Publishing the University of Economics‘ academic bibliography database as linked data

Jitka Hladká,
FIS VŠE Praha
Jindřich Mynarz,
Národní technická knihovna v Praze

All about…

converting bibliographic records

to RDF-based data

Presentation Outline

- Academic bibliography database at the beginning
- Linked data
- Transformation process
- Interlinking
- Linked data user interface
- Design issues
- Legal issues
- Future challenges and possibilities

At the beginning…

Academic bibliography database:

- Bibliographic information
- Project / research & development information added

All provided in :

MARC standard format for bibliographic description

Academic bibliography database

- Publishing activities of the academic staff of the University of Economics, Prague
- journal articles, conference papers, lecture notes, monographs and monograph chapters
- Interesting resource of scientific research and development activities

…but is it ready to the Web environment?

Linked data

= „best practices for publishing and connecting structured data on the Web“

- Using URIs to identify things
- HTTP as a retrieval mechanism for resources
- RDF is a graph-based data model for representing structure and linking data about things

Linked data benefits

- linking the pieces of data, information, knowledge on the Web from disparate resources
- Machine processable information ready to use by web applications
- Improved retrieval and discovery services

What about our bibliographic data?

- RDF is familiar to people outside the library community
- Integration of bibliographic information into the Web
- Involve the trusted high-quality scientific and research outputs information into the growing Semantic Web

Transformation process

- Data preparation
- Data modelling
- Interlinking

…our experiences, comments and suggested problems' solutions to share

Problems of original data

- Bibliographic format - MARC
- „Library data“ made by „non-librarians“
- No cataloguing standard used
- No established classification scheme used
- No controlled vocabulary used: names used in many variations, translations and abbreviations

```
<marc:record>
  <marc:controlfield tag="001">00009792</marc:controlfield>
  <marc:controlfield tag="003">CZ-PrVSE</marc:controlfield>
  <marc:datafield tag="040" ind1="" ind2="">
    <marc:subfield code="a">ABA006</marc:subfield>
  </marc:datafield>
  <marc:datafield tag="041" ind1="0" ind2="">
    <marc:subfield code="a">cze</marc:subfield>
  </marc:datafield>
  <marc:datafield tag="044" ind1="" ind2="">
    <marc:subfield code="a">CZ</marc:subfield>
  </marc:datafield>
  <marc:datafield tag="100" ind1="1" ind2="">
    <marc:subfield code="a">Hindls, Richard</marc:subfield>
    <marc:subfield code="u">FIS</marc:subfield>
  </marc:datafield>
  <marc:datafield tag="245" ind1="0" ind2="">
    <marc:subfield code="a">Statistika - kvantitativní metody
  </marc:datafield>
  <marc:datafield tag="250" ind1="" ind2="">
    <marc:subfield code="a">1. vyd.</marc:subfield>
  </marc:datafield>
  <marc:datafield tag="260" ind1="" ind2="">
```

Publisher names

- Synsets created
- Joining all variant names of a single publisher together
- Preferred form of the name selected according to the **Publishers' Directory** maintained by the National Library of the Czech republic

Is there a need to transform **all the information** captured in MARC format?…

- Complex task
- Information discovery and retrieval on the Web
- Various domains of knowledge

…not required and not necessary, we need to select the **most essential and widely used** information.

Selection tools

Tags' frequency analysis of **MARC fields/subfields** used

→Final list of the **most frequent** fields and subfields

= the most essential bibliographic data pieces that we would like to use in the further transformation

Data modelling

Transformation into the RDF data model:

1. Entity types specification
2. Find the most appropriate rdf:type for the specified entities

…in available RDF vocabularies

and domain ontologies

Resource types

- Concept (skos:Concept)
- Document (bibo:Document)
- Organization (foaf:Organization)
- Person (foaf:Person)
- Project (arpfo:Project)
- Record (irw:InformationResource)

Selection criteria

1. **popularity** of vocabulary (ontology)

Well established and most popular vocabularies are preferred.

