4 Metadata, markup, and data description

This chapter delves into the crucial role of metadata in classifying, describing, and organizing information across various domains. It underscores how different fields have unique metadata standards, from bibliographic records in libraries to cultural objects, which embody local customs and expertise. This chapter explores the use of structured data and formal languages such as XML, TEI, and JSON, highlighting their flexibility and value in managing and transmitting information. It also addresses the inherent challenge of ambiguity in humanistic documents and the ethical considerations of reducing complex, multifaceted information into rigid, machine-readable formats.

4a Metadata and classification

Exercise 4.1: Standard metadata

Try applying the Dublin Core metadata standard (https://www.dublincore.org/specifications/dublin-core/usageguide/elements/) to documents or objects in an academic field or discipline. Make records for half a dozen of the documents or artifacts. What information that is relevant to your work cannot fit into this scheme? Which elements of the scheme are confusing or difficult to use consistently?

How-to example

Applying the Dublin Core metadata standard to documents or artifacts within a specific academic field, such as art history, involves creating structured records for each object or document using the fifteen core elements of Dublin Core. Below is an example of how Dublin Core might be applied to six artifacts or documents in this field.

Example records:

1. **Object**: Oil painting—*Starry Night* by Vincent van Gogh
   - **Title**: Starry Night
   - **Creator**: Vincent van Gogh
   - **Subject**: Post-Impressionism, Landscape, Night Sky
   - **Description**: An iconic oil painting depicting a swirling night sky over a quiet village.
   - **Publisher**: N/A
   - **Contributor**: Museum of Modern Art
   - **Date**: 1889
   - **Type**: Physical object—painting
   - **Format**: Oil on Canvas, 73.7 cm × 92.1 cm
   - **Identifier**: MoMA_472.1941
   - **Source**: N/A
   - **Language**: N/A
   - **Relation**: N/A
   - **Coverage**: Saint-Rémy-de-Provence, France
   - **Rights**: Public Domain
2. **Document**: Research article—Imagining: Creating Spaces for Indigenous Ontologies
   - **Title**: Imagining: Creating Spaces for Indigenous Ontologies
   - **Creator**: Marisa Elena Duarte and Miranda Belarde-Lewis
   - **Subject**: Indigenous knowledge, knowledge organization, cataloging, classification of knowledge, Indigenous peoples
   - **Description**: A scholarly article exploring how the decolonizing methodology of imagining can guide knowledge organization toward Indigenous community-based approaches for developing alternative information structures.

   - **Publisher**: *Cataloging & Classification Quarterly*
   - **Contributor**: N/A
   - **Date**: 2015
   - **Type**: Text—journal article
   - **Format**: PDF
   - **Identifier**: doi:10.1080/01639374.2015.1018396
   - **Source**: N/A
   - **Language**: English
   - **Relation**: N/A
   - **Coverage**: Contemporary cataloging practice
   - **Rights**: Copyright © 2015 Marisa Elena Duarte and Miranda Belarde-Lewis

3. **Object**: Textile—Brazil, indigenous cloak (https://apnews.com/article/brazil-indigenous-cloak-lula-denmark-repatriation-artifact-6fbf5074e4b332b4a795e94ed44064b5)

   - **Title**: Tupinambá Feather Cloak
   - **Creator**: Tupinambá People
   - **Subject**: Indigenous Brazilian Culture; Ceremonial Garments
   - **Description**: A rare ceremonial cloak made from the vivid red feathers of the scarlet ibis, traditionally worn by the Tupinambá people of Brazil for rituals and as a symbol of authority.
   - **Publisher**: N/A
   - **Contributor**: National Museum of Denmark
   - **Date**: 17th century (estimated)
   - **Type**: Physical object—ceremonial garment
   - **Format**: Featherwork; Scarlet ibis feathers and natural fibers
   - **Identifier**: DK_NMD_TupinambaCloak_17C
   - **Source**: National Museum of Denmark
   - **Language**: N/A
   - **Relation**: Part of repatriation agreement between Brazil and Denmark
   - **Coverage**: Brazil (origin); Denmark (historical collection)
   - **Rights**: Repatriated to Brazil; held under Brazilian cultural heritage protection

4. **Document**: Dissertation—*The Inner Workings of Brush and Ink: A Study on Huang Binhong (1865–1955) as Calligrapher with Special Respect to the Concept of Interior Beauty (neimei)*

   - **Title**: The Inner Workings of Brush and Ink: A Study on Huang Binhong (1865–1955) as Calligrapher with Special Respect to the Concept of Interior Beauty (neimei)
   - **Creator**: Shoa-Lan Hertel
   - **Subject**: Huang Binhong; Chinese Calligraphy; Literati Aesthetics; Interior Beauty (neimei)
   - **Description**: A dissertation examining the calligraphy of Huang Binhong, analyzing how the artist embodies the Chinese aesthetic ideal of "interior beauty" (neimei) through his brushwork and philosophical approach to art.
   - **Publisher**: Academia.edu (online repository); Freie Universität Berlin
   - **Contributor**: Huang Binhong (primary subject)
   - **Date**: 2017
   - **Type**: Text—Dissertation/thesis
   - **Format**: PDF; digital text
   - **Identifier**: academia.edu/40223608
   - **Source**: Academia.edu (https://www.academia.edu/40223608/The_Inner_Workings_of_Brush_and_Ink_A_Study_on_Huang_Binhong_1865_1955_as_Calligrapher_with_Special_Respect_to_the_Concept_of_Interior_Beauty_neimei_)
   - **Language**: English
   - **Relation**: Related to studies on Chinese literati painting and calligraphy; references works by and about Huang Binhong
   - **Coverage**: China (historical and cultural focus); 19th–20th century
   - **Rights**: © Shoa-Lan Hertel; all rights reserved by the author

5. **Object**: Manuscript—*Leonardo's Notebook Codex Forster II*

   - **Title**: *Codex Forster II*
   - **Creator**: Leonardo da Vinci
   - **Subject**: Renaissance, Scientific Studies, Anatomy, Inventions
   - **Description**: A collection of Leonardo da Vinci's notes, sketches, and studies on various topics.

