe a specific form of tablet
2. What about a data dictionary? structural
3. What if I made metadata on fires:
   - When created: Feb. 17, 2022 administrative
   - Who can prevent it: Only you
   - Type of fire: Forest

Subsets of Administrative Metadata

- 1. Rights Management – Intellectual Property rights (license), and/or permissions (what groups have access to this data)
- 2. Preservation Metadata – Information about what an archive would require to preserve this data (what OS it was created on; with what program; what is the provenance of this data)
  - Related to Vint Cerf and Digital Vellum TED talk

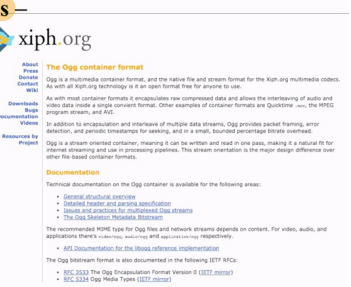
School of Engineering

Metadata can describe a single item of data, or a collection

- CCSDS - Consultative Committee on Space Data Systems
  - Reference Architecture for Space Information Management – CCSDS 312.0-G-1
  - Reference Model for an Open Archival Information System (OAIS) – CCSDS 650.0-M-2
- Data Object – "the data"
- Metadata Object – "metadata"
- Information Object – 1..* metadata, 1..1 data object

Package Formats

- **ZIP/CAB/BZ2/TAR Archives** – Single File, with many files "inside" of the package
  - Metadata describes the file "index", and names, and so forth
- **Container Formats**
  - **OGG** – Multimedia container format - https://xiph.org/ogg/
  - **MP4** – Multimedia Container format

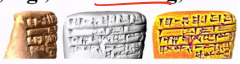


Some Design Choices

- **Embed the Metadata in a Single Object**
  - HTML documents embed metadata in HTML <head> tags

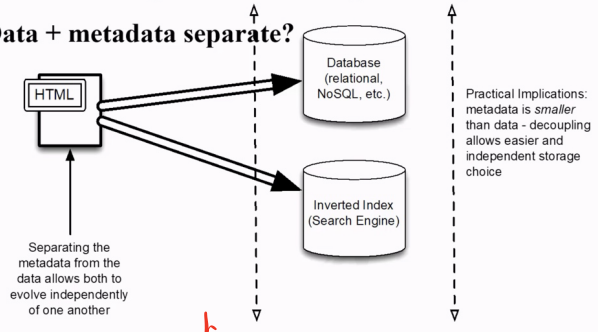
```
<head>
<meta charset="UTF-8">
<meta name="description" content="Free Web tutorials">
<meta name="keywords" content="HTML,CSS,XML,JavaScript">
<meta name="author" content="Hege Refsnes">
</head>
```

  - Sometimes this is not possible since object is immutable
- **Store the Metadata separately, e.g., in a Catalog, or Search Index, or Database**
  - E.g., cuneiform metadata



Storing Metadata separately

- **Data + metadata separate?**



What does Metadata Do?

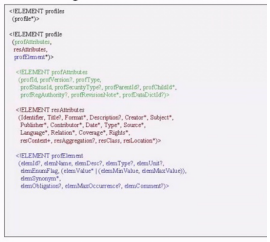
- 1. **Resource Discovery** – search and retrieval (e.g., within a certain *time* or *geospatial* range) – locate data resources in a network.
  - E.g., what is the song or tablet?
- 2. **Organizing Electronic Resources** – in Big Data realm, delineating files and content extremely important. Mostly done automatically with metadata. (compare analyzing raw web content to organized content)
- 3. **Interoperability** – Promote machine and human readability and understandability. Provide smaller summarization of data to computer programs, spiders, bots, agents, etc.
- 4. **Digital Identification** - identifying resources by a URL/URI, by a digital object identifier, or by PURL (persistent URL).
- 5. **Archiving and Preservation** – preserve information about formats, about software environment that data was generated in, and provenance necessary to enable reproducibility and lineage. (e.g., song)

Key Expectations: Resource Discovery

- Allow resources to be found by given criteria
- Identify resources (so that they can be found)
- Bring similar resources together (clustering)
- Distinguish dissimilar resources (deduplication and similarity)
- Give Location Information

Resource Discovery

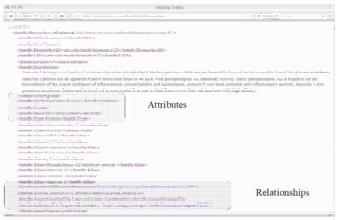
- Apache OODT Resource Profile System [Hughes2005, Hughes2010]
- ResAttributes
  - Dublin Core elements along with
  - Class (type) such as dataset, paper, software/ server
  - Context (project)
  - Aggregation (Singleton, Collection)
  - Location (URL or service)



