DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications 2014

Describing Theses and Dissertations Using Schema.org

Jeff Mixter
OCLC, USA
mixterj@oclc.org

Patrick OBrien
Montana State University, USA
patrick.obrien4@montana.edu

Kenning Arlitsch
Montana State University, USA
kenning.arlitsch@montana.edu

Abstract

This report discusses the development of an extension vocabulary for describing theses and dissertations, using Schema.org as a foundation. Instance data from the Montana State University ScholarWorks institutional repository was used to help drive and test the creation of the extension vocabulary. Once the vocabulary was developed, we used it to convert the entire ScholarWorks data sample into RDF. We then serialized a set of three RDF descriptions as RDFa and posted them online to gather statistics from Google Webmaster Tools. The study successfully demonstrated how a data model consisting of primarily Schema.org terms and supplemented with a list of granular/domain specific terms can be used to describe theses and dissertations in detail

Keywords: Schema.org; RDF; linked data; institutional repositories; semantic web; search engine optimization; data modeling.

1. Introduction

As academic institutions realize the value of their intellectual output, well-organized and discoverable institutional repositories are increasingly viewed as strategic assets. The intellectual output of an academic institution is diverse and ranges from student theses and dissertations to conference proceedings, presentations, books, journal articles, and the datasets that support research conclusions. It is crucial for purposes of discovery to publish the metadata in a format that is easily understood, consumed and indexed by search engines and other machine-based data aggregators.

This project builds on research whose initial aim was to improve visibility of digitized special collections in commercial search engines, and was partially funded by the Institute of Museum and Library Services (IMLS). The first phases of research were successful in developing search engine optimization (SEO) strategies and methods, and led to the publication of a book (Arlitsch & OBrien, 2013). Beyond digitized special collections the research also revealed that institutional repositories (IRs) pose unique and complex problems to scholarly search engines, and as a result many IRs were not being consistently harvested and indexed. The project described in this report examines a specific subset of IR content, theses and dissertations. The scope of the project was to create a set of extension terms for Schema.org[1] that can be used to describe theses and dissertations and to create a process model that explains how we converted the existing Montana State University Dublin Core metadata into Linked Data. Following this proof of concept, we plan to explore how to integrate the new vocabulary into existing IRs so that they can provide search engines with more meaning and context, ultimately resulting in more accurate search results for users.

1.1. Data Sample

We used the Montana State University ScholarWorks IR dataset to drive and validate the modeling process that expanded and implemented the Schema.org vocabulary. This approach provided the group with a multitude of rich modeling examples and use cases but it also helped

[1] http://schema.org

https://doi.org/10.23106/dcmi.952136542

keep the process of modeling firmly grounded in the requirements presented by the data. The ScholarWorks dataset that was used for the study was a collection of student theses and dissertations. There were 1909 records in the sample, which had originally been described using Dublin Core (DC) and, where necessary, additional DC extensions for granular details. It should be noted that prior to use in this study, the ScholarWorks metadata was cleaned up to ensure that all of the fields were populated with information, where appropriate, and that the fields were used according to their proper definitions. This prior work mitigated the need to perform an initial review and cleanup in order to use the data, but IR managers who plan to implement structured metadata should be aware that this cleanup is a crucial first step.

2. Extension Vocabulary Development

In our initial review of the dataset, we tried to use existing vocabularies to describe theses and dissertations. It became evident when reviewing the sample data extracted from ScholarWorks that existing vocabularies alone were not robust enough to fully describe the items. Application Profiles were an attempt by the larger metadata community to develop a set of vocabulary terms that can be used within a specific context to describe unique items. The idea was that a metadata schema could be developed from a variety of existing schemas, modified if needed and then used to describe a unique set of items within the context of a specific application or domain (Heery & Patel, 2000). Sir Tim Berners-Lee referred to this same type of modeling as "cherry-picking" at the Gov 2.0 Expo in 2010, suggesting that nearly all of the vocabulary terms that one would need to describe an item already exist (Berners-Lee, 2010). The work around application profiles was recently restarted within the context of developing RDF application profiles. A DCMI Task Group has begun to investigate how RDF application profiles could be created and used to help with data validation.[2] An early example of picking and choosing RDF terms from a variety of vocabularies can be found in the British Data Model (Hodson, Deliot, Danskin, Rosie & Ashton, 2012). In this model, terms are taken from fifteen different vocabularies and combined to form a comprehensive model for describing bibliographic items.

