Describing Theses and Dissertations Using Schema.org

Project Report Presentation and Update

DCMI 2014 Austin, Texas

October 10, 2014

Jeff Mixter - OCLC Research

Patrick OBrien - Montana State Univeristy

Kenning Arlitsch - Montana State University

Background

- This project is based on an IMLS Grant that Kenning and Patrick were awarded in 2010
- Initial scope was to improve indexing and visibility of digital collections in Search Engines
- Since the release of Schema.org in 2011 the scope has expanded to include modeling IR material in a way that make them more visible to traditional search engines

Schema.org

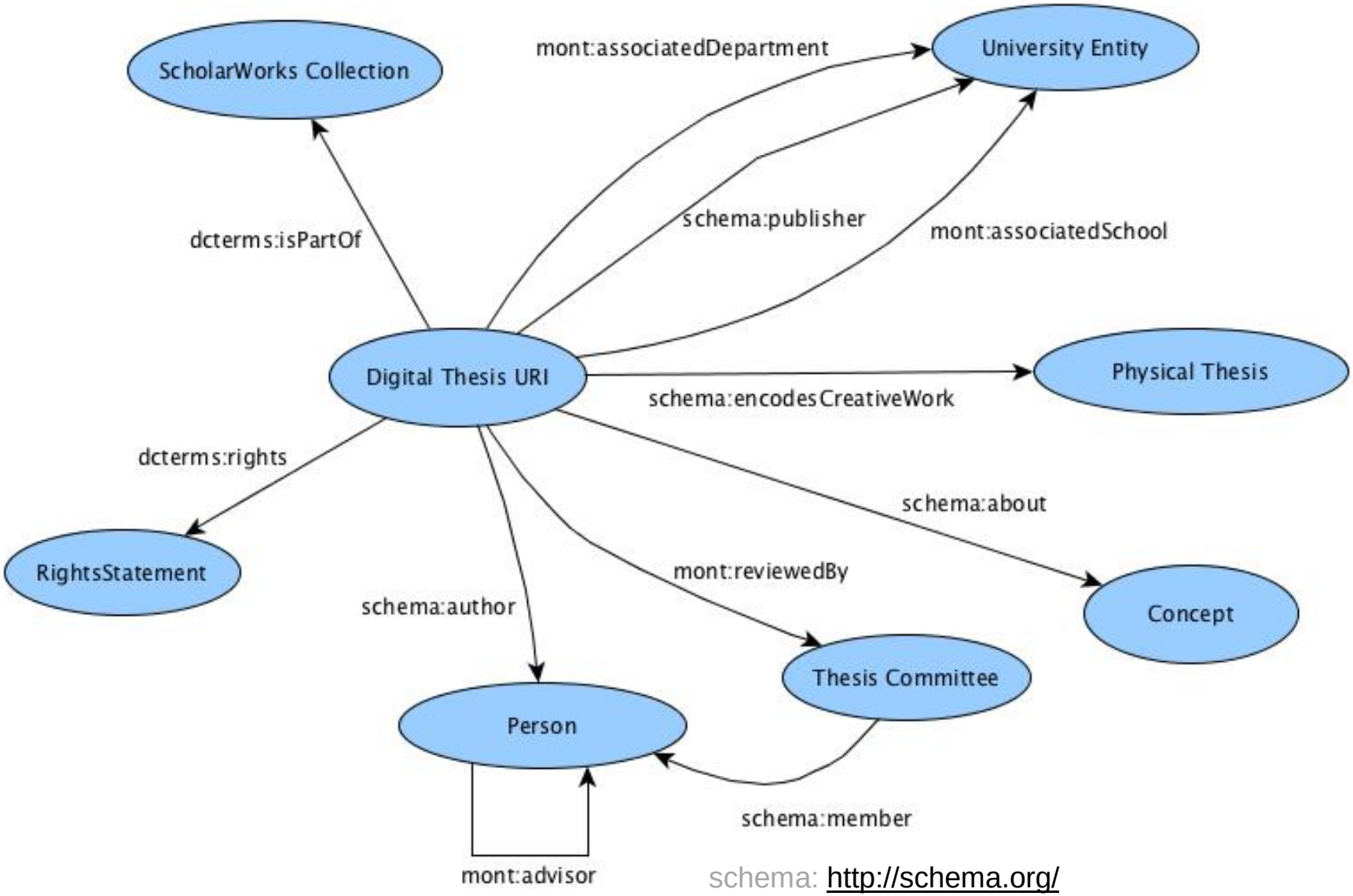
- Released in 2011 by Bing, Google, Yahoo and Yandex
- Lingua franca for describing things on the web
- W3C Working Group SchemaBibExtend was created to help make bibliographic recommendations and suggestions to Schema.org

Data Sample

- 1,909 DC records from the Montana State University ScholarWorks IR
- They had already undergone extensive metadata clean-up

Data Model

- Started with Schema.org as the base
- We created an extension vocabulary using the same mechanics and conventions used in Schema.org
  - RDFS vocabulary
  - It is published as RDFa
  - http://purl.org/montana-state/library/



schema: http://schema.org/
dcterms: http://purl.org/dc/terms/
mont: http://purl.org/montana-state/library/