- **Publisher**: N/A
- **Contributor**: Victoria and Albert Museum
- **Date**: c.1495–1497
- **Type**: Physical object—manuscript
- **Format**: Paper, Ink, Various Dimensions
- **Identifier**: MSL/1876/Forster/141/II
- **Source**: N/A
- **Language**: Italian
- **Relation**: N/A
- **Coverage**: Italy, Renaissance
- **Rights**: Public Domain

6. **Document**: Article—*Africanizing A Modern African Art History Curriculum from the Perspectives of an Insider, Freeborn O. Odiboh*

    - **Title**: *Africanizing A Modern African Art History Curriculum from the Perspectives of an Insider*
    - **Creator**: Odiboh, F. O.
    - **Subject**: Modern African Art; Art History Curriculum; Decolonization; African Studies; Pedagogy
    - **Description**: This article critiques the continued marginalization and misinterpretation of modern African art within Western scholarship and calls for a reform of African art history curricula in tertiary institutions. It argues for African-centered terminologies, classifications, and approaches that reflect indigenous knowledge systems and experiences. Drawing from an insider perspective, the article advocates for a historically grounded, unbiased, and developmentally relevant art history pedagogy across Africa.
    - **Publisher**: International Association of African Researchers
    - **Contributor**: *African Research Review* (journal); *African Journals Online* (AJOL—digital host)
    - **Date**: 2009-06-23
    - **Type**: Text—peer-reviewed journal article
    - **Format**: PDF; application/pdf
    - **Identifier**: OCLC: 806758989; URL: https://www.ajol.info/index.php/afrrev/article/view/43591/27114
    - **Source**: *African Research Review*, Vol. 3, No. 1 (2009)
    - **Language**: English
    - **Relation**: Related to discourses on decolonizing African art history and higher education curricula
    - **Coverage**: Africa; Postcolonial academic contexts
    - **Rights**: © F. O. Odiboh; distributed by the International Association of African Researchers

Relevant information that cannot fit into Dublin Core:

1. **Paradata (data on process)**: Details on how the artifact or document was created, digitized, or manipulated (e.g., what tools were used, how it was processed, or decisions made during digitization), don't have a specific place within Dublin Core.
2. **Ethical considerations**: For data associated with sensitive or contemporary subjects (e.g., privacy issues, ethical concerns, or re-use considerations), Dublin Core lacks fields that directly address these dimensions.
3. **User interaction or engagement data**: If the document or artifact is part of an interactive digital exhibition, data on how users engage with the content (e.g., clicks, duration spent) is not covered in Dublin Core.
4. **Multiple provenance layers**: While Dublin Core allows for capturing some provenance details (like *Creator* and *Contributor*), it can be difficult to track complex layers of provenance (e.g., successive ownership or changes in custodianship over time) without additional metadata standards.

Confusing or difficult elements:

1. **Relation**: The *Relation* element can be ambiguous when documenting objects or texts that have complex relationships with other works. Determining how to define and standardize relationships can be challenging, especially for interdisciplinary or multimedia projects.
2. **Coverage**: *Coverage* is another potentially confusing element, as it can refer to either temporal or spatial coverage. For certain objects (e.g., art), deciding between geographic location, time period, or cultural scope may require additional clarity or guidelines for consistent application.
3. **Source**: The *Source* field is tricky when documenting original objects versus derivative or digital copies. Distinguishing the role of *Source* from other fields like *Format* or *Identifier* can be confusing when dealing with reprints, reproductions, or digital surrogates.

Exercise 4.2: Customized metadata

Start creating a taxonomy and/or classification system for an area in which you have a high level of knowledge—such as your favorite music, family photographs, memorabilia, or other personal collection. What terms, references, or resources would you want to cross-reference repeatedly? Which should be in a pick list, so you could use them consistently? Is Dublin Core sufficient?

How-to example

Let's say I'm creating a taxonomy and classification system for a collection of family photographs and memorabilia. This system would need to capture a wide range of data, including people, places, dates, events, and types of media. Here's an approach to developing a taxonomy and identifying key terms and categories, as well as how Dublin Core might be integrated.

Primary categories (top-level classification):

1. **People**
    - **Names** (e.g., family members, ancestors, significant others)
    - **Generations** (e.g., grandparents, parents, children)
2. **Places**
    - **Locations** (e.g., home addresses, vacation spots, countries)
    - **Historical locations** (e.g., old family homes, ancestral villages)
3. **Dates**
    - **Exact dates** (e.g., January 1, 2000)
    - **Time periods** (e.g., 1990s, World War II era)
    - **Event-specific dates** (e.g., birth, weddings, anniversaries)
4. **Events**
    - **Family events** (e.g., birthdays, weddings, reunions)
    - **Holidays** (e.g., Christmas, Thanksgiving)
    - **Milestones** (e.g., graduation, first steps)
5. **Types of media**
    - **Photographs** (e.g., printed, digital)
    - **Documents** (e.g., birth certificates, letters)
    - **Videos** (e.g., home videos, digital clips)
    - **Memorabilia** (e.g., souvenirs, childhood toys)
6. **Themes**
    - **Life events** (e.g., childhood, coming of age, retirement)
    - **Travel** (e.g., vacations, trips abroad)
    - **Cultural or religious traditions** (e.g., baptisms, Diwali)

Subcategories and terms for cross-referencing:

- **Photograph types**: Candid, portrait, group photo, landscape
- **Memorabilia types**: Keepsake, heirloom, awards, letters
- **People by relationship**: Parent, sibling, cousin, in-law
- **Locations by type**: Home, vacation, landmark, country of origin

Pick list fields (controlled vocabularies):

- **People's names**: Ensure consistency in naming (e.g., "John M. Doe" instead of variations like "Johnny" or "J. Doe")
- **Events**: Predefined event types like "wedding," "anniversary," and "holiday celebration"
- **Locations**: Standardized place names (e.g., "New York City, NY" instead of "NYC" or "New York")
- **Media types**: Consistent labeling (e.g., "Photograph," "Video," "Document")
- **Time periods**: Fixed time ranges (e.g., "1990–99," "2000–09")

Dublin Core fields and suitability for this system:

Dublin Core is generally useful for basic metadata like **title, creator, date, and format**, but it may not be entirely sufficient for this personal collection. Here's how Dublin Core elements apply:

Dublin core element	*How it applies*
Title	Title of each photograph, video, or item (e.g., "Wedding 1995" or "Grandfather's Medal")
Creator	The person who took the photograph or created the document
Subject	Relevant subjects like "family reunion," "graduation," "birthday"
Description	Brief description of the photo or item (e.g., "Grandma's Eightieth birthday party")
Date	Date the photograph or item was created (e.g., "March 15, 1990")
Type	Type of media (e.g., "photograph," "video," "document")
Format	Physical or digital format (e.g., "JPEG," "printed photo," "MP4")
Identifier	Unique ID for each item (e.g., "Photo_00123")
Rights	Ownership rights or copyright, if applicable (e.g., "family-owned")
Coverage	Geographic coverage, such as the location where the photo was taken (e.g., "Paris, France")
Language	For documents or videos with spoken/written language, specify language (e.g., "English")

