Wrap Up / Future Directions

LIS 598
Data Curation II

Agenda

- Some key takeaways from the course
- Tables, Trees, and Triples
- Triples: …always two years away
- Future directions

DC 1: Curation is the active and on-going maintenance of data through a lifecycle of use.

DC 2 : If data curation is a form of intervention (in that it mediates a relationship between a data producer and a data user) then we have to be grounded in a set of principles about whom and what to privilege, when, and for what reasons.

Curation should privilege:

- Efficiency through standardization. This is made possible through cleaning, normalization, and adhering to standards when encoding and describing data.

- Fitness for purpose. The amount of time and effort needed to prepare a dataset for sharing, reuse, etc are requisite to the importance of the dataset

- Ease computation. The work we conduct should make it possible to compute against data in ways that are easy, useful, and for the good of our designated community.

Data Representation

- Three levels of abstract for thinking about how data are represented

    - (level 1) Computing systems use encodings for binary (bits) that allow us to use something like UTF-8's **`01001000`** to represent an **H**

    - (level 2) We have sets of data that are encoded and use an exchange format to represent values, and structure of a dataset (e.g. CSV)

    - (level 3) We have formal data structures - like a relational database - where values, links (or relationships between) values, and functions of operations are clearly defined.

“The one-size-fits-all dilemma looks to be one of the biggest challenges to metadata management and creation. But I think it brings up a great question: why do we need or even want one schema to be representative of everything? Management of that schema would be a nightmare, and it would eventually fall apart”

This is exactly right.
We want to develop generalized approaches that allow for extension to meet particular needs.

- So, we choose INTEROPERABLE element sets that
  - 1. Ease access and reuse;
  - 2. Are expressive without becoming too difficult to mange;
  - 3. Enable efficient computation (and by that we might also mean discovery)

Tables.Trees.Triples

- Tables: Structure matrix like data. With relational structures we can create complex relationships, and enable ‘data independence’ - But, we do so at the expense of difficult to construct queries and complex normalization tasks.

- Trees: Hierarchical structure of HTML, XML, etc. allows for efficient and simple nesting of data. We can represent complex ideas - but they are computationally expensive, and we have more difficulty with data independence.

- Triples: Allow for assertions, reasoning, and efficient querying. We can express all kinds of relationships (not just hierarchies). But, the overhead of transforming the infrastructures for creating and managing tables and trees is going to

…Triples…

Or, why the promise of the semantic web is always 2 years away

- The web is replete with domain-specific informal / formal informational schemes… (e.g. baseball statistics, USA congress, Tax codes, street addresses, chemical properties of drugs, etc)

- Think of the power that could come with trying to find a domain-agnostic language that would allow anyone to assert anything at anytime (the rules of the web) … but following a set of conventions that enable **intelligent** ‘linking’ of this content…

- We already have a “structure” for linking with HTML … What linked data, RDF, etc. promise to do is provide both a structure and a semantics for this linking….

- Here is the problem… propositional statements are already numerous on the web (made through HTML)
    - We often assert that X is_a Y (e.g. Donald Trump - is - The President of the United States of America)
    - These assertions are often informal and unstructured. We don’t use HTML tags, for example, to identify anything other than a link to a new location (that is a direction for our browser to fetch some new information)
        - We can’t go rewrite all of the web…But these assertions are somewhat simple and easy to model in a formal language
- Subject - object - predicate; X is_a Y …
- Linked data

- **Nic Weber** is_a **Professor**…
    - Subject: Nic Weber = URL = nicweber.info
    - Predicate: is_a = URL = perl.namespace.info/is_a
        - Object: Professor = URL = perl.namespace.info/professionaltitles/Professor

- By assigning a URL or a namespace to all propositional content on the web… we have the ability to make assertions in a formal language….