Classes

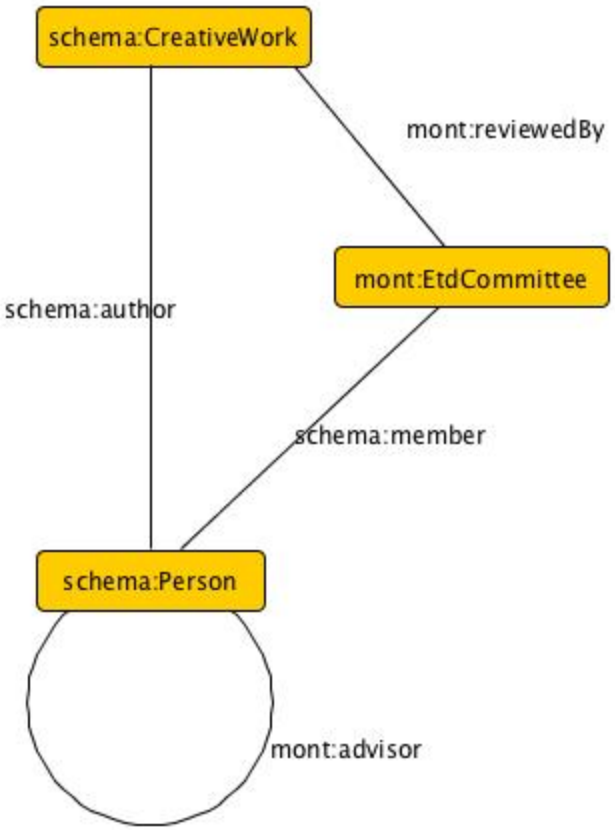
- There was a need to add more specificity to the existing Creative Work branch classes
  - Mont:Thesis
  - Mont:Concept
- There was also a need to describe entities unique to IRs and Universities that are not covered in Schema.org's current vocabulary
  - Mont:InstitutionalRepository
  - Mont:AcademicDepartment

Properties

- Create more granular relationships between classes
  - Mont:committeeMember
- Describe important attributes of Theses and Dissertations that were not included in Schema.org*
  - Mont:firstPage**
- Highlight and model unique relationships that were otherwise locked in the metadata records
  - Mont:advisor

* Schema.org underwent an update following the publication of the project report

** This property has since been replaced by the schema:pageStart



- Inferring additional information from the record
- This has the potential of allowing Universities to aggregate a large amount of data about Academic Output and use it for reviews/marketing
- This highlights the idea of developing a graph of university entities

Process Model

- Data was loaded into OpenRefine
- Data was reconciled against Dbpedia.org, LCSH and VIAF
  - Matching was made easier by the specific metadata fields that the records used
  - dc:subjects.lcsh matched 78%
- Generated our own internal URIs***

*** The URI pattern for the current production data differs from that used in the example data presented in the project report

Syndication of RDF data

- Data from three records was published online along with an HTML page that described all of the entities referenced in the CBDs
  - Serialized at RDFa
- Since then we have loaded all 1,909 RDF descriptions back into the ScholarWorks repository and tweaked the Dspace instance to pull over and display JSON-ld data
- All newly created entities are loaded into a Triple Store with a Pubby front end
  - http://54.191.234.158:8081/resource/department/7

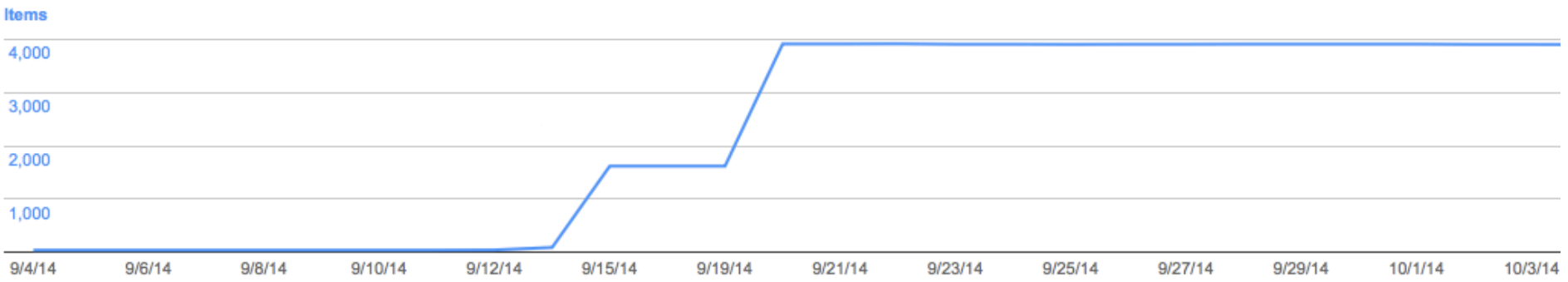
Google Webmaster Tools

Structured Data

Status: 10/6/14

3,922 Items
on 1,299 pages

0 Items with Errors
on 0 pages



Download

Show 25 rows 1-1

Data Type	Source	Pages	Items	Items with Errors
CreativeWork	Markup: schema.org	1,299	1,301	–

Next Steps

- Setup a more production ready Pubby interface
- Make modifications to the ScholarWorks structured data
- Make libraries visible on the Web
  - Build the presence of the library and its sub-organizations on the Semantic Web
  - Kenning, A., OBrien, P., Clark, J. A., Young, S. W. H. & Rossmann, D. (2014). Demonstrating Library Value at Network Scale: Leveraging the Semantic Web With New Knowledge Work. *Journal of Library Administration*, 54(5), 413-425.

Questions?

Thank You!

Jeff Mixter
mixterj@oclc.org
@JeffMixter
614-761-5159