We used the same approach to develop the extension vocabulary for the theses and dissertations sample set. Below is a table showing the vocabularies that we used.

TABLE 1: Vocabularies used in the project

Vocabularies used in the project	
Prefix	Namespace
schema	http://schema.org
dcterms	http://purl.org/dc/terms/
pto	http://www.productontology.org/id/
rdfs	http://www.w3.org/2000/01/rdf-schema#
mont	http://purl.org/montana-state/library/

In addition to Table 1, we created and published a VoID dataset description.[3] It includes information about the sample datasets, including dataset statistics. The extension vocabulary that we developed was not designed to be prescriptive. Rather, it was meant to be used with the entire Schema.org vocabulary. In this sense, our extension vocabulary provides a descriptive way for rationalizing existing descriptions of theses and dissertations as Linked Data without adding any constraints or validation requirements. As Linked Data graphs continue to grow in size, validation will obviously become an important topic and requirement for systems/services. Over the next few years, it will be interesting to observe the path that the RDF Application Profile Task Group

[2] http://wiki.dublincore.org/index.php/RDF_Application_Profiles
[3] http://purl.org/montana-state/scholarworks/sampledataset

https://doi.org/10.23106/dcmi.952136542

takes in dealing with validation requirements. The full list of extension classes and properties are available online.[4]

2.1. Classes

The new classes we developed for the extension vocabulary were divided into two unique categories. The first category included class extensions that were used to add a more granular description of the item being described. The labels for these classes were derived from the 'Appendix III – Types' controlled vocabularies used in the Citation Style Language.[5] Table 2 lists the first category of classes.

TABLE 2: Citation Style Language terms

Extension Classes derived from Citation Style Language terms
mont:JournalArticle
mont:MagazineArticle
mont:NewspaperArticle
mont:Bill
mont:Chapter
mont:ConferencePaper
mont:Entry
mont:Figure
mont:Graphic
mont:Interview
mont:LegalCase
mont:Legislation
mont:Manuscript
mont:MusicalScore
mont:Pamphlet
mont:Patent
mont:PersonalCommunication
mont:Report
mont:Speech
mont:Thesis
mont:Treaty

The second category of classes that was developed for the extension vocabulary included terms that were not covered by existing popular vocabularies but were required for the description of theses and dissertations. Table 3 lists the second category of classes that were created for the extension vocabulary.

TABLE 3: Extension Classes not covered by other vocabularies

Extension Class
mont:AcademicDepartment
mont:Collection
mont: School
mont:Concept
mont:DigitalCollection
mont:DoctoralThesis
mont:EtdCommittee
mont:InstitutionalRepository
mont:MasterThesis
mont:ScholarlyWork
mont:SpecialCollection

[4] http://purl.org/montana-state/library
[5] http://citationstyles.org/downloads/specification.html#appendix-iii-types

https://doi.org/10.23106/dcmi.952136542

A diagram of the classes and relationships used in the project can be found in Appendix I.

2.2. Properties

Although Schema.org has a wide variety of properties, the ScholarWorks instance data helped us identify use cases that required more granular terms to properly describe the item. We were able to create relationships between entities that were otherwise mashed together in the Dublin Core records. Figure 1 illustrates how we were able to identify individual people and committees and also define how they were related to each other.