What Dublin core lacks for this use case:

1. **People relationships**: Dublin Core doesn't have a specific way to indicate familial relationships. For a family collection, you may need to track how individuals are related (e.g., mother, father, cousin), which would require a custom field.
2. **Emotional or sentimental value**: It's important in personal collections to capture the *emotional significance* or *sentimental value* of an item, which Dublin Core doesn't address. A custom field could be created for "emotional context" or "family importance."
3. **Multiple layers of description**: Some items, like a photograph with multiple people, events, and places, require more detailed, hierarchical descriptions. Dublin Core is flat and doesn't handle these complexities well.
4. **Versions/updates**: For collections that may be digitized or modified (e.g., adding restoration details to an old photo, like making it darker or lighter so you can see people better), it's important to track versions, which is not well-handled by basic Dublin Core.
5. **Privacy considerations**: Personal family collections may include sensitive information (e.g., private letters, medical records), and Dublin Core doesn't offer fields to handle privacy concerns or restrictions on viewing.

Additional fields beyond Dublin core:

To enhance the taxonomy, I would add the following custom fields:

- **Relationship to owner**: How the people in the photo or item are related to the owner (e.g., "grandparent," "sibling").
- **Emotional value**: A subjective field for the family to note the emotional or historical significance of the item.
- **Restoration history**: If photos or memorabilia were restored, this would track what was done (e.g., "digitally restored in 2023").
- **Condition**: Document the physical or digital condition of the item (e.g., "fragile," "digitally restored").

Example record using this taxonomy:

Field	*Example entry*
Title	"Grandparents' wedding photo"
Creator	Unknown (likely a family member)
People	John Doe, Mary Doe
Relationship to owner	Grandparents
Event	Wedding
Date	June 12, 1950
Location	New York City, NY
Format	Black and white photograph, 8 × 10 inches
Type	Physical object—photograph
Identifier	Photo_0001
Condition	Fragile, slight fading
Emotional value	This is one of the few photos of the wedding ceremony.
Rights	Family-owned
Restoration history	Digitally scanned in 2019 for preservation
Accessible version	A relief version of this image has been 3D printed.

Dublin Core provides a solid base for cataloging basic information like titles, creators, dates, and formats. However, for personal collections like family photographs and memorabilia, it lacks the granularity needed for handling relationships, emotional significance, privacy, and item condition. Therefore, a combination of Dublin Core and custom fields would be ideal for an effective taxonomy system in this case.

Exercise 4.3: Multiple ontologies

What area of your own experience—health, emotional conditions, knowledge of a neighborhood or travel route, etc.—lends itself to the concept of "fluid" ontologies? Why would different naming conventions be of use?

How-to example

One area of experience that lends itself well to the concept of "fluid" ontologies is travel. Travel experiences can vary widely based on numerous factors such as personal preferences, cultural contexts, modes of transportation, and the dynamic nature of destinations. Here's how fluid ontologies manifest in this context and why different naming conventions would be useful:

Fluid ontologies in travel:

1. **Dynamic contexts**:
   - **Travel preferences**: Each traveler has unique interests, such as adventure, relaxation, history, or food. These preferences can change over time or based on different trips, making travel categories fluid.
   - **Cultural sensitivity**: Places may have different meanings and significance to various cultures, which can affect how travelers interact with them.
2. **Evolving experiences**:
   - **Travel routes**: The popularity of routes can shift due to factors like tourism trends, local events, or environmental changes (e.g., road closures, new attractions).
   - **Changing accommodations**: Different lodging options (hotels, hostels, Airbnbs) may evolve or appear over time, requiring flexible classifications.
3. **Personal growth**:
   - As people travel more and experience new places, their perspectives can shift. What once was a priority may become less important, or vice versa. This evolution could affect how they categorize their travel experiences or recommendations.
4. **Mobility issues**:
   - What are the wheelchair accessible options for the route? Where would a mobility-impaired or limited individual find obstacles? Are there adequate facilities to accommodate wheelchair-bound persons? Do the maps indicate their location?

Use of different naming conventions

1. **Contextual relevance**:
   - Different naming conventions can help tailor information to specific audiences. For instance, a travel guide for families might use terms like "kid-friendly" or "family-oriented," while a guide for backpackers might focus on "budget-friendly" or "off-the-beaten-path."
2. **Cultural appropriateness**:
   - Names and terms can carry different connotations in various cultures. Using localized naming conventions ensures sensitivity and respect, helping travelers connect more meaningfully with their destinations.
3. **Specialization**:
   - Travel can be categorized in many ways (e.g., eco-tourism, luxury travel, cultural heritage). Different naming conventions allow for the creation of specialized subsets within broader travel categories, helping travelers find exactly what they're looking for.

4. **Facilitating communication**:

    - When discussing travel with friends or on social media, using fluid naming conventions allows for richer conversations. For example, one might use “hidden gems” to describe lesser-known attractions rather than just “sights,” enhancing the discussion.

5. **Organizational flexibility**:

    - For travel blogs or apps, using different naming conventions can improve user experience. Tags like “food experiences,” “adventure activities,” or “cultural insights” can be fluidly applied to various destinations, making it easier for users to find relevant content.

Example of fluid ontology in travel:

- **Categories**:
    - **Type of travel**: Adventure, leisure, cultural, business
    - **Location type**: Urban, rural, coastal, mountainous
- **Naming conventions**:
    - **Adventurous activities**: “Hiking” vs. “trekking” vs. “mountain climbing”
    - **Dining options**: “Fine dining,” “street food,” “local cuisine,” “vegan-friendly”
    - **Transport modes**: “Public transit,” “car rentals,” “bicycle tours”

The fluidity of ontologies in travel reflects the diverse and changing nature of experiences. Different naming conventions enhance communication, increase relevance, and facilitate exploration in a way that resonates with individual preferences and cultural contexts. This adaptability is key to creating meaningful and relevant travel experiences, whether in personal reflection, travel planning, or community sharing.

Recommended readings

Baca, Murtha. 2016. “Intro to Metadata.” In *Introduction to Metadata*, edited by Murtha Baca. 3rd ed. Los Angeles: Getty Publications. https://www.getty.edu/publications/intrometadata/.

Hoffman, Gretchen. 2013. “How Are Cookbooks Classified in Libraries?” *NASO* 4 (1). www.researchgate.net/publication/272962458_How_are_Cookbooks_Clas sified_in_Libraries_An_Examination_of_LCSH_and_LCC.