1. concept **description match**

*The description or meaning of the concept should **match** as closely as possible to MARC bibliographic data element meaning.*

Looking for the best concept…

Vocabularies and domain ontologies:

- representing bibliographic information
- common vocabularies providing domain agnostic terms widely used
- specialized on scientific research, academic environment and project's description

We felt free to combine various vocabularies to find the most appropriate descriptions for our data.

List of vocabularies and domain ontologies used

- Bibliographic Ontology (bibo) http://purl.org/ontology/bibo/
- Dublin Core (dc) http://purl.org/dc/elements/1.1/
- Friend of a Friend (foaf) http://xmlns.com/foaf/0.1/
- Academic Research Project Funding Ontology (arpfo) http://vocab.ouls.ox.ac.uk/projectfunding#
- Semantic Web for Research Communities (swrc) http://ontoware.org/swrc/swrc/SWRCOWL/swrc_updated_v0.7.1.owl
- …

…Not found?

Vocabularies

- under reconstruction
- newly established

- PlaceOfPublication problem – the appropriate property missing

`dc:coverage, dc:spatial`?

Project information modelling

- bibliographic data are enriched by information about the projects related to the described documents
- **ARPFO**, *Academic Research Project Funding Ontology* – describing the project funding structure of academic research
- **SWRC**, *Semantic Web for Research Communities* – modelling entities of research communities.

FRBRization?

Functional Requirements for Bibliographic Records (FRBR) abstract data model

= „*framework for relating the data in bibliographic records and defining the basic level of functionality for records created by national bibliographic agencies*“ (1)

- **Expression of Core FRBR Concepts in RDF** vocabulary describing basic concepts and relations http://vocab.org/frbr/core.html

FRBRization?

- classical library catalogue: *framework for relating the data in bibliographic records, collocation of multiple variations, versions, formats or editions of a single work to improve retrieval and displaying of bibliographic information*

Academic bibliography database ≠ classical library data

- Multiple versions, formats or editions of worhks occurring only rarely, if so
- Work, expression, manifestation, item
- not very suitable for our data

Interlinking

- Deterministic – exact matches
  - Country codes, languages
- Non-deterministic – probabilistic matching
  - Authors, documents
  - With manual revisions

Interlinking

- MARC Codes
- Geonames
- VIAF
- DBLP
- KEG web

Linked data user interface

- {3. principle} "*When someone looks up a URI, provide useful information, using the standards* (RDF, SPARQL)." (2)
- Various representations of a resource
  - HTML, RDF/XML, RDF/Turtle
- Various access mechanisms
  - SPARQL, search interface

Database of publication activity
University of Economics, Prague

cs en

Diversity of the fatty acids of the Nostoc species and their statistical analysis

URI	http://keg.vse.cz/dpc/id/document/26885
Types	bibo:Document
	bibo:DocumentPart
Title	Diversity of the fatty acids of the Nostoc species and their statistical analysis
Is part of	http://keg.vse.cz/dpc/id/document/788513
Creator	http://keg.vse.cz/dpc/id/person/913390
Contributors	http://keg.vse.cz/dpc/id/person/190977
	http://keg.vse.cz/dpc/id/person/856198
	http://keg.vse.cz/dpc/id/person/784076
Key words	http://keg.vse.cz/dpc/def/concept/216380
	http://keg.vse.cz/dpc/def/concept/59331
	http://keg.vse.cz/dpc/def/concept/740230
	http://keg.vse.cz/dpc/def/concept/122123
	http://keg.vse.cz/dpc/def/concept/349972
	http://keg.vse.cz/dpc/def/concept/326899
Language	eng
Page start	308
Page end	321
Document record	http://keg.vse.cz/dpc/doc/document/26885