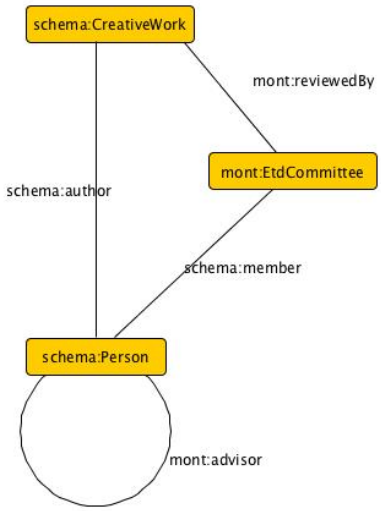


FIG 1: Relationships derived from DC records

Table 4 contains all of the properties that were created for the extension vocabulary as well as the type of Web Ontology Language (OWL) property that should be interpreted for each.

TABLE 4: List of Properties and OWL equivalencies

Extension vocabulary property	Object or Data property
mont:associatedDepartment	Object
mont:associatedSchool	Object
mont:adviser	Object
mont:campus	Object
mont:committeeChair	Object
mont:committeeMember	Object
mont:curates	Object
mont:facultyMember	Object
mont:hadDepartment	Object
mont:hasEtdCommittee	Object
mont:hasLibrary	Object
mont:reviewedBy	Object
mont:callNumber	Data
mont:degreeGrantedForCompletion	Data
mont:degreeGranted	Data
mont:firstPage	Data
mont:lastPage	Data

https://doi.org/10.23106/dcmi.952136542

3. Testing And Implementing The Model

After the model was developed, the entire ScholarWorks dataset was converted into Linked Data using a modified version of OpenRefine[6] called LODRefine.[7] Once the data were imported into LODRefine, a variety of data cleanup tasks were conducted and finally the Schema.org and extension vocabulary were imported and used to generate Linked Data. The first major cleanup task was to separate cells that contained multiple values into individual cells. After completing the cleanup we attempted to reconcile named entities to existing Linked Data datasets. We queried several datasets, including LCSH, VIAF and DBpedia. The most successful matching came from values that were included in the 'subjects', 'subjects.lcsh' and 'coverage.spatial' fields. The 'subject.lcsh' terms had a particularly high match rate (78% match to LCSH URIs) while the other fields matched at a lower rate (40% matched to DBpedia.org). The one problem with querying LCSH terms was that there were many pre-coordinated headings. Since the LCSH Linked Data dataset only includes terms that are part of the LCSH Authority files, there were quite a few terms that did not match up correctly. A solution to this problem would be to coin local URIs for the pre-coordinated headings and then include dc:hasPart or rdfs:seeAlso properties pointing out to the individual LCSH URIs that are referenced in the compound heading.

For the named entities that did not reconcile to the aforementioned datasets, local URIs were coined. These URIs followed a set pattern and then used the string value of the field as the identifier token. Figure 2 is an example of one of the URIs that was created when we could not match it to an existing Linked Data dataset.

http://scholarworks.montana.edu/doc/entities.html#person/Angie_Keesee

FIG 2: Sample URI coined for string value

More information about how to clean up dirty data and generate Linked Data using OpenRefine can be found in (Vorborgh & De Wild, 2013). In order to publish the Linked Data in a web-friendly serialization and to begin to test how much structured data search engines can mine, we converted three of the descriptions into RDFa and published them on ScholarWorks.[8] For all of the entities that did not have existing metadata records, such as people, places, organizations, etc, a single HTML page was generated that has a list of entity descriptions. The page is anchored with the URI tokens that appear after the #, so if one of these 'extra entity' URIs is resolved in the browser it will position the user in the appropriate portion of the page. The list can be found at Montana Scholar Works.[9]

3.1. Instance Data Example

In order to give a better understanding of the results of the modeling, this section walks through one of the sample records that was converted into Linked Data. The full RDFa description of this record is available online.[10] Figure 3 on the following page provides a graphic representation of the terms used to describe the item. The sample pictured in Figure 3 is also expressed in Turtle in Appendix II. The diagram does not list all of the properties and classes that can/should be used to describe theses and dissertations. A complete list of all of the terms used in the sample collection can be found in the Appendix III.