Kardos, Ann. 2024. *Introduction to Metadata*. Librarires, Amherst: University of Massachusetts. https://guides.library.umass.edu/intro_to_metadata

Liu, Fangchao, and John Hindmarch. 2023. “A Review of the Cultural Heritage Linked Open Data Ontologies and Models.” *The International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences.* XLVIII-M-2. 10.5194/isprs-archives-XLVIII-M-2–2023–943–2023.

Miller, Steven J., and Chartered Institute of Library and Information Professionals (Great Britain). 2022. *Metadata for Digital Collections*. Second edition. UK edition. London: Facet Publishing.

Parthenos Project. “Formal Ontologies: A Complete Novice’s Guide.” https://training.parthenos-project.eu/sample-page/formal-ontologies-a-complete-novices-guide/.

Tibaut, Andrej, and Sara Guerra de Oliveira. 2022. “A Framework for the Evaluation of the Cultural Heritage Information Ontology.” *Applied Sciences* 12 (2). https://www.mdpi.com/2076-3417/12/2/795.

4b Markup: XML, TEI, KML, JSON, IIIF, and other standards

Bonus content: discussion of TEI, a special subset

One form of XML termed TEI (Text Encoding Initiative) takes its name from the group of scholars who were responsible for creating it as a standard for marking up texts. As in the case of other professional organizations who maintain standards, the TEI is a consortium of nonprofit groups who support and guide its activities. Their guidelines were first released in 1994, and are meant, like other standards, to make digital documents interoperable. That means that the use of standard tags in accord with set rules and protocols allows these materials to be integrated into search, analysis, and research. The TEI recognizes that customization is an aspect of humanities research, and their structure allows for and supports this. **[Exercise 4.5: TEI]**

Bonus content: Encoded Archival Description, an alternative for archives

As networked platforms became more common, the requirements for standardized markup became more important. Another standard commonly used in the humanities is Encoded Archival Description, or EAD. It is different from

TEI in its attention to the physical features of archival objects and it has its own vocabulary for description. OAI, the Open Archives Initiative is used to “harvest” information, particularly metadata, for integration into specialized, discipline-specific, search engines. Markup is ubiquitous in digital publishing and scholarship.

Bonus content: JSON: describing objects rather than inserting markup

Markup is exceptionally well suited to the analysis and interpretation of textual elements, and thus used in literary and humanities work, as well as critical editing in all fields. XML formats are also used to create metadata and thus to structure and describe the contents of records as was discussed in the first part of this chapter. But collections of digital collections lend themselves to other kinds of structured data description in file formats designed specifically for information exchange of data. The most popular of these is JSON format.

Not all data formats are text, and many files created in JavaScript contain information that cannot be readily transferred between a server and a browser and won’t fit easily into XML. JSON (JavaScript Object Notation) was developed to address this problem. It is text-based but can describe many data formats. It is thus a very useful standard with which to be familiar to facilitate exchange. JSON is used for storage as well as exchange and has the advantage that it is relatively simple to learn and is highly customizable.

JSON’s notation scheme was designed so that it could describe data objects effectively. That means that its syntax contains ways to identify data types so they are not confused with each other. Also, JSON’s syntax is very similar to that of Python and other languages used to develop web applications, making it easy for those familiar with programming to learn and use it. Metadata, classification, and markup schemes have certain features in common. They are all ways of disambiguating data types, formats, and structures so that they can be processed and either analyzed or displayed properly—or exchanged and integrated into new projects. The crucial principle is that code of any kind conforms to formal rules that disambiguate data, content, and information to make their statements computable.

Where XML and TEI are used to markup parts of texts—either thematically or formally, as we have seen—JSON syntax identifies types of structured data. It builds on JavaScript but is a notation system (hence its name) for describing or representing what is in a data file. XML and JSON serve different purposes, but both are components of digital research projects. If you are working with already structured content that contains complex data structures, files, or a range of formats, JSON is a useful way to describe it for transmission. JSON has a large user base and is likely to be supported as a format for the foreseeable future. These are also important considerations. **[Exercise 4.8: Contrast JSON and XML formats]**

Bonus content: discussion of Linked Open Data and the concept of interoperability

As the scope and scale of online resources grew in the last decades, the realization arose that making material “interoperable” posed certain challenges (Blaney 2020). The principle of interoperability is that resources described and stored in one networked environment ought to be able to be integrated and used simultaneously with materials in other environments. If the Museum of Modern Art in New York classified its assets with one set of terms and the São Paulo Museum of Modern Art and one in Seoul Korea used others, how could they be searched in aggregate? What if they used different digitization standards and file formats? The legacy data in collection management systems (the “back end”) is considerable. Though increasingly, there are standards for the cataloging of cultural objects, the problem of differences remains. One perceived solution to this problem was to create a structured information format that would be applied to all of these collections and allow them to be searched at the same time.

Many scholarly projects were built in siloed environments. This means that while considerable discipline-specific expertise went into these projects, they could not “talk” to each other or share information. Gradually, consortia began to form. One example is the international Pelagios Network created to integrate geographical work on ancient places, including cartography, archaeology, codicology, and other fields. A project called Europeana created a networked infrastructure in Europe. In Canada and Australia, networked infrastructure has been built with the goal of exposing collections to greater access. The documentation in these projects shows their shared recognition of the need for aggregation and sharing.

Linked Open Data (LoD) fosters data integration by creating a standard for libraries, museums, archives, and other cultural institutions that is legible to machines. The goal is a standard metadata vocabulary that can be universally adopted, including application to legacy information. For instance, one feature of LoD is an International Authority File—a single index of all named entities such that every individual person, or place, who might be referenced in any data source has a single unique ID called a Uniform Resource Identifier. The structure used in LoD allows for considerable analysis of relations (the data is structured in what are known as “triples” of subject/predicate/and object). Think of the challenges of names spelled many ways, in different languages, and using different transcription conventions—the difficulties of aligning these are daunting.

Bonus content: discussion of Resource Description Format

LoD uses a data standard known as RDF, Resource Description Format. Like many of the standards already mentioned, it is supervised by the W3C as part of the development of Semantic Web Technologies. These are approaches that were meant to make data machine-readable, and thus analyzable, across sites and repositories. Keep in mind, all of this grew up without precedents or standards, as multiple and varied institutions began to move their assets into online formats. Whether LoD will succeed depends on the extent to which it is adopted and whether alternatives arise that can be automated to do some of the same work of integrating information that was created in a wide range of formats and structures.

