Publishing Datasets from an Open Access Repository As Linked Data

10th International Conference on Open Repositories

Hui Zhang
Oregon State University
Agenda

- Background
- DSpace DB Schema
- Proposed RDF Model
- Mapping Relation DB to RDF
- Mapping Results
Electronic Theses & Dissertations

Digital representations of student master's theses and doctoral dissertations.
Research Data Share and Reuse

- Challenges
  - How to describe meaning of the data
  - How to facilitate trust
    - Producer
    - Purpose
    - Procedure
Research Data Share and Reuse

Possible solutions

- Treat data set as primary object, not bits
- Link data set to the research works in context
- Use controlled vocabularies in descriptive metadata
Project Motivation

- Facilitate metadata enhancement
- Investigate different models for representing data sets
- Build a prototype for new discovery interface
DSpace DB Schema
Portland Common Data Model

pcdm:Collection

pcdm:hasMember (m:m)
ore:aggregates (m:m)

pcdm:File

pcdm:hasFile (1:m)
pcdm:hasRelatedFile (1:m)

Access
Bitstream
Descriptive
Technical
Mapping RDB to RDF #1

SELECT h.handle AS thehandle, b.name AS filename, b.sequence_id AS sequence_id, bfr:mimetype AS mimetype, mv.text_value AS text_value
FROM handle AS h, item AS i, item2bundle AS ib, bundle AS bundle, bitstream AS b, bundle2bitstream AS bb, metadatavalue AS mv, metadataschemaregistry AS msr, metadatafieldregistry AS mfr, bitstreamformatregistry AS bfr
WHERE
  h.resource_type_id = 2 AND
  h.resource_id = ib.item_id AND
  ib.bundle_id = bb.bundle_id AND
  h.resource_id = i.item_id AND i.withdrawn = 'f' AND
  bb.bitstream_id = b.bitstream_id AND
  b.bitstream_format_id = bfr.bitstream_format_id AND
  bundle.bundle_id = ib.bundle_id AND
  bundle.name = 'ORIGINAL' AND
  b.deleted = 'f' AND
  i.in_archive = TRUE AND
  h.resource_id = i.item_id AND
  h.resource_type_id = 2 AND
  msr.metadata_schema_id = mfr.metadata_schema_id AND
  mfr.metadata_field_id = mv.metadata_field_id AND
  mv.text_value is not null AND
  i.item_id = mv.item_id AND
  msr.namespace = 'http://dublincore.org/documents/dcmi-terms/' AND
  mfr.element = 'rights' AND
  mfr.qualifier IS NULL
Mapping RDB to RDF #2

- @prefix map: <#>.
- map:bitstreams
  rr:logicalTable <#bitstream-view>;
  rr:subjectMap [ 
    rr:template 'http://sa.library.oregonstate.edu/bitstream/handle/{"thehandle"}/
    {"sequence_id"}';
    rr:class pcdm:Object;
   ];
  rr:predicateObjectMap [ 
    rr:predicate edm:dataProvider;
    rr:objectMap [ rr:constant 'Oregon State University' ];
  ];
  rr:predicateObjectMap [ 
    rr:predicate pcdm:hasFile;
    rr:objectMap [ rr:template 'http://sa.library.oregonstate.edu/handle/{"thehandle"}/
      bitstream/{"sequence_id"}/"filename"';
    rr:termType rr:IRI] ;
  ];
Mapping RDB to RDF #3

./dump-rdf -u dspace_db_user
  -p dspace_db_passwd
  -f TURTLE
  -j jdbc:postgresql://dspace_db_ip:port/dspace_db_name
dspace-r2rml-mapping.ttl > dspace-r2rml-output.ttl
Next Step: Consume Results

- Pick a RDF triple store
- Pick a software library for RDF
- build a web application that interacts with triple store with features:
  - ingest, update, and delete RDF records
  - provide a user interface to query SPARQL endpoint and access results supporting content negotiation
Thanks!

- [https://github.com/osulp/dspace2rdf-dev](https://github.com/osulp/dspace2rdf-dev)
- hui.zhang@oregonstate.edu