[6] http://openrefine.org/
[7] http://code.zemanta.com/sparkica/
[8] http://scholarworks.montana.edu/doc/index.html).
[9] http://scholarworks.montana.edu/doc/entities.html
[10] http://scholarworks.montana.edu/doc/SampleWork1.html

https://doi.org/10.23106/dcmi.952136542

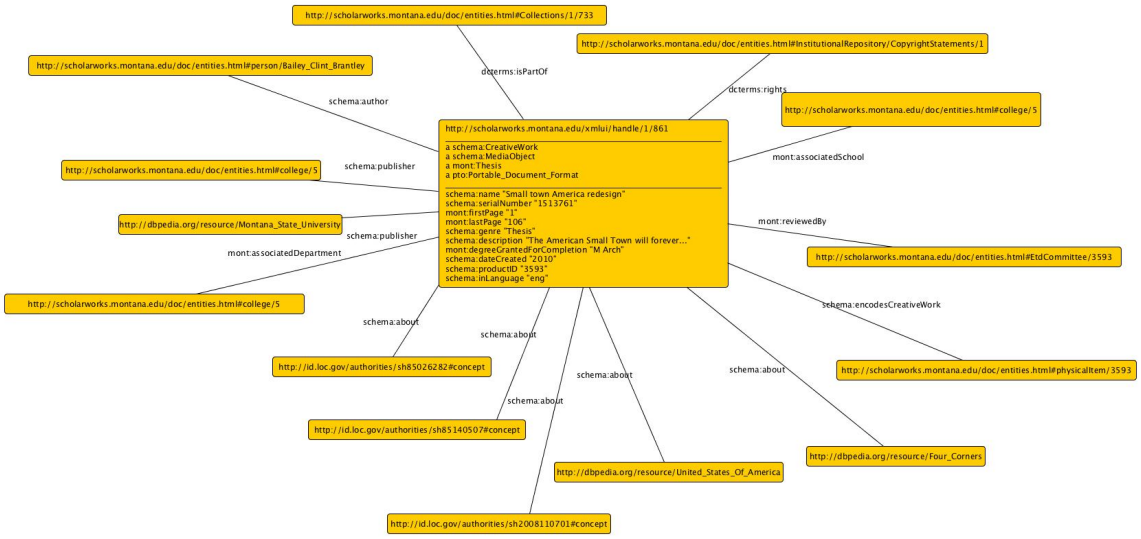


FIG 3: Graphical representation of sample item

4. Conclusion

We were able to successfully map the theses and dissertations metadata into Schema.org and, when needed, supplemented existing Dublin Core fields with terms we created as part of an extension vocabulary for Schema.org. The extension terms followed the same standards and practices as those in Schema.org and every attempt was made to position extension terms as sub-classes or sub-properties of existing Schema.org terms. The project has thus far successfully developed an extension vocabulary to describe theses and dissertations and show how to apply the vocabulary to existing metadata. Since modeling is an iterative process, the next step in the project will be to apply the vocabulary to more sets of theses and dissertations and make additions/changes. We also plan to publish more RDFa and begin to track the amount of structured data that is harvested by search engines using tools such as Google Webmaster Tools.[11]

Acknowledgements

This work could not have been completed without the help and assistance of Dr. Jean Godby and Jeff Young of OCLC Research.

References

Arlitsch, K., & OBrien, P. S. (2013). Improving the visibility and use of digital repositories through SEO. Chicago: ALA TechSource, an imprint of the American Library Association. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=578551

Berners-Lee, T. (2010). Open, Linked Data for a Global Community. Washington D.C. Retrieved from https://www.youtube.com/watch?v=ga1aSJXCFe0&feature=player_embedded

Heery, R., & Patel, M. (2000). Application profiles: mixing and matching metadata schemas. Ariadne, 25, 27-31.