One example of a project that is built on LoD principles is LOUD. The acronym stands for Linked Open Usable Data and promotes usability. By contrast, LoD's publishing data is meant to make it easy to exchange information among systems. Usability implies other activities—for instance, is a file really usable if it can only be displayed at one size? What if a video has no fast-forward button? Or an audio file cannot be bookmarked? By now you will not be surprised to find that a common standard is JSON-LD, a JSON format for LoD. LOUD is meant to be more useful for consumers, including curators and individuals working with art objects. The LinkedArt site contains useful documentation, including information on projects and consortia (PHAROS consortium of Photo Archives, Linked Conservation Data, American Art Collaborative, etc.) and a long list of prestigious institutions with broad international representation. The descriptive metadata on the site (accessed under the Model tab) gives a clear sense of the community from which it derives, and the specific domains in which it is useful. LOUD is related to International Image Interoperability Framework (IIIF), which will be taken up immediately in the next section. **[Exercise 4.9: LOUD, an example of LOD]**

Bonus content: discussion of the IIIF

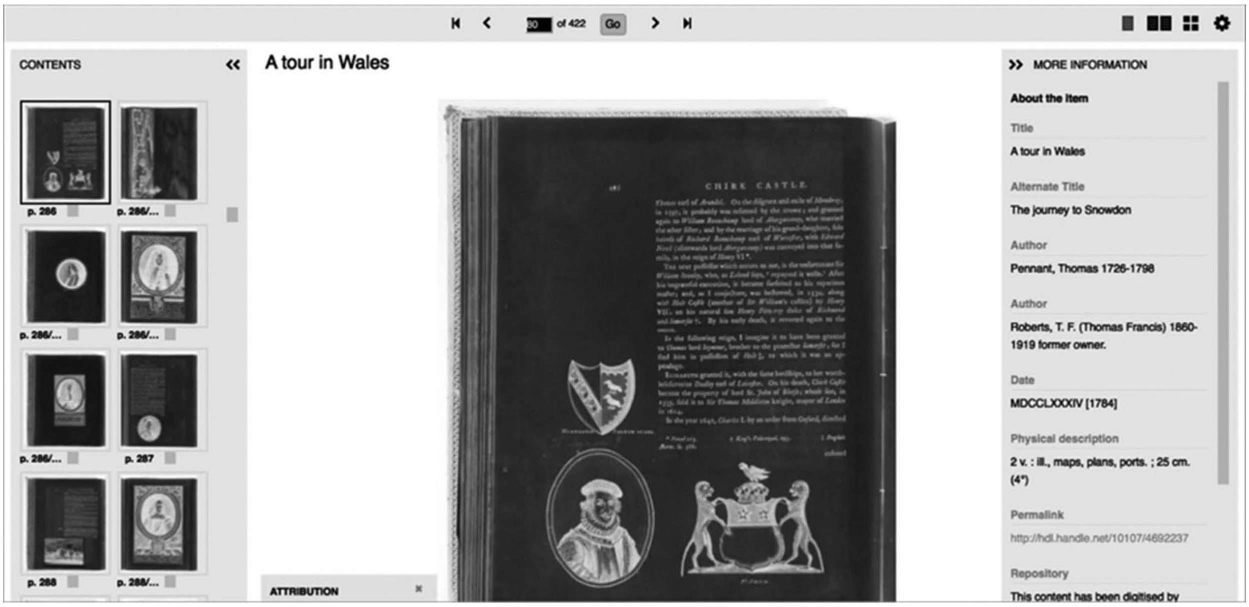
The need to standardize image formats for exchange and transmission lead to the development of IIIF (https://iiif.io/). This standard is widely accepted in museums, galleries, and repositories containing image and audio-visual files. The problem this was addressed to design was that if, for instance, a museum digitized its materials in a format that was not widely used, the files would not be able to open in another platform or environment. This would be equivalent to trying to open an MP3 file in Microsoft Word. Since the sharing of cultural heritage materials is one of the advantages of web-based repositories, information professionals realized that agreeing on a standard would save time, labor, and other resources. As the IIIF community states on the website, the goal was to break down silos.

Each of the words from which the acronym IIIF is taken is suggestive: International shows that this is a project that has many global partners. A glance at the map of participating institutions confirms this, and though the majority are in the northern hemisphere and in Europe, Canada, the United Kingdom, and the United States, participants from the global south are also present. Image is perhaps misleading, because though the project began with focus on image management, it has expanded to audio and video formats as well using the same principles. Interoperability points to the commitment to make resources usable across institutions and platforms in ways that are reliable (links will remain, references are stable, and the assets have standard formats and metadata). Framework, the final term indicates that the project is designed to provide standards for storage, access, delivery, and use through a number of APIs (you remember these, Applied Programming Interfaces).

The main APIs in this framework are for Image (viewing specifications) and Presentation (metadata and sequencing of images). More on both of these see below. But the concept of the framework is important because it is designed to anticipate use and integrate standards and services into a single environment. For example, it imagines that a scholar might want to take images from more than one institution and work with them in a common environment, adding custom notes and organization. Or, that if someone was looking at a video, you might want to be able to fast forward or put in bookmarks for later reference. These two examples describe an integration of digital resources (or assets: images and videos) and services for their use (annotation or search).

IIIF was initiated to "give scholars an unprecedented level of uniform and rich access to image-based resources hosted around the world."* For instance, the Image API allows an international user community to access images and do things like edit them, resize, or cut and paste. While this sounds simple enough, designing a standard that makes this possible requires considerable cooperation. In addition, IIIF is open-source and does not require vendor subscriptions or proprietary systems to be implemented. The system was designed so that access and other actions (linking, updating, versioning) could be managed by host institutions in accord with their own policies for log-in and permissions.

IIIF compliant repositories make images available to be used in certain standard ways: as thumbnails, for zooming, or as regions for detailed viewing. The derivatives—versions of the image in various sizes for onscreen viewing, detailed study, print publication, and so on—are already built into the collection as sized image files. The format allows the images to be scaled, rotated, cropped, and manipulated in other ways while keeping them as part of web linked data. Annotation layers can be stored—and shared—on images, thus making detailed scholarship and analysis available. By standardizing the ways that objects are represented in repositories, as well as the formats in which they are stored and accessed, the IIIF makes image access easier and more reliable. It also means that institutional repositories can make their resources available once for reuse by multiple individuals or other institutions, rather than having to answer requests for image download and use on a repeated basis. The savings to institutional repositories are considerable, and benefits to those accessing the images are that they know what to expect. Reliability and consistency add value to the assets but preparing these digital materials requires an investment of resources. IIIF assets can be accessed without a high-speed internet connection because of their reliable standards for availability.