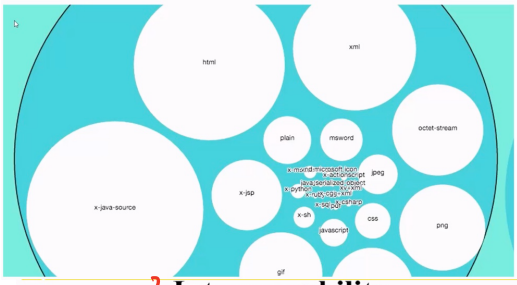
- J. S. Hughes, D. Crichton and C. Mattmann. Ontology-Based Information Model Development for Science Information Reuse and Integration. ITSSA, Vol 6., No. 2/3, pp. 200-211, August 2010.
- J. Steven Hughes, D. Crichton, S. Kelly, C. Mattmann, J. Crichton, and T. Tran. Intelligent Resource Discovery using Ontology-based Resource Profiles. Data Science Journal, Vol. 4, pp. 171-188, December 2005.

Resource Discovery (2)

- Modern representations
  - OODT profile represented as Resource Description Framework
  - RDF itself was a replacement
- Grid Resource Allocation Manager (GRAM)
- UDDI
- JSON, software as a service
- *Search Engines*



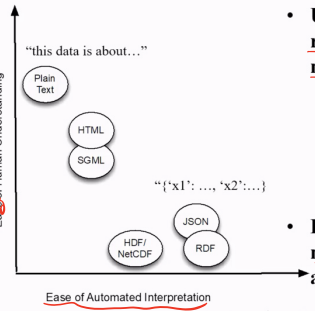
2. Organizing Electronic Resources



Organizing Electronic Resources

- May involve structural metadata and resource descriptive metadata
- Suggest common types
- Suggest commonality between metadata fields
  - All files edited in a similar way
  - All files generated in a similar way (e.g., folders)
- Suggests ways of interchanging *metadata*
- Too difficult to do this by hand with Big Data, must do so automatically

3. Interoperability



- Use common formats to represent and exchange metadata
  - Tradeoff here – represent metadata in such a way that a human can easily read and interpret it, but so that a machine can do so automatically
- Ensure minimal loss of metadata during automated processing

Interoperability (2)

- Two approaches for metadata interoperability
  1. **Cross System Search (Federated Search)**
     - Z39.50 protocol commonly used for this
     - https://en.wikipedia.org/wiki/Z39.50
     - New Efforts
       - ZING 39.50 REST-based with HTTP rather than custom protocol
       - OAI, SPARQL
  2. **Metadata Harvesting**
     - Open Archives Initiative (OAI) and its harvesting protocol (OAI-PMH 1.0)
     - https://en.wikipedia.org/wiki/Protocol_for_Metadata_Harvesting
     - Grew out of Santa Fe Convention in 1999 proposed by Herbert Van de Sompel (Ghent University), extended by CONI and DLF to OAI initiative

Interoperability (3)

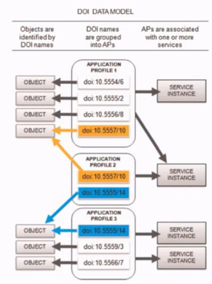
- Naming Challenges
  - Title
  - Name
  - Headline
- What is the *right name?*
- How do we mitigate naming differences in metadata fields
  - Preserve mapping between model and name

4. Digital Identification

- Digital Object Identifiers (DOIs)
  - Increasingly common in science and in literature (digital libraries)
- Persistent URLs (pURL) – aka "permalinks" such as DOI.org for papers
- Uniform Resource Locator (URL)
  - Introduced by Tim Berners Lee in RFC 1738
  - http://www.hjp.at/doc/rfc/rfc1738.html
  - URLs written as:
    - <scheme>:<schema specific type>
    - Common Internet Scheme Syntax: //<user>:<password>@<host>:<port>/<url-path>

DOIs

- Consist of 4 parts
  - Syntax specification defining the construction of a String/DOI name - ANSI/NISO Z39.84-2005
    - prefix / suffix e.g., 10.1000 / 123456
  - Resolution component
    - Registry that maps DOI name to URL
  - Metadata component
    - Minimal recommended set, e.g., Identifier, and identification of domain data dictionary
  - Social infrastructure for federation
    - Identify management responsibility



Paskin, Norman. "Digital object identifier (doi) system." Encyclopedia of library and information sciences 3 (2008): 1586-1592.

