Viable Vocabularies

An Introduction to Controlling, Cleaning, and Linking Your Data

Instructors

- Metadata and Discovery Unit
  - Mike Bolam - Head, Metadata & Discovery Unit
  - Gabi Gulya - Metadata Librarian
  - Heajin Kim - Authority Control Specialist
  - Staci Ross - Visiting Cataloging/ Metadata Librarian

Metadata & Discovery Unit

The Metadata and Discovery Unit provides metadata consultation, development, and production support for students, faculty, and staff. As a center for expertise in descriptive, technical, and administrative metadata, we work with researchers at the University of Pittsburgh to support their metadata needs.

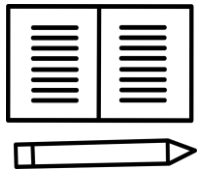
If you are starting a new project, want to improve the quality of existing metadata, or want to learn more about what metadata can do for you, specialists in the Metadata and Discovery Unit are available for consultations to help you get started. Training on standards and best practices, using controlled vocabularies, and data cleaning tools are also available through scheduled consultations. Contact us to make an appointment.

Introductions

Introduce yourself and tell us why you are here today.

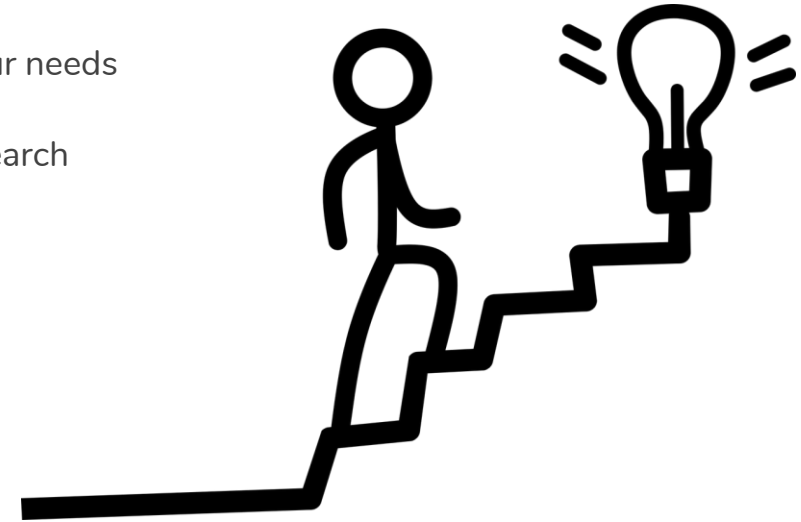
Learning Objectives

- Participants will be able to define controlled vocabulary and identify its purpose in the research process
- Participants will be able to find appropriate vocabularies for their research needs.
- Participants will be able to select appropriate terms from a controlled vocabulary.
- Participants will be able to perform the basic tasks of cleaning and linking data using OpenRefine.



Agenda

- Introduction to controlled vocabularies
- Finding the best vocabularies to meet your needs
- Using controlled vocabularies in your research
- Introduction to linked data
- Using OpenRefine to clean and link data



Workshop Files

- Workshop Slides: http://bit.ly/vv_slides
- Workbook: http://bit.ly/vv_workbook
- OpenRefine: http://www.openrefine.org
- Workshop Data Set: http://bit.ly/vv_data
- Named Entity Recognition Extension: http://freeyourmetadata.org/named-entity-extraction/

What are controlled vocabularies?

A **controlled vocabulary** is "an organized arrangement of words and phrases used to index content and/or to retrieve content through browsing and searching." (Patricia Harpring). To put it simply: "Find out what they call it, then select it." (Allen Smith) There are many types of controlled vocabularies, created and maintained by professionals in the subject area.

It is usually given to content when the content is cataloged/published by a cataloger, publisher or researcher. Therefore, controlled vocabulary is a tool for cataloger/publisher/researcher to organize information,and it is a key to easily retrieve accurate information for the user.

What about thesauri?