A tour in Wales

MORE INFORMATION

About the Item

Title
A tour in Wales

Alternate Title
The journey to Snowdon

Author
Pennant, Thomas 1726-1798

Author
Roberts, T. F. (Thomas Francis) 1860-1919 former owner.

Date
MDCCLXXXIV [1784]

Physical description
2 v. : ill., maps, plans, ports. ; 25 cm. (4°)

Permalink
http://hdl.handle.net/10107/4692237

Repository
This content has been digitised by

Figure 4.1 IIIF information in Tour of Wales by Thomas Pennant, digitized by the National Library of Wales. Note the sidebar of information. (By permission of Llyfrgell Genedlaethol Cymru/The National Library of Wales).

To be used effectively, the resources in IIIF have to be accessed in a viewer installed in a local environment such as a desktop or web page. The viewers are programs (not people), and many off-the-shelf products exist. Some are open-source and others are proprietary. Leaflet-IIIF, for instance, is made to work with images. IIIF has its own vocabulary to describe the way its framework is organized into what are termed manifests. Understanding these technical details is essential for those building IIIF repositories, but for those using them, tutorials are more immediately valuable (Figure 4.1).

IIIF offers considerable support for its framework, including tutorials and demonstrations of its capacities. Above is an image taken from one such tutorial, showing the "sidebar" column with metadata next to the image of the resource. The difference between accessing images through IIIF and finding them in a random search engine result is that the institutional source and information about the dimensions, materials, authorship, and other significant data are included (Look at the sidebar in Tour of Wales above). Search engines offer image results that are often far from the original materials, remediated multiple times, and do not have information for tracking owners for permissions and use. Inaccuracies in the image resource itself, such as cropping of important information (signatures, frames, dates, or other contextual elements) also occur. Professionally managed assets presented in a framework are the gold standard for digital scholarship. **[Exercise 4.10: Tutorial Introducing IIIF and Viewers]**

Note that the information about IIIF can be found on GitHub (https://github.com/IIIF), a useful open-source repository for programming code (it was recently bought by Windows, but still operates), about which more will be discussed later. (GitLab (https://about.gitlab.com/) is an alternative.) GitHub (https://github.com/) provides a service by hosting projects and code and supporting version control. Learning the new terminology, acronyms, and technical back-end pieces can feel overwhelming, but finding a viewer with which to work is an easier way into the experience. Scholars do not have to be developers but gleaning the basics of a framework's design from tutorials allows for more effective use of these tools.

**For those with an interest in rule-based systems and the highest level of document type specification, understanding the syntax of SGML boils down to a way to "declare" the contents of files and determining the ways they can be used.*

Bonus content: discussion of CIDOC-CRM

Every domain of knowledge and different discipline has its specific needs and demands when it comes to metadata and standards. Over several decades, the cultural heritage field has developed a "theoretical and practical tool for information integration" in its field. This is known as CIDOC-CRM. CRM stands for Conceptual Reference Model, a phrase that does not bode well for easy uptake of what it is (sorry). However, translated into the language of digital scholarship, this means that it is a way to describe cultural heritage materials using a formal, structured, and standard language. As with many of these standards, it was developed to be useful for information systems, rather than human readers. A look at the use cases that are part of the documentation on its site is instructive. They are institutionally based projects, mostly with library or museum technical support, and contain substantial intellectual contributions by a team of scholars as well as technical experts. With CIDOC-CRM, we are squarely in the realm of professionals whose expertise has helped to build standards for infrastructure. Because it is such a widely used and powerful model and is an ISO (International

Standards Organization) standard, knowing about the role it plays in museums and other repositories of cultural heritage materials is useful.

What you need to understand about CIDOC-CRM at this point, given the discussions previously, is that standards have been created for data in networked cultural heritage environments and that any project that is being conceived at this point in time in digital scholarship benefits from being built within these standards and frameworks.

Exercise 4.4: XML

XML schema exists in many fields and disciplines. Look at this list (https://en.wikipedia.org/wiki/List_of_types_of_XML_schemas), pick one, and begin to see how it identifies elements of documents for this field. While you are looking at the list, compare it to the Library of Congress (https://en.wikipedia.org/wiki/Library_of_Congress_Classification#Design_and_organization) and to the Dewey Decimal System (https://en.wikipedia.org/wiki/Dewey_Decimal_Classification#Design). Why do you think the XML schema list is so different? What does the list tell you about which fields are most focused on standards for publishing information? What fields are missing? Dance?

How-to example

XML schemas are concerned with the structure of the document itself—breaking it down into its constituent parts (title, abstract, references, etc.) for standardization and machine readability. In contrast, the LOC and DDS classify resources based on the content or subject of the documents and how they relate to broader knowledge categories. XML schemas don't typically classify the document's content into subjects; instead, they describe its format and structure. XML schemas tend to be field-specific because different disciplines have unique needs for how they structure their documents. For example, the structure of a scholarly article in medicine (*JATS*) is different from how legal texts are structured (e.g., Akoma Ntoso, an XML schema for legislative documents). LOC and DDS aim to be universal systems, applied across all disciplines but are not concerned with how documents are structured internally. The extensive use of XML schemas in certain fields, such as academia (JATS), publishing (EPUB), archives (EAD), and business (XBRL—eXtensible Business Reporting Language), highlights that these fields are especially focused on the standardization of information for digital preservation, dissemination, and interoperability. Some fields, particularly in the arts and humanities, may be underrepresented in XML schema development. While there are schemas for film, music, and theater, there's no widely adopted XML schema specifically for dance notation or choreography. This could reflect the challenge of encoding the highly embodied and performative nature of dance, which does not lend itself as easily to textual or data-centric standards. Dance information is often stored as video recordings, images, or in specialized notation systems (e.g., Labanotation), which may not align well with the existing textual XML schema paradigms.

Exercise 4.5: TEI

TEI is a highly specialized XML standard for marking up literary and textual documents. It is divided by genre and its tags must be used in particular orders, hierarchies, and are highly defined. Looking at the TEI text body standards (https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html), compare the elements for seven. Performance Scripts, and eleven. Manuscript Descriptions. What are some ways the schema shifts to better suit the content?