5. Archiving and Preservation

- Digital information is fragile
  - Sometimes, human oriented: Apollo 11 missing tapes https://en.wikipedia.org/wiki/Apollo_11_missing_tapes
  - Sometimes due to bit rot
    - http://www.economist.com/node/21553445

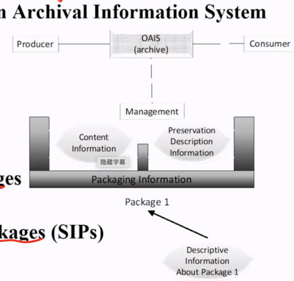
Archiving and Preservation

- Strong focus on *reproducibility*
  - *for example, know what the Georgian song is about*
  - All the information required to reproduce the object
  - Lineage
    - Where it came from
    - The operations that changed it over time
    - Also known as "Provenance" (e.g., Overleaf.org preserves who wrote what)
    - W3C has a provenance group that produced a PROV model
    - https://www.w3.org/TR/prov-overview/
- Other organizations include National Library of Australia, British Cedars Project (CURL Exemplars in Digital Archives), Joint OCLC Working Group, Research Libraries Group (RLG)
- PREMIS (Preservation Metadata: Implementation Strategies)
  - http://www.loc.gov/standards/premis/

OAIS – Reference Model

small vision of metadata

- Reference Model for an Open Archival Information System (OAIS) – CCSDS 650.0-M-2
- Describes packaging model as well as archival model
- Information Objects
- Data Objects
- Archival Information Packages (AIPs)
- Submission Information Packages (SIPs)



Structuring Metadata

- Metadata Schemes (Schema)
  - Typically Associated with File Types
  - Prescribe Names of Elements and their semantics
  - Value Constraints
    - Discrete eg: 编写这个doc的人名
    - Continuous (or Range) eg: the time the doc edited
    - Controlled Vocabulary
- May prescribe encoding syntax
  - XML
  - JSON

① Classic Example: Dublin Core

of structural metadata

- Started at 1995 workshop sponsored by OCLC and NCSA in Dublin Ohio
- Continuing Development => Dublin Core Metadata Initiative
- Originally 13 core elements, grown to 15 later
  - *Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights*

Dublin Core Example

```
Title="Metadata Demystified"
Creator="Brand, Amy"
Creator="Daly, Frank"
Creator="Meyers, Barbara"
Subject="metadata"
Description="Presents an overview of
    metadata conventions in
    publishing."
Publisher="NISO Press"
Publisher="The Sheridan Press"
Date="2003-07"
Type="Text"
Format="application/pdf"
Identifier="http://www.niso.org/
    standards/resources/
    Metadata_Demystified.pdf"
Language="en"
```


Dublin Core Significance

- 15 elements to describe *any* electronic resource
- Generic yes, powerful yes
- First widespread use of common ISO-11179 to specify metadata elements
  - Range, Controlled Value set
  - Specification for expressing metadata in a registry
  - https://en.wikipedia.org/wiki/ISO/IEC_11179
- Inspired Metadata capture syntax in
  - RDF, OODT Profiles, Schema.org, etc.

Text Encoding Initiative (TEI)

- Develop guidelines for marking up electronic texts including novels, plays and poetry primarily to support human research in the humanities
  - http://www.tei-c.org/index.xml
- TEI specified a *Guidelines for Electronic Text Encoding and Interchange*
  - Specifies a header portion of the resource that consists of metadata about the work
  - TEI defined as SGML syntax and rules and in a Document Type Definition (DTD)
  - *TEI Lite* often used b/c TEI specification is so large

Metadata Encoding and Transmission Standard (METS)

- Standard data structure for digital library objects
- XML Schema for XML documents representing digital library objects
  - Descriptive Metadata
  - Administrative Metadata
  - Names and Locations of files in the Digital Object
- Originally the outgrowth of the *Making of America II project* http://www.loc.gov/standards/mets/
  - Digitization project of major research libraries for textual and image based works

Learning Object Metadata (LOM)