- A **thesaurus** represents all the concepts for a specific domain in a consistent manner and labels each concept with a preferred term
  - Controlled term, but includes synonyms and variant labels
  - Discipline-specific
  - Hierarchical relationships between concepts are expressed
    - Example: “apartments” from Getty’s Art & Architecture Thesaurus (AAT)

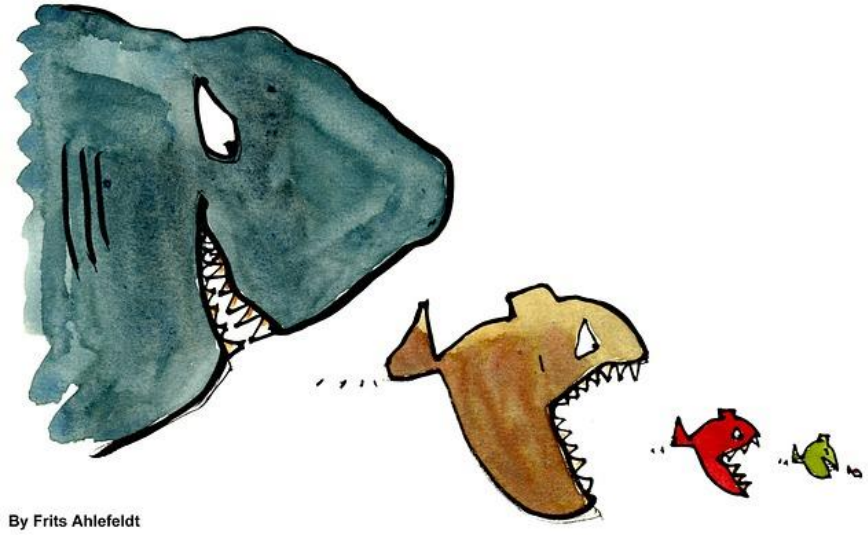
apartments

Hierarchical Position:
Objects Facet
.... Components (hierarchy name) (G)
..... components (objects parts) (G)
..... <components by specific context> (G)
..... building divisions (G)
..... rooms and spaces (G)
..... <rooms and spaces by location or context> (G)
..... interior spaces (spaces by location) (G)
..... combinations of rooms (G)
..... apartments (G)

- **Keywords** and **tags** may not come from controlled terms, so there is a higher chance of variance and inconsistency

What about taxonomies?

- A **taxonomy** is “the rules or conventions of order or arrangement”
- Allows for classification according to a pre-determined system.
    - Helps you organize your data into hierarchical relationships
    - Example : King Philip Came Over For Good Soup : Kingdom -> Phylum -> Class -> Order -> Family -> Genus -> Species.



Lambe, P. (2007). *Organising Knowledge: Taxonomies, Knowledge and Organisational Effectiveness*. Oxford, UK: Chandos.

Examples of controlled vocabulary and natural words (from LC headings)

AUTOMOBILES

SUVs

Sedans

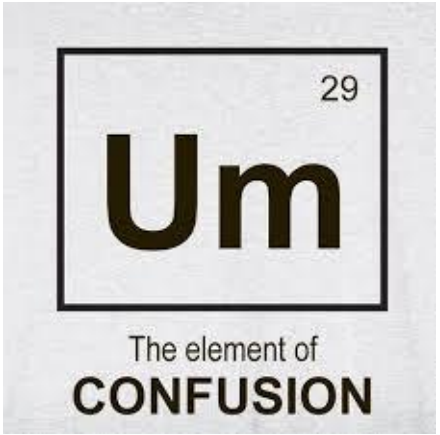
Sports cars

Controlled Vocabulary Activity

Why use controlled vocabularies in your research?

The need for controlled vocabulary is caused by the ambiguity of our language.

We are often confused by different places or persons having the same names. We sometimes use the same terms in unrelated disciplines. And there are variant spellings for the same word.