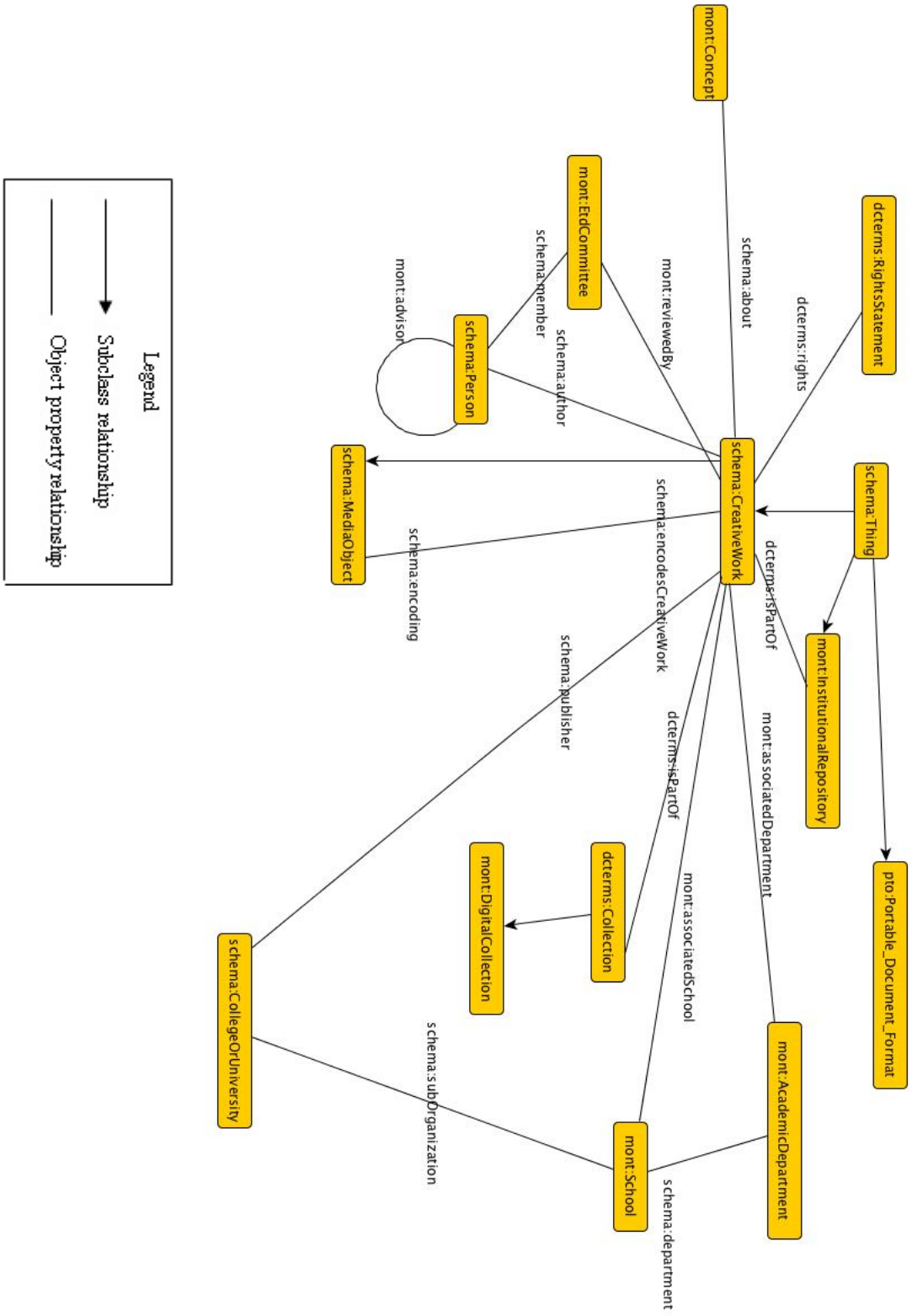
Hodson, T., Deliot, C., Danskin, A., Rosie, H., & Ashton, J. (2012). British Data Model – Book. Retrieved from http://www.bl.uk/bibliographic/pdfs/bldatamodelbook.pdf

Vorborgh, R., & De Wilde, M. (2013). Using OpenRefine: The essential OpenRefine guide that takes you from data analysis and error fixing to linking your dataset to the Web. Packt Publishing Ltd.

