Beyond TEI:
Metadata for Digital Humanities

Carolyn Hansen & Sean Crowe

DHSI 2017

Course Description

High-quality metadata is essential for the description, discovery, and preservation of DH projects. While TEI is the most used metadata standard in DH, there is so much more to learn and explore! This course will introduce metadata schemas and standards such as Dublin Core, VRA, controlled vocabularies, and linked data and RDF. We will also discuss ontologies, ethics of standardization, data management, and digital preservation. Hands-on work with participants' own datasets will be given to practice metadata/data cleaning with OpenRefine, creating custom schemas, and linking to external authorities. Students need no prior experience with metadata or programming.

How to prepare for this course:

There are a few things that you can do to prepare for this course.

First, come prepared by installing OpenRefine. You can find instructions for installing OpenRefine for your operating system at http://openrefine.org/download.html. Download and install version "OpenRefine 2.7-rc2 Release Candidate 2." Help will be provided on the first day.

Second, pick a dataset in your field that you would like to work with, or better yet, several. We will go over how to create a metadata schema, choose controlled vocabularies, clean data, and create linked data and RDF. We will provide a dataset for the in-class exercises, but you may wish to work with your own data.

Third, try to get through the readings. Some of them are technical, but don't be discouraged! Even if you don't understand everything you're reading, even being familiar with the vocabulary will help you to get more out of this course. Also, visit some of the sites for various metadata standards (indicated in the "For Reference" sections).

Schedule:

Day 1 | Introduction to Metadata and OpenRefine

Morning/Early Afternoon –
Introduction to metadata. Participants will share their project goals, what kinds of documents or data they are working with, and what schemas they are considering. We will also give an overview of the course and discuss (broadly!) types of metadata, standards and schemas, ontologies, controlled vocabularies, ethics, and editorial policy.

Readings:

- "Setting the Stage." Ann J. Gilliland. In *Introduction to Metadata.* Edited by Murtha Baca. Getty. 2016.
- "OpenRefine." Wikipedia. Last modified March 24, 2017. https://en.wikipedia.org/wiki/OpenRefine
- Short videos: "1. Explore Data"; "2. Clean and Transform Data"; and "3. Reconcile and Match Data." Available at: http://openrefine.org/ (Note: Videos use the application's old name "GoogleRefine")

For Reference:

- OpenRefine Wiki: https://github.com/OpenRefine/OpenRefine/wiki
- OpenRefine Manual: https://www.amazon.com/Using-OpenRefine/dp/1783289082/ (Note: Quite a bit of this book is out of date, which is why we won't be using it in class)

Afternoon – Hands-on time. An introduction to the OpenRefine tool for cleaning metadata will be provided and participants will be given project-based exercises to learn OpenRefine's basic functionality when working with humanities data.

Day 2 | Schemas, Controlled Vocabularies, and Ontologies

Morning – Choosing metadata schemas. We will discuss Dublin Core, MODS, and VRA in addition to interoperability and machine-readability. Exercises will be provided to give participants experience working with different schemas as well as to see how their data changes depending on the schema used.

Readings:

- "Nine Questions to Guide You in Choosing a Metadata Schema." Marie R. Kennedy. *Journal of Digital Information*. Vol 9., No. 1 (2008). https://journals.tdl.org/jodi/index.php/jodi/article/view/226/205
- "A Gentle Introduction to XML." TEI: Text Encoding Initiative. Version 3.1.0. December 2016. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SG.html#SG11

For Reference:

- Dublin Core: http://dublincore.org/
- MODS: http://www.loc.gov/standards/mods/
- VRA Core: https://www.loc.gov/standards/vracore/

Afternoon – Introduction to ontologies and controlled vocabularies. We will discuss LC and LAC authorities, VIAF, ISNI, the Getty, as well as the curatorial and political implications of "naming." Exercises will be provided to give participants experience working with different controlled vocabularies as well as creating their own ontologies.

Readings:

- "The Power to Name: Representation in Library Catalogs." Hope A. Olson. *Signs*. Vol. 26, No. 3 (Spring 2001).

For Reference:

- LC Authorities: http://authorities.loc.gov/
- LAC Authorities: http://www.bac-lac.gc.ca/eng/services/canadian-subject-headings/Pages/about-csh.aspx
- VIAF: https://viaf.org/
- ISNI: http://www.isni.org/
- Getty: http://www.getty.edu/research/tools/vocabularies/

Day 3 | Linked Data and RDF

Morning/Early Afternoon – An introduction to linked data, the Semantic Web, RDF, SPARQL, APIs, and Linked Data Fragments will be provided.

Readings:

- "Chapter 1: Introduction" from *Linked Data: Evolving the Web into a Global Data Space* (1st edition). Tom Heath and Christian Bizer. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool. 2011. http://linkeddatabook.com/editions/1.0/
- "RDF 1.1 Primer. W3C Working Group. June 2014. https://www.w3.org/TR/rdf11-primer/
- Chapters 1-3 from *An Introduction to APIs*. Brian Cooksey. Zapier, Inc. 2014.

For Reference:

- The Semantic Web Made Easy: https://www.w3.org/RDF/Metalog/docs/sw-easy
- Linked Data Fragments: http://linkeddatafragments.org/

Afternoon – Hands-on time. Participants will learn how to create RDF and linked data using OpenRefine.

Day 4 | TEI, Reconciliation, and OpenRefine

Morning/Early Afternoon – TEI and OpenRefine. Participants will learn how to use OpenRefine for standardization and reconciling hierarchical TEI metadata with external authorities. Introduction of final project/course e-exhibit.

For Reference:

- TEI: http://www.tei-c.org/index.xml
- XML: https://www.w3schools.com/xml/xml_whatis.asp

Afternoon – Hands-on time with participants, including exercises using TEI and OpenRefine. A TEI dataset will be provided for the exercises, but participants are welcome to use their own data.

Day 5 | Data Management and Digital Preservation

Morning -- We will discuss best practices for creating metadata and file systems that can be migrated and preserved, including PREMIS, METS, and BagIt. Long-term hosted storage solutions such as institutional repositories and digital consortiums, will be examined. Additional hands-on time and work on final project/e-exhibit.

Readings:

- "A Digital Dark Ages? Challenges in the Preservation of Electronic Information." Terry Kuny. 63rd IFLA Council and General Conference. 1997.
- "Significant Properties of Digital Objects: Definitions, Applications, Implications." Margaret Hedstrom and Christopher A. Lee. Proceedings of the DLM-Forum 2002. https://ils.unc.edu/callee/sigprops_dlm2002.pdf

For Reference:

- PREMIS: http://www.loc.gov/standards/premis/
- METS: http://www.loc.gov/standards/mets/
- BagIT: https://en.wikipedia.org/wiki/BagIt

Early Afternoon – Set up for final project and e-exhibit showcase/lunch.

Instructor Contact Information

Carolyn | hansen.caro@gmail.com | @meta_caro
Sean | sean.crowe@uc.edu | @crowesn