Examples of controlled vocabulary and natural words (from LC headings)

Railroad trains

Passenger trains

Freight trains

Examples of controlled vocabulary and natural words (from LC headings)

Aurelius Augustine,

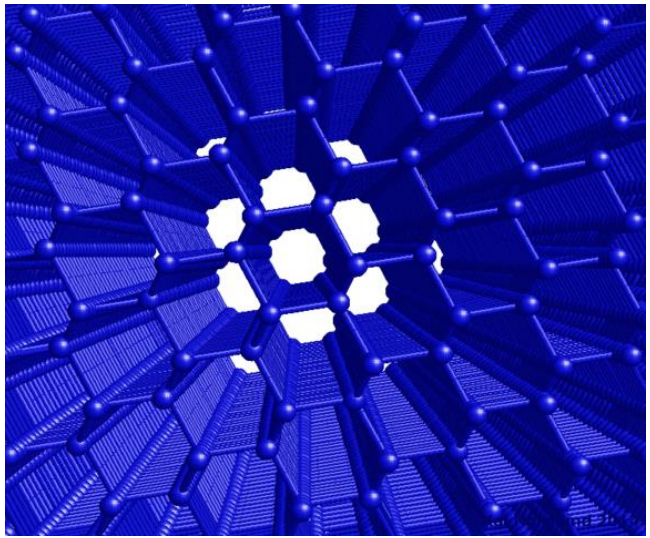
St. Augustine,

St. Augustine of Hippo,

St. Augustine Bishop of Hippo

Augustine, of Hippo, Saint, 354-430

Channeling : in Physics vs. in Spiritualism



Channeling (Physics)

Channeling (Spiritualism)

Michael Jackson, who is…..

a whisky expert, singer, and didjeridu player?

Jackson, Michael, 1942-2007

Jackson, Michael, 1958-2009

Jackson, Michael (Didjeridu player)

Advantages of using controlled vocabulary

Using controlled vocabulary can make your research more findable. When you use controlled vocabulary in your searches, you can get more meaningful results.

“gardening to attract butterflies” vs. “butterfly gardening”

Advantages of using controlled vocabulary

Controlled vocabulary can be used in any industry that collects and uses information: research, libraries, corporations, etc.

ex: The World Bank - WBG Thesaurus, WBG Business Taxonomy, WBG Topical Taxonomy [vocabulary.worldbank.org]

Advantages of using controlled vocabulary

Controlled vocabulary decreases possible misunderstandings in your research when a user searches for it. If you do research on a certain city, and there are several other cities having the same name, and you would be able to clarify which city it is in a glance.

Advantages of using controlled vocabulary

Pittsburgh (Pa.)

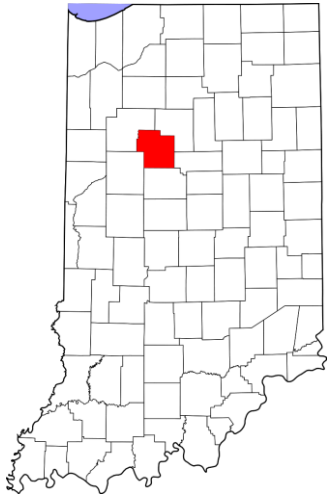
Oakland (Pittsburgh, Pa.)

South Side (Pittsburgh, Pa.)

Have you ever been to Pittsburgh?

Pittsburgh (Pa.)

Pittsburgh (Ont.)



Pittsburgh (Sullivan County, Ind.)

No more confusion!

To avoid misunderstandings we need to differentiate persons and places, confirm the usage of terms in a certain discipline and choose one spelling over the others.

Controlled vocabulary made by authorized institutions diminishes these uncertainties, which makes searches more efficient.

Vocabulary Finding Activity

Method to find the best vocabularies

If there is an existing controlled vocabulary list in your discipline.... (such as LC headings, Art & Architecture Thesaurus, MeSH etc.)

1) Go to one of the lists of controlled vocabularies for your discipline.

2) Think of possible term(s) of your research.