[11] https://www.google.com/webmasters/tools/home?hl=en

https://doi.org/10.23106/dcmi.952136542

Appendix I: Visual graph of the vocabulary terms used



https://doi.org/10.23106/dcmi.952136542

Appendix II: Sample data serialized as Turtle

```
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix ns1: <http://purl.org/montana-state/library/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix schema: <http://schema.org/> .
@prefix xhv: <http://www.w3.org/1999/xhtml/vocab#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://scholarworks.montana.edu/xmlui/handle/1/861> a schema:CreativeWork,
      schema:MediaObject, mont:Thesis,
      <http://www.productontology.org/id/Portable_Document_Format> ;
   dcterms:isPartOf <http://scholarworks.montana.edu/doc/entities.html#Collections/1/733> ;
   dcterms:rights <http://scholarworks.montana.edu/doc/entities.html#InstitutionalRepository/CopyrightStatements/1> ;
   ns1:associatedDepartment <http://scholarworks.montana.edu/doc/entities.html#college/5> ;
   ns1:associatedSchool <http://scholarworks.montana.edu/doc/entities.html#college/5> ;
   ns1:degreeGrantedForCompletion "M Arch" ;
   ns1:firstPage "1" ;
   ns1:lastPage "106" ;
   ns1:reviewedBy <http://scholarworks.montana.edu/doc/entities.html#EtdCommittee/3593> ;
   schema:about <http://dbpedia.org/resource/Four_Corners>,
      <http://dbpedia.org/resource/United_States_Of_America>,
      <http://id.loc.gov/authorities/sh2008110701#concept>,
      <http://id.loc.gov/authorities/sh85026282#concept>,
      <http://id.loc.gov/authorities/sh85140507#concept> ;
   schema:author <http://scholarworks.montana.edu/doc/entities.html#person/Bailey_Clint_Brantley> ;
   schema:dateCreated "2010" ;
   schema:description "The American Small Town will forever have a place in the undertones of American culture and
in the American psychy. The small town has become an identifing piece of the fabric that the overall American Society
as a whole uses to project its own image, not only to the world but to its self. This study is an examination of key
elements of the American Small town and an exploration into why these places are disappearing. The study goes on to
utilize this information to derive a plan for a small town that is free of modern day plights, such as sprawl and
redundency. In the end, it proposes a plan for the community of Four Corners, M.T. This case study re-design is an
example of how small communities can be shaped early on to prevent waste, maximize efficiency and quality of life." ;
   schema:encodesCreativeWork <http://scholarworks.montana.edu/doc/entities.html#physicalItem/3593> ;
   schema:genre "Thesis" ;
   schema:inLanguage "eng" ;
   schema:name "Small town America [electronic resource] : a re-design / by Clint Brantley Bailey.",
      "Small town America redesign" ;
   schema:productID "3593" ;
   schema:publisher <http://dbpedia.org/resource/Montana_State_University>,
      <http://scholarworks.montana.edu/doc/entities.html#college/5> ;
   schema:serialNumber "1513761" .
```


https://doi.org/10.23106/dcmi.952136542

Appendix III: List of classes and properties used in the study

Classes
schema:Intangible
schema:Person
schema:Organization
schema:CreativeWork
schema:CollegeOrUniversity
schema:EducationalOrganization
schema:MediaObject
pto:Portable_Document_Format
dcterms:RightsStatement
dcterms:Collection
mont:Concept
mont:EtdCommittee
mont:School
mont:InstitutionalRepository
mont:DigitalCollection
mont:AcademicDepartment
Object Properties
schema:subOrganization
schema:encoding
schema:author
schema:member
schema:encodesCreativeWork
schema:about
schema:department
schema:publisher
dcterms:isPartOf
dcterms:rights
mont:advisor
mont:associatedDepartment
mont:associatedSchool
mont:reviewedBy
Data Properties
schema:genre
schema:dateCreated
schema:inLanguage
schema:url
schema:serialNumber
schema:name
schema:productID
schema:description
mont:firstPage
mont:lastPage
mont:degreeGrantedForCompletion
mont:degreeGranted
rdfs:label

https://doi.org/10.23106/dcmi.952136542