- IEEE Learning Technology Standards Committee (LTSC)
- https://en.wikipedia.org/wiki/Learning_object_metadata
- Learning Object Metadata (IEEE 1484.12.1-2002)
  - Minimal set of attributes required to manage, locate, and evaluate learning objects such as computer-based training and distance learning
- Information types
  - General – about the object as a whole
  - Lifecycle – evolution of the object
  - Technical – technical detail and requirements
  - Educational – pedagogical attributes
  - Rights – intellectual property rights
  - Relation – related objects
  - Annotation – Date/Author/Comments

Geospatial Metadata Schemata

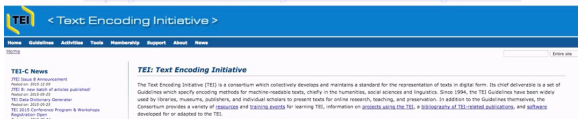
- Federal Geographic Data Committee
  - Specifies FGDC-STD-001-1998 for geographic datasets which include GIS systems and cartography based files
- ISO 19115
  - Geographic Information Metadata
  - More commonly used today as it allows both ISO and FGDC compliance (includes Dublin Core information)
  - Location (latitude/longitude), Dataset Description information, provenance
- Both are specified in XML

Quality Control for Metadata

- *Framework of Guidance for Building Good Digital Collections*
  - *http://www.niso.org/apps/group_public/workgroup.php?wg_abbrev=digcoll*
- Some suggestions about metadata checks and "good" metadata

TEI continued

- TEI specifies
  - Header – used to record bibliographic information about both electronic and non electronic version of the text
  - TEI extended for citations and remainder of the work
- Grobid Extraction System uses TEI as its core metadata format
  - http://grobid.readthedocs.org/en/latest/TEI-encoding-of-results/



Metadata Object Description Schema

- MODS – derivative of MARC 21
  - MARC (Machine Readable Cataloging) is a standards and specification for digital formats cataloged by libraries, e.g., books
- Includes subset of MARC fields
- Expressed using XML syntax http://www.loc.gov/standards/mods/

A MODS Record Example

```
<mods>
    <titleInfo>
        <title>Metadata demystified</title>
    </titleInfo>
    <name type="personal">
        <namePart type="family">Brand</namePart>
        <namePart type="given">Amy</namePart>
        <role>
            <roleTerm authority="marcrelator" type="text">author</roleTerm>
        </role>
    </name>
    <typeOfResource>text</typeOfResource>
    <originInfo>
        <dateIssued>2003</dateIssued>
        <place>
            <placeTerm type="text">Bethesda, MD</placeTerm>
        </place>
        <publisher>NISO Press</publisher>
    </originInfo>
    <identifier type="isbn">1-880124-59-9</identifier>
</mods>
```

MPEG Multimedia Metadata

- ISO/IEC Moving Pictures Expert Group (MPEG)
- MPEG-7 defines metadata elements, structures, and relationships used to described audiovisual objects https://en.wikipedia.org/wiki/MPEG-7
- MPEG-21 framework brings it together
  - Vision, Technologies and Strategy
  - Digital Item Declaration
  - Digital Item Identification
  - Intellectual Property Management and Protection
  - Rights Expression Language
  - Rights Data Dictionary
  - Digital Item Adaptation

Creating Metadata

- How do we create metadata?
- Creation (Curation) Tools
  - Templates – allows a user to generate metadata based on a template.
  - Mark-up tools – structure metadata attributes in its representation format (XML, JSON, etc.) according to a schema.
  - Extraction Tools – unlocks metadata from existing file types and/or external metadata catalogs.
  - Conversion tools – converts metadata from one format to another.

"Good" Metadata

- Appropriate to the materials, and users in the collection and describe the object's intended use
- Supports interoperability (common representation, and abilities to resolve naming conflicts)
- Uses standard controlled vocabularies
- Identifies terms and conditions of use
- Is an object itself and allows for unique identification and preservation
- Allows for long term management of objects in collections

Some Challenges for "Good" Metadata

- User interface tools and support for Data Curators
- "Crosswalks" between different metadata schemata and software and approaches to deal with that
- Data Curator training
- Support for MIME types taxonomy in classification
- Keeping up with the guidelines and metadata specifications

Key Resources

- NISO.org, Understanding Metadata http://www.niso.org/publications/press/UnderstandingMetadata.pdf
- CCSDS - Consultative Committee on Space Data Systems
  - Reference Architecture for Space Information Management – CCSDS 312.0-G-1
  - Reference Model for an Open Archival Information System (OAIS) – CCSDS 650.0-M-2

Summary

- What is Metadata?
- Types of Metadata
- What does Metadata do?
- Specific Metadata Models and Tradeoffs
- Creating Metadata
- Quality Control for Metadata