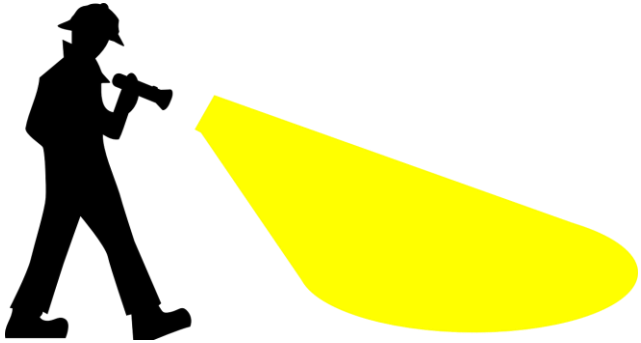
3) Search for those terms in the list.

4) When you get the search result, read the explanation carefully.

5) If a search result is different from your expectation, try synonyms/variants of the terms you had searched until you find a right one.

*When you need to choose a term, you can consider the purpose of your research, possible users, and relevance to existing research.

Finding Vocabularies



- RDA Metadata Standards Directory lists hundreds of standards, extensions, tools, and use cases. The directory can be browsed by discipline and subject area.
- Seeing Standards: A Visualization of the Metadata Universe provides a visualization of relationships between over 100 metadata standards used by cultural heritage organizations (libraries, museums, archives, galleries, etc.) The glossary provides links and brief descriptions for each of the standards represented.
- Digital Curation Centre's Disciplinary Metadata links to information about these disciplinary metadata standards, including profiles, tools to implement the standards, and use cases of data repositories currently implementing them.
- Metadata & Discovery @ Pitt LibGuide

Selected controlled vocabularies by discipline

General purpose

- Names
  - Library of Congress Name Authority File (LCNAF)
  - Virtual International Authority File (VIAF)
- Concepts, Objects, Topics, etc.
  - Library of Congress Subject Headings (LCSH)

Selected controlled vocabularies by discipline

Art & Humanities

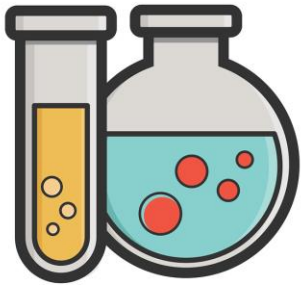
- Art & Architecture Thesaurus
- Rare books and special collections vocabulary
    - Union List of Artist Names (ULAN) - bibliographic info about artists and architects
    - Getty Thesaurus of Geographic Names (TGN) - names of places important for the study of art and architecture.
    - Library of Congress Thesaurus for Graphic Materials (LCTGM)
    - UNESCO Thesaurus

Selected controlled vocabularies by discipline

Transportation Information

- Transportation Research Thesaurus (TRT) developed by the National Transportation Library

Sciences

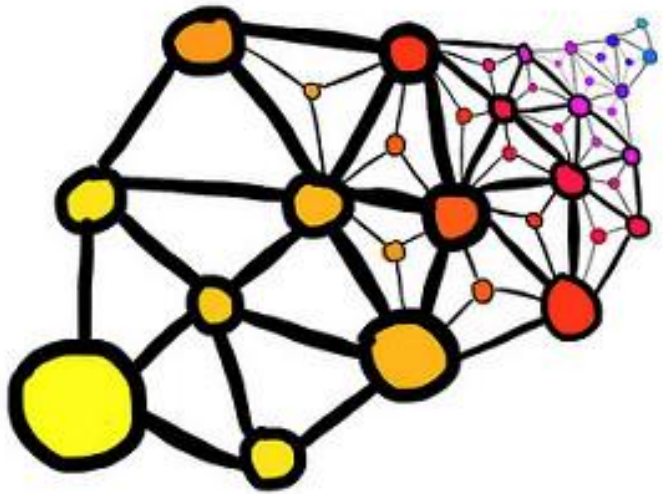


- International Classification of Disease (ICD)
- Medical Subject Headings (MeSH)
- National Agriculture Library Thesaurus (NALT)
- Unified Astronomy Thesaurus (UAT) from the American Astronomical Society