How-to example

Performance Scripts focus on the structural components of a performance text, such as acts, scenes, and speeches, which are essential for representing the flow and organization of a performance. Manuscript Descriptions, on the other hand, focus on the physical and historical aspects of a manuscript, providing detailed metadata about its physical state, origin, and content. In addition, Performance Scripts are designed to capture the dynamic nature of a performance, with elements like speaker, stage directions, and grouped speeches. Manuscript Descriptions are more static, focusing on the static attributes of a manuscript and detailing information under headings like physical description, provenance, and content summary (e.g., <physicalDescription>, <provenance>, <summary>). These shifts in the schema reflect the different needs and purposes of encoding performance texts versus manuscript descriptions. Performance Scripts need to capture the dynamic and structured nature of performances, while Manuscript Descriptions need to provide detailed and comprehensive metadata about physical objects.

Many pioneering digital humanities projects were the work of literary scholars for whom TEI was a useful tool of analysis. TEI could be applied to a local project hosted on an institutional server but could also be integrated into larger projects. A quick look at the TEI Guidelines gives a sense of how closely it conforms to the study of bibliographical objects—books and manuscripts in particular. The major divisions of the TEI are Front Matter, Back Matter, and Body. The guidelines are specific to literary and humanistic texts, so there are standard tags for verse that include metrical analysis and rhyme, for instance, or metrical analysis within stanzas. The tag sets can become very elaborate to indicate every genre and format of text and the many components specific to each one (just think of all of the different ways the language of a play can be described from its stage directions, set descriptions, speakers, narrator, chorus and so on.). TEI is useful for formalizing the interpretation and study of texts, but it forces the analysis of a literary work into a formal and explicit framework that can be problematic on intellectual and ethical grounds. **[See Exercise 4.6: Orlando Project and 4.7 TEI title page elements]**

Exercise 4.6: Orlando Project

The Orlando Project (http://www.artsrn.ualberta.ca/orlando/) advances feminist literary scholarship through collaborative and multidisciplinary efforts, producing a richly searchable textbase that encompasses extensive interpretive information on women's writing and culture. Look at the documentation for the XML scheme and think about the customization of its tags and features.

How-to example

The Orlando Project, which advances feminist literary scholarship, applies a feminist methodology in its XML schema by prioritizing inclusivity, intersectionality, and the representation of marginalized voices within literature and culture. Focus your close looking on the project's glossary of tags (https://orlando.cambridge.org/about/glossary). Based on the project's documentation, here are a few key ways this feminist methodology is applied through the customization of its tags and features:

1. Customization for inclusivity and representation:
    - The Orlando Project's XML schema is designed to encode not just the texts themselves but also rich contextual and interpretive information about women's writing and culture. This means that the schema captures gendered, racialized, and class-based dimensions of authorship and literary production.
    - Tags are customized to represent women writers and their experiences in a way that highlights their often overlooked contributions to literary history. For example, the project's database tracks how women's writing has intersected with other social movements and historical events, capturing this in ways that more traditional schemas may overlook.
2. Intersectional categories:
    - A feminist methodology is apparent in how intersectional categories are created and encoded. Women's identities are not just reduced to gender; instead, the project captures multiple intersecting factors such as race, class, nationality, and sexuality. This approach prevents reductive categorizations and highlights how multiple social identities shape the experiences and contributions of women writers.
    - The XML tags reflect a nuanced understanding of women's lives in historical contexts, allowing scholars to explore how categories like "race," "gender," or "political activism" interrelate in the textual database.
3. Focus on interpretation and subjectivity:
    - Feminist methodology embraces the subjectivity of interpretation and acknowledges that data is never neutral. The Orlando Project explicitly encodes interpretive data in its XML schema, allowing researchers to explore not only the text but the meanings and contexts behind it.
    - The tags allow for the encoding of thematic interpretations, including the presence of gender dynamics, power relations, and cultural contexts in the texts. This brings feminist interpretive lenses into the digital structure of the project, fostering critical analysis from multiple perspectives.
4. Collaborative, multidisciplinary effort:
    - Feminist scholarship often emphasizes collaboration and collective knowledge production, which is reflected in the development of the Orlando Project. The collaborative nature of the project across scholars from diverse disciplines ensures that the schema evolves to reflect a broad range of feminist concerns and methodologies.
    - This collaborative practice shapes the customization of the XML schema to ensure it remains responsive to the needs of scholars from different fields, such as history, cultural studies, and gender studies.
5. Highlighting marginalized and under-represented voices:
    - A key aspect of feminist methodology is the focus on recovering marginalized and underrepresented voices. The Orlando Project's schema includes tags and features that allow the identification and tracking of writers from marginalized communities, such as women of color, working-class women, and women writing outside of dominant cultural centers.
    - The project's encoding strategies make it possible to search for patterns of exclusion and visibility, revealing how certain voices were historically marginalized or erased in literary scholarship.

6. Temporality and social change:

    - Feminist methodology often considers temporality and how gender roles evolve over time. The Orlando Project's XML schema reflects this by encoding historical and cultural shifts in gender roles and women's participation in literature. It tags the historical periods, movements, and social changes that influenced the writing of women, emphasizing how feminist movements intersect with literary production.

Exercise 4.7: TEI title page elements

Look at the elements for TitlePage in TEI (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DS.html#DSTITL) and think about how using these terms limits the kinds of materials that can be marked up. Are there publications that would not follow this form? Elements that might need to appear for it to work? What about artists' books? Or zines?

How-to example

The TEI TitlePage elements are designed to capture the traditional components of a title page, such as the title, author, publication information, and edition details. This structure is well-suited for conventional books and academic publications but might not accommodate the unique formats of artists' books and zines. Artists' books often have unconventional formats, mixed media, and non-linear narratives that might not fit neatly into the standard TitlePage structure. Elements like <graphic> or <figure> might be more appropriate to capture visual elements. Zines are typically self-published, small-circulation works that might not have formal publication details. Elements like <byline> or <imprint> might be less relevant, and additional elements to capture the DIY nature and unique content might be needed. To better accommodate these types of publications, the TEI schema could include more flexible and descriptive elements, such as:

- <artwork> for visual art components,
- <collaboration> to capture multiple contributors,
- <format> to describe the physical characteristics of the publication, and
- <context> to provide background information on the creation and distribution of the work.

Exercise 4.8: Contrast JSON and XML formats

Examine the contrasts in the JSON and XML formats of the same information. The main difference is the syntax and the way that the values are specified. Here are two excellent examples: www.w3schools.com/js/js_json_xml.asp and https://json.org/example.html. Describe a scenario when you might use JSON instead of XML, and a second scenario when you might opt for XML instead of JSON.