- That was a fictional example…
- Here is a real one…Cheryl Metoyer

- https://www.wikidata.org/wiki/Q56855501

field of work	library science ▸ 1 reference	edit
	knowledge organization ▸ 1 reference	edit
		+ add value

employer	University of Washington ▸ 1 reference	edit
		+ add value

ethnic group	Cherokee ▸ 1 reference	edit
		+ add value

+ add statement

Identifiers

VIAF ID	113894620 ▸ 1 reference	edit
		+ add value

Identifiers

VIAF ID	113894620	edit
	1 reference	
	stated in: Virtual International Authority File	
	retrieved: 2 October 2018	
	+ add reference	
	+ add value	
Library of Congress authority ID	n92093176	edit
	1 reference	
	retrieved: 2 October 2018	
	stated in: Virtual International Authority File	
	+ add reference	
	+ add value	

{"entities":{"Q56855501":
{"pageid":56773023,"ns":0,"title":"Q56855501","lastrevid":919233307,"modified":"2019-04-19T20:20:00Z","type":"item","id":"Q56855501","labels":{"en":
{"language":"en","value":"Cheryl Metoyer-Duran"},"nl":{"language":"nl","value":"Cheryl Metoyer-Duran"},"es":{"language":"es","value":"Cheryl Metoyer-
Duran"},"ast":{"language":"ast","value":"Cheryl Metoyer-Duran"}},"descriptions":{"en":{"language":"en","value":"Cherokee researcher and professor of library and
information science"}},"aliases":{"en":[{"language":"en","value":"Cheryl Metoyer"},{"language":"en","value":"Cheryl A. Metoyer"},{"language":"en","value":"Cheryl
Metoyer Duran"}]},"claims":{"P31":[{"mainsnak":{"snaktype":"value","property":"P31","datavalue":{"value":{"entity-type":"item","numeric-
id":5,"id":"Q5"},"type":"wikibase-entityid"},"datatype":"wikibase-item"},"type":"statement","id":"Q56855501$B09F2511-F531-4963-98FC-
9AD6D19CC477","rank":"normal","references":[{"hash":"cfe647bfa43ced2e5aa986acf647f8ff15e93dde","snaks":{"P248":
[{"snaktype":"value","property":"P248","datavalue":{"value":{"entity-type":"item","numeric-id":54919,"id":"Q54919"},"type":"wikibase-
entityid"},"datatype":"wikibase-item"}],"P813":[{"snaktype":"value","property":"P813","datavalue":{"value":
{"time":"+2018-10-02T00:00:00Z","timezone":0,"before":0,"after":0,"precision":11,"calendarmodel":"http://www.wikidata.org/entity
/Q1985727"},"type":"time"},"datatype":"time"}]},"snaks-order":["P248","P813"]}]}],"P21":[{"mainsnak":{"snaktype":"value","property":"P21","datavalue":{"value":
{"entity-type":"item","numeric-id":6581072,"id":"Q6581072"},"type":"wikibase-entityid"},"datatype":"wikibase-item"},"type":"statement","id":"Q56855501$596ABCAA-
CF6B-4407-A4F0-A5F54F290ADB","rank":"normal"}],"P106":[{"mainsnak":{"snaktype":"value","property":"P106","datavalue":{"value":{"entity-type":"item","numeric-
id":121594,"id":"Q121594"},"type":"wikibase-entityid"},"datatype":"wikibase-item"},"type":"statement","id":"Q56855501$ae52624f-4f8d-7d62-843e-
335c085c764d","rank":"normal","references":[{"hash":"d963a5f40dfad5db26cf056ad88b4421f09bc141","snaks":{"P813":
[{"snaktype":"value","property":"P813","datavalue":{"value":
{"time":"+2018-10-02T00:00:00Z","timezone":0,"before":0,"after":0,"precision":11,"calendarmodel":"http://www.wikidata.org/entity
/Q1985727"},"type":"time"},"datatype":"time"}],"P854":[{"snaktype":"value","property":"P854","datavalue":{"value":"https://ischool.uw.edu/people/faculty/profile
/metoyer","type":"string"},"datatype":"url"}]},"snaks-order":["P813","P854"]}]}],"P214":[{"mainsnak":{"snaktype":"value","property":"P214","datavalue":
{"value":"113894620","type":"string"},"datatype":"external-id"},"type":"statement","id":"Q56855501$a8b36e6b-407f-fd49-7812-
fae54825ef68","rank":"normal","references":[{"hash":"cfe647bfa43ced2e5aa986acf647f8ff15e93dde","snaks":{"P248":
[{"snaktype":"value","property":"P248","datavalue":{"value":{"entity-type":"item","numeric-id":54919,"id":"Q54919"},"type":"wikibase-
entityid"},"datatype":"wikibase-item"}],"P813":[{"snaktype":"value","property":"P813","datavalue":{"value":
{"time":"+2018-10-02T00:00:00Z","timezone":0,"before":0,"after":0,"precision":11,"calendarmodel":"http://www.wikidata.org/entity
/Q1985727"},"type":"time"},"datatype":"time"}]},"snaks-order":["P248","P813"]}]}],"P244":[{"mainsnak":{"snaktype":"value","property":"P244","datavalue":
{"value":"n92093176","type":"string"},"datatype":"external-id"},"type":"statement","id":"Q56855501$061b0640-4891-4391-0085-
f125cc2fd9dc","rank":"normal","references":[{"hash":"cfe647bfa43ced2e5aa986acf647f8ff15e93dde","snaks":{"P813":
[{"snaktype":"value","property":"P813","datavalue":{"value":