Linked Data:
An Introduction

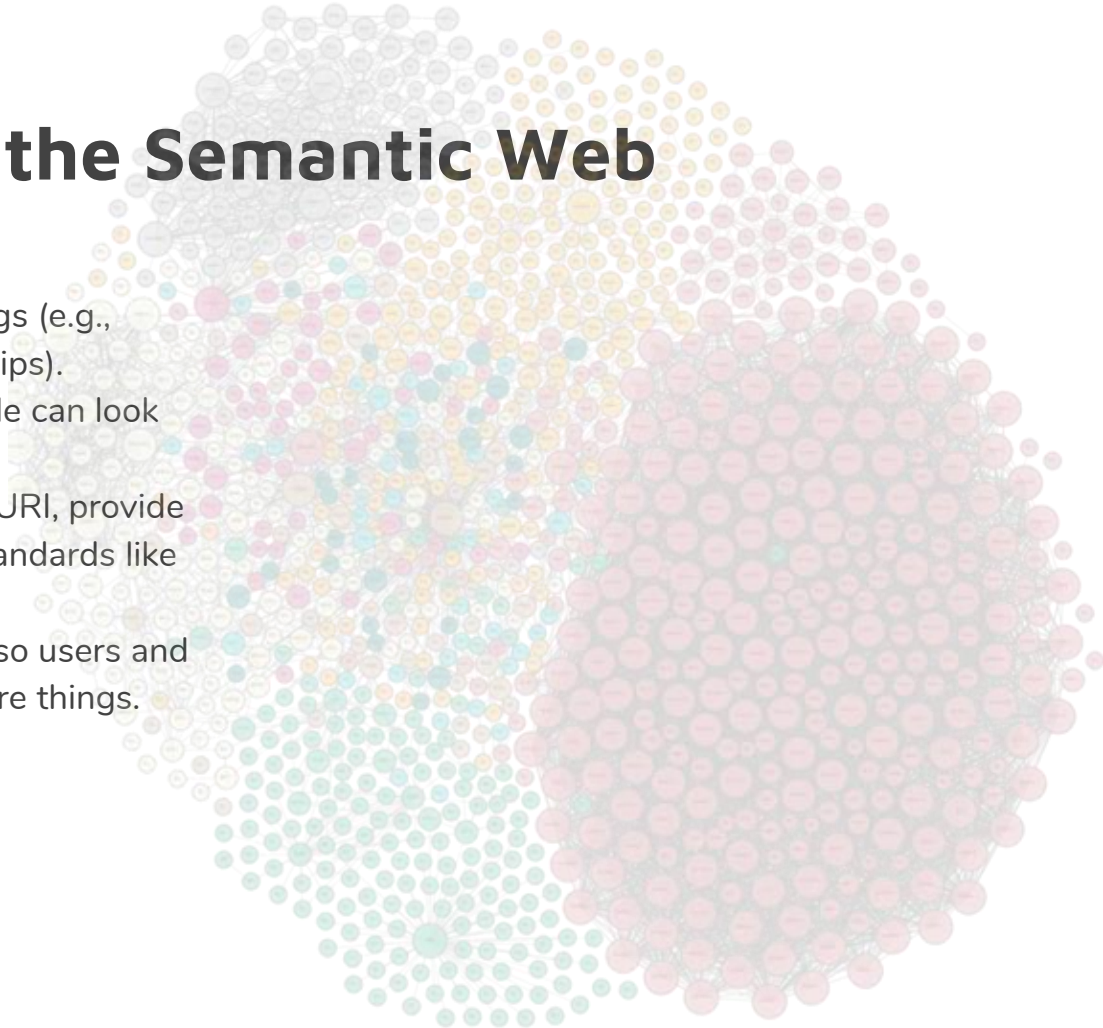
What is Linked Data?