Homepage About

Database of publication activity

University of Economics, Prague

cs en

Springer

URI	http://keg.vse.cz/dpc/id/organization/903703
Name	Springer
Type	foaf:Organization
Based near	http://sws.geonames.org/2761369/
	http://sws.geonames.org/4167147/
	http://sws.geonames.org/5391811/
	http://sws.geonames.org/5128581/
	http://sws.geonames.org/4164138/
	http://sws.geonames.org/2950159/
	http://sws.geonames.org/2907911/
Is publisher of	A Fast Probabilistic Bidirectional Texture Function Model
	A Pattern-Bassed Framework for Uncertainty Representation in Ontologies
	A Reasoning-Based Support Tool for Ontology Mapping Evaluation
	A Study in Empirical and 'Casuistic' Analysis of Ontology Mapping Results
	AIME 2007 : 07.07.2007 - 11.07.2007, Amsterdam
	Academic KDD Project LISp-Miner
	Accounting Reform in Transition and Developing Economies
	Accounting Reform in Transition and Developing Economies
	Accounting Reform in Transition and Developing Economies
	Accounting Reform in the Czech Republic
	Action Rules and the GUHA Method: Preliminary Considerations and Results
	Advances in Data Management
	Advances in Data Management
	Advances in Data Mining : 14.07.2007 - 18.07.2007, Lipsko
	Advances in Data Mining : 16.07.2007 - 17.07.2007, Lipsko

Database of publication activity

University of Economics, Prague

cs en

MSM6138439910

URI	http://keg.vse.cz/dpc/id/project/284721
Identifier	MSM6138439910
Type	arpfo:Project
	swrc:Project
Outcome documents	A Pattern-Bassed Framework for Uncertainty Representation in Ontologies
	A Pattern-based Framework for Representation of Uncertainty in Ontologies
	A Reasoning-Based Support Tool for Ontology Mapping Evaluation
	A Study in Empirical and 'Casuistic' Analysis of Ontology Mapping Results
	AOQL Plans by Variables when the Remainder of Rejected Lots is Inspected
	AOQL Plans by Variables when the Remainder of Rejected lots is Inspected
	AQUA (Assiting Quality Assessment): A system based on Semantic web and information extraction technologies to support medical quality labelling agencies
	Acceptance sampling by variables when the remainder of rejected lots is inspected - exact calculation of the LTPD plans
	Acceptance sampling by variables when the remainder of rejected lots is inspected - exact calculation of the LTPD plans
	Acceptance sampling by variables when the remainder of rejected lots is inspected - exact calculation of the LTPD plans
	Action Rules and the GUHA Method: Preliminary Considerations and Results
	Advanced approaches to XML document validation
	Ageing of the Population of the Czech Republic and its Economic Consequences in the Sphere of Pension Security and Financing of Health Care
	Amfiteatru Economic
	An Architecture for Mining Resources Complementary to Audio-Visual Streams
	An Impact of Longer Life Caused by Higher Attained Level of Education on the Social Insurance System
	Analysing Ontological Structures through Name Pattern Tracking
	Analysis of Consumption of Czech Households
	Analysis of PX index time series

Design issues

- Individual page for each resource
- Content negotiation (HTTP 303 redirects)
- Providing labels (title, name) instead of URIs for resources.
- Performance of SPARQL queries (caching is needed)
- Specific templates for distinct RDF types

URI patterns

- URI pattern for each resource type
  - /id/{resource type}/{resource ID}.{format}
    - e.g., /id/person/123, /def/concept/456.rdf
- Hierarchical pattern
- Neutral identifiers

Connecting with applications

- Data can be easily repurposed by other applications
- Added metadata
  - COinS (Context Objects in Spans)
    - OpenURL resolvers (SFX)
    - Citation managers (Zotero)
  - RDFa

Legal issues

- In Czech Republic public domain is not an option
- Open data licences have not been adapted to the Czech legislation
- Traditional Creative Commons licences are not a good fit for data

Future challenges and possibilities

- Further interlinking
- Data cleaning
- Synchronization with the original dataset
- Better user interface (visualizations)
- More appropriate licence

References

1) IFLA. Functional Requirements for Bibliographic Records – Final report 1998 [online]. Latest revision: 11 April 2010. Available from WWW: http://archive.ifla.org/VII/s13/frbr/frbr1.htm

2) BERNERS-LEE, Tim. Linked data [online]. W3C, Last change 2009/06/18. Available from WWW: http://www.w3.org/DesignIssues/LinkedData.html

Thank you for your attention.

…*any questions?*