https://www.wikidata.org/wiki/Special:EntityData/Q56855501.json

```
entities:
  Q56855501:
      pageid:                 56773023
      ns:                     0
      title:                  "Q56855501"
      lastrevid:              919233307
      modified:               "2019-04-19T20:20:00Z"
      type:                   "item"
      id:                     "Q56855501"
    labels:
      en:
          language:           "en"
          value:              "Cheryl Metoyer-Duran"
      nl:
          language:           "nl"
          value:              "Cheryl Metoyer-Duran"
      es:
          language:           "es"
          value:              "Cheryl Metoyer-Duran"
      ast:
          language:           "ast"
          value:              "Cheryl Metoyer-Duran"
  descriptions:
      en:
          language:           "en"
        value:                "Cherokee researcher and professor of library and information science"
```

Q56855501 - Cheryl Metoyer-Duran

Searching link targets on 1 Wikipedias...

Toggle existing labels

Item	Label [en]	Wikilabels	Link #
Q7414787 [CC |]	San Manuel Band of Mission Indians	San Manuel Band of Mission Indians	1
Q121594 [CC |]	professor	Professor	1
Q108266 [CC |]	Navajo people	Navajo	1
Q6973500 [CC |]	National Indian Education Association	National Indian Education Association	1
Q213422 [CC |]	Seneca people	Seneca people	1
Q1075148 [CC |]	University of California, Riverside	University of California, Riverside	1
Q1540453 [CC |]	Yakama Nation	Yakama	1
Q54919 [CC |]	Virtual International Authority File	Virtual International Authority File	1
Q7896574 [CC |]	University of Washington Information School	University of Washington Information School	1
Q219563 [CC |]	University of Washington	University of Washington	1
Q116971 [CC |]	Mohawk people	Mohawk people	1

https://tools.wmflabs.org/wikidata-todo/cloudy_concept.php?q=Q56855501&lang=en

Reasonator

Random item | English | Find | Other

Cheryl Metoyer-Duran (Q56855501)

Cheryl Metoyer | Cheryl A. Metoyer | Cheryl Metoyer Duran

Cherokee researcher and professor of library and information science

Cheryl Metoyer-Duran is a professor.
Her field of work includes library science and knowledge organization

knowledge organization

Q1929761

field of study related to Library and Information Science

library science

Other properties

field of work	library science field that applies the practices, perspe…gy, education, and other areas to libraries
	knowledge organization field of study related to Library and Information Science
employer	University of Washington public research university in Seattle, Washington, United States
ethnic group	Cherokee Native American people indigenous to the Southeastern United States

https://tools.wmflabs.org/reasonator/?q=Q56855501

- Encoding schemes like JSON-LD are making the promise of the semantic web, and linked data a closer reality

- Curation will increasingly need to attend to the real value created by the linked data world … But, honestly I took this same exact class over a decade ago and Allen Renear said the same thing to me then too…

- So, know linked data principles. Invest in some blogs that keep you up to data. When you teach this class in 2029 you can tell your class its just a few years away (but hopefully you also have some cool tools to show off)

- Future Directions for Data Curation
    - Bringing DC to the public sector
    - Curating, preserving, and working with Web Archives
    - Publishing to the web w/ APIs that are more interoperable and valuable
    - Privacy enhancing technologies will be a boon to researchers, and a big hurdle for curators.
    - DC will educate about best practices in managing sensitive data, but also need to understand how to work with poorly managed / secured data.