- **Linked data** is the framework behind the **Semantic Web**, which uses metadata to infer relationships between entities: people, concepts, and resources
- In the Linked Data environment, controlled vocabularies include **URIs** (Uniform Resource Identifiers) to help humans and computers identify and reconcile names, concepts, and objects

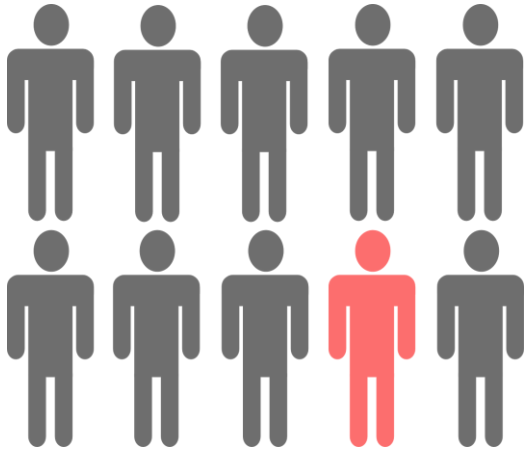


Four Rules for the Semantic Web

1. Use URIs as names for things (e.g., names, concepts, relationships).
2. Use web URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using standards like controlled vocabularies.
4. Include links to other URIs, so users and computers can discover more things.



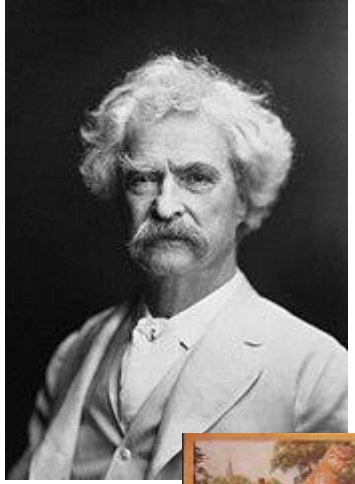
So... what is a URI?

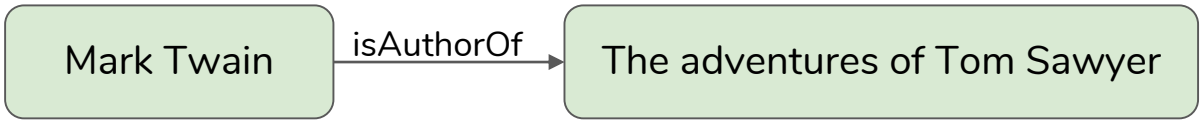


- A **URI**, or Uniform Resource Identifier, is a unique, controlled term used to identify something
- Types of URIs:
  - **URN** (Uniform Resource Name): identifies an entity by name in a particular namespace
    - Example: Twain, Mark, 1835-1910 (VIAF)
    - Example: Mark Twain (Wikidata)
  - **URL** (Uniform Resource Locator): identifies an entity by a web address
    - Example: https://viaf.org/viaf/50566653

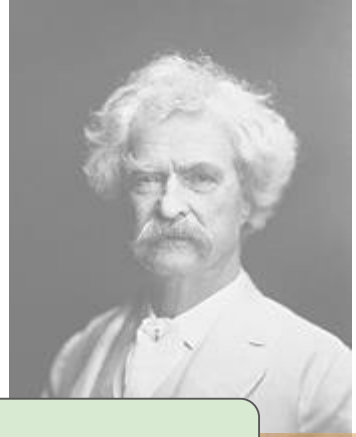
What about triples?

- Relationships between entities are structured as three-part subject-predicate-object statements, called **triples**
  - Example: Mark Twain -- is author of -- *The Adventures of Tom Sawyer*
- In a linked data system, each part of a triples statement is expressed as a URI, so computers can infer relationships between entities





Triples and URIs



URN:

Twain, Mark, 1835-1910 —isAuthorOf→ The adventures of Tom Sawyer

is equivalent to

URL:

https://viaf.org/viaf/50566653 —http://schema.org/author→ https://viaf.org/viaf/178211495