How-to example

JSON SCENARIO

Imagine a digital humanities project involving a large-scale analysis of social media interactions to study contemporary cultural trends. JSON would be ideal here because it's lightweight, easy to parse, and works seamlessly with JavaScript and other programming languages commonly used in web development. For example, you could use JSON to gather, store, and analyze tweets related to a specific event, enabling quick data manipulation and visualization for your research. This would be particularly useful for creating a web-based, publicly accessible data dashboard presentation.

XML SCENARIO

Imagine a digital edition of a historical manuscript with intricate annotations, footnotes, and cross-references. XML would be more appropriate because it allows for highly detailed and structured markup, preserving the complex hierarchy and relationships within the document. This is crucial for accurately representing the manuscript's content and providing robust metadata for scholarly analysis.

These choices reflect the need for simplicity and efficiency in social media research versus the requirement for detailed, structured documentation in historical manuscript studies. JSON excels when you need speed and interactivity, making it ideal for projects in the digital humanities that focus on real-time visualizations or data manipulation on the web. XML is a better fit for text-based projects in the humanities where preserving the structure of the data and attaching rich metadata are essential, such as for archiving historical documents or encoding complex manuscripts. In these contexts, the choice between JSON and XML depends on whether the research requires interactive data exploration or long-term structured archiving and textual representation.

Exercise 4.9: LOUD, an example of LoD

One example of a project that is built on LoD principles is LOUD (https://linked.art/loud/). The acronym stands for Linked Open Usable Data and promotes usability. LOUD is meant to be more useful for curators and individuals working with art objects. The LinkedArt site contains useful documentation, including information on projects and consortia (PHAROS consortium of Photo Archives, Linked Conservation Data, American Art Collaborative, etc.) and a long list of prestigious institutions with broad international representation. Can you imagine a project in which you would make use of LOUD? What would be the benefits and liabilities?

How-to example

Look at the parts of the LOUD model (https://linked.art/model/) and create a workflow for creating metadata for a photograph collection you have inherited from your grandparents from their travels around the world in 1935. Could you use the model effectively? How about if you want to make metadata for your parents' collection of concert posters from their wild days in the 1960s and early 1970s when they were going to rock music performances? What are the differences between these two projects and how does each have challenges for metadata standards?

Exercise 4.10: Tutorial introducing IIIF and viewers

Look at the list of viewers available for IIIF and see what you learn reading their descriptions. Build a vocabulary for understanding open-source and proprietary formats. Note that some are built in JavaScript and HTML5, others make use of JSON formatted data (the suffix.js is frequently present). Learn to evaluate feature sets for these applications in relation to your project needs: comparison, annotation, sequencing, high-resolution image displays, and so on.

Read through this workshop (http://ronallo.com/iiif-workshop-new/image-api.html) to understand IIIF. Ask yourself what you do and do not understand and whether you think you can use this resource within your research.

How-to example

Explore these IIIF workshop videos:

- Teaching and Research with Digital Collections—Part 1 (2023) (https://youtu.be/SuPGRmryCqI?si=npytm97c1xsRKkyR)
- Teaching and Research with Digital Collections—Part 2 (2023) (https://youtu.be/_xOnkUfSpM8?si=81jPDTTFqfSF6BxP)
- Teaching and Research with Digital Collections—Part 3 (2023)(https://youtu.be/N1ZuugRm04E?si=8M8XoPnViWqlHCeC)

Recommended readings

"JSON vs XML." 2024. "W3schools." www.w3schools.com/js/js_json_xml.asp.

Schwartz, Michelle, and Constance Crompton. 2018. "Remaking History: Lesbian Feminist Historical Methods in the Digital Humanities." In *Bodies of Information: Intersectional Feminism and the Digital Humanities*, edited by Elizabeth Losh and Jacqueline Wernimont, 131–56. Minneapolis: University of Minnesota Press. https://rshare.library.torontomu.ca/articles/chapter/Remaking_History_Lesbian_Feminist_Historical_Methods_in_the_Digital_Humanities/24135024?file=42342756.

Bibliography

Blaney, Jonathan. 2020. "Introduction to the Principles of Linked Open Data." *The Programming Historian*. https://programminghistorian.org/en/lessons/intro-to-linked-data.

Breakthrough Staff. 2017. "How Genetically Related Are We to Bananas?" *Breakthroughs*. https://www.pfizer.com/news/articles/how_genetically_related_are_we_to_bananas.

IIIF Gain Richer Access to the World's Image and Audio/Visual Files. n.d. https://iiif.io/.

Prescod, Paul, Ben Feuer, Andrii Hladyki, Sean Paulk, and Arjun Prasad. 2023. Balisage Paper: Auto-Markup BenchMark: Towards an Industry-Standard Benchmark for Evaluation Automatic Document Markup. https://www.balisage.net/Proceedings/vol28/html/Prescod01/BalisageVol28-Prescod01.html.

Wallack, Jessica, and Ramesh Srinivasan. 2009. "The Local and the Global: Reconciling Mismatched Ontologies in Development Information Systems." In *HICSS Hawaii International Conference on Systems Science*s, 1–10.

Zhang Allison B., and Don Gourley. 2009. "Creating Metadata: Metadata Cross-Walk." In *Creating Digital Collections*. Oxford: Chandos Publishing. www.sciencedirect.com/topics/computer-science/crosswalk.

Resources

- A compilation of sample standard forms used in the 19th Century (https://mrcc.purdue.edu/FORTS/samples)
- Dublin Core (https://dublincore.org/)
- JSON vs. XML (W3Schools) (http://www.w3schools.com/js/js_json_xml.asp)
- KML Tutorial (https://developers.google.com/kml/documentation/kml_tut)
- Library of Congress Classification (http://www.loc.gov/catdir/cpso/lcco/)
- MARC Records (http://www.loc.gov/marc/bibliographic/)
- Mark-Up Languages (Wikipedia) (https://en.wikipedia.org/wiki/List_of_XML_markup_languages)
- Metadata Crosswalks (https://guides.lib.utexas.edu/metadata-basics/crosswalks)
- Metadata Standards, European Union (http://www.dcc.ac.uk/guidance/standards/metadata/list)
- SGML Overview (http://www.w3.org/MarkUp/SGML/)
- SGML Tutorial (http://www.w3.org/TR/WD-html40-970708/intro/sgmltut.html)
- TEI (https://tei-c.org/)
- TEI Guidelines (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html)
- TEI’s Intro to XML (https://tei-c.org/release/doc/tei-p5-doc/en/html/SG.html)
- Weather History for Los Angeles, CA (Almanac) (http://www.almanac.com/weather/history/CA/Los%20Angeles/2018-07-02)