Expressing Triples through Turtle

- One way to express triples is through **Turtle** (Terse RDF Triple Language), which is both machine- and human-readable
- Turtle provides a concise way to group three URIs to make a triple, and provides ways to abbreviate such information, for example by factoring out common portions of URIs

Example of Triples Expressed as Turtle

```
<http://www.w3.org/People/Berners-Lee/card#i>
        <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
        <http://xmlns.com/foaf/0.1/Person> .
<http://www.w3.org/People/Berners-Lee/card#i>
        <http://xmlns.com/foaf/0.1/name>
        "Tim Berners-Lee"@en .
<http://www.w3.org/People/Berners-Lee/card#i>
        <http://xmlns.com/foaf/0.1/name>
        "Τιμ Μπέρνερς Λι"@gr .
```

\- Same as -

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

<http://www.w3.org/People/Berners-Lee/card#i> a foaf:Person ;
        foaf:name "Tim Berners-Lee"@en , "Τιμ Μπέρνερς Λι"@gr .
```

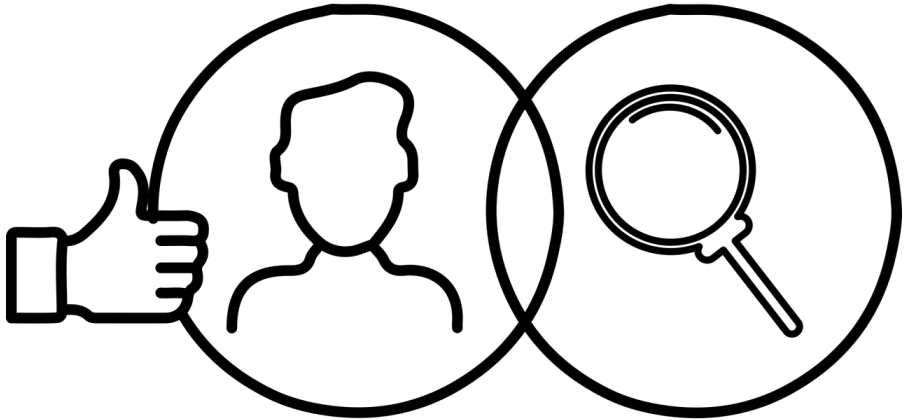
Example of Triples Expressed as Turtle

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix corp: <http://www.example.org/corp#> .

<http://www.pipian.com/people/pipian/card#me> a foaf:Person;
        foaf:name "Ian Jacobi"@en, "イアン・ジャコービ"@jp ;
        foaf:age 24 ;
        <http://purl.org/vocab/relationship/collaboratesWith> [
        is corp:director of [
        rdfs:label "World Wide Web Consortium"@en ] ] .
```

Linking Data: Reconciliation

- Vocabulary **reconciliation** is a process where automated systems use terms from unstandardized metadata to search controlled vocabularies and return URIs.
- Reconciliation is an integral part of linking data to other datasets.



OpenRefine for Data Cleaning and Linking

OpenRefine is powerful tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data.

You can use OpenRefine to:

- Get an overview of a data set
- Split data up into more granular parts
- Clean up tabular data by removing inconsistencies in format and terminology
- Enhance a data set with data from other sources

Installing OpenRefine

http://www.openrefine.org → Download → OpenRefine 2.8

- Windows Kit
  - Download the ZIP archive.
  - Unzip & extract the contents of the archive to a folder of your choice.
  - To launch OpenRefine, double-click on openrefine.exe.
- Mac Kit
  - Download the DMG file.
  - Open the disk image & drag the OpenRefine icon into the Applications folder.
  - Double-click on the icon to start OpenRefine.

Installing OpenRefine

OpenRefine runs locally on your computer. It does not require an internet connection, unless you want to reconcile your data with external sources.

If you close your browser, you can get back to OpenRefine by pointing it here: **http://127.0.0.1:3333/** or **http://localhost:3333**

Your data is not stored online or shared with anyone.

Sample Data Set

The data set we'll be using is a subset of the Powerhouse Museum Objects Database, which was made available under a CC-BY-NC license. The original data set has over 75,000 objects, but for simplicity, the test data set is a csv file with 1095 objects.

http://bit.ly/vv_data

OpenRefine Activity

Getting Started

Creating a project

- File formats, create/open/import

Exploring your data

- Rows, column headers, menus, “All” column, quick data review to make sure it looks ok

Manipulating columns

- Collapse, expand, rearrange, remove

Analyzing and Fixing Data

Faceting

- Facet by text on height (we'll do more with this later)
- Facet by blank on categories
- Duplicates facet on record id

Text filter

- Find a word or phrase (e.g. USA in title)

Analyzing and Fixing Data

Simple Cell Transformations

- Demo Trimming whitespace on Title
- Explain collapse consecutive/Unescape HTML/etc.

Splitting Multivalue cells

- Categories - demo facet unsplit
- Split - demo facet again

Analyzing and Fixing Data

Understanding row vs. record

- Demo with categories facet. Sometimes you want to be working with records and sometimes you want to be working with rows.

Clustering

- Demo clustering with split categories. Make sure facets are cleared.

Linking to Another Data Set

Linking to Wikidata (built into OpenRefine 2.8)

- Wikidata is a free and open knowledge base that can be read and edited by both humans and machines.
- Wikidata acts as central storage for the structured data of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wikisource, and others.
- Wikidata also provides support to many other sites and services beyond just Wikimedia projects! The content of Wikidata is available under a free license, exported using standard formats, and can be interlinked to other open data sets on the linked data web.

Linking to Another Data Set

- Add column “CatId” based on this column:
  “https://www.wikidata.org/wiki”+cell.recon.match.id
  - The wikidata reconciliation is returning the id. We want the URI
  - cell.recon = an object encapsulating the reconciliation results for that cell
  - recon.match = null, or the recon candidate that has been matched against this cell
  - id = the captured identifier for the object
- Merge:
  - Add column “CatAndId” based on this column two columns
  - cells['Categories'].value + ' (' + cells['CatId'].value + ')'
  - Join cells

Named Entity Extraction

Reconciliation works when your data is already in a structured format.

However, many fields (e.g. “description”) contain unstructured text, yet they usually convey a high amount of interesting information. To capture this in machine-processable format, named entity recognition can be used.

Named Entity Extraction Extension

http://freeyourmetadata.org/named-entity-extraction/ → Refine extension → Download

- Download ner-extension.zip and unzip it
- Copy the unzipped folder to your extensions folder
    - To find your extensions folder, choose Browse workspace directory from the Refine interface, and navigate to the folder extensions (which you should create if it doesn't exist yet).
- Start or restart Refine.
- Open or create a project
- Restart OpenRefine
- Open your project

Named Entity Extraction

- Run on Description Field
- Demonstrate links to DBPedia
    - DBpedia (from "DB" for "database") is a project aiming to extract structured content from the information created in the Wikipedia project. This structured information is made available on the World Wide Web. DBpedia allows users to semantically query relationships and properties of Wikipedia resources, including links to other related datasets. Tim Berners-Lee described DBpedia as one of the most famous parts of the decentralized Linked Data effort.
- Get uri --
    - Add column based on this column: cell.recon.match.id
- Join cells back together

Named Entity Extraction

- Get uri --
  - Add column “NERId” based on this column: cell.recon.match.id
- Merge:
  - Add column “CatAndId” based on this column two columns
  - cells['DBpedia Spotlight'].value + ' (' + cells['NERId'].value + ')'
  - Join cells

Exporting a Project

- Top left menu-
  - Export project lets you save everything, including history of edits (undo/redo). Share with others or move to another device
- Variety of data export options
  - Triple Loader and MQLWrite – must align with pre-existing schema (outside scope)
  - Custom table exporter – Tight control over export. Can select and order columns exported, control date format, reconciliation results
  - Templating – More advanced control using JSON (outside scope)

Metadata Problems?

Get

M.A.D.

ULS METADATA AND DISCOVERY UNIT