ODI-RDF - TROUBLE WITH TRIPLES

IASSIST 2012-06-06, Washington - DC

Olof Olsson, Swedish National Data Service, Olof.Olsson@snd.gu.se
Thomas Bosch, GESIS - Leibniz Institute for the Social Sciences, Thomas.Bosch@gesis.org
Benjamin Zapilko, GESIS - Leibniz Institute for the Social Sciences, Benjamin.Zapilko@gesis.org
Arofan Gregory, ODaF - Open Data Foundation, Arofan.Gregory@earthlink.net
Joachim Wackerow, GESIS - Leibniz Institute for the Social Sciences, Joachim.Wackerow@gesis.org
Background/History

• There is an increasing demand for data (in the broadest sense) to become part of the Web of Linked Data (RDF)
  – Driven by Open Data initiatives
  – Driven by interest from researchers/students

• Over the past 3 years, an RDF vocabulary was developed for the Web of Linked Data based on SDMX
  – Called the Data Cube Vocabulary
  – Now part of the W3C’s Open Government project
  – Addressed to statistical aggregates and related metadata
  – While this work was initiated, it was understood that micro-data (especially metadata) would form an additional piece of work
Background/History (2)

• In 2011, one of the Schloss Dagstuhl workshops focused on an expression of DDI in RDF for describing microdata
  – Building on earlier discussions at Dagstuhl and elsewhere
  – Building on a decision to focus on this topic by the DDI Alliance
• At EDDI 2011, another workshop was held on the same topic
• Two outputs will be forthcoming:
  – A DDI-based vocabulary to support discovery of microdata sets
  – A DDI-based vocabulary, extending SKOS, for describing standard classifications
  – These will be published initially by the DDI Alliance
Why RDF for DDI?

• To increase visibility of data holdings using mainstream Web technologies (RDF)
  – Based on a proven metadata model – DDI!
  – Using an approach in line with best practice in the Linked Data community

• To increase the connection between research data sets and other resources
  – Users can provide layers of additional linking
  – Similar to some methods used in qualitative research

• To better identify opportunities for merging datasets and other emerging functionality such as inferencing

• To improve the quality of approaches to research data within the Linked Data community
  – Based on a single coherent set of standards
  – Leveraging the experience and knowledge of the DDI community
The Goal

• To have a single, proven, standard way of describing microdata within the Web of Linked Data using RDF

• To leverage existing metadata holdings within archives, data libraries, and data producers
  – DDI in all versions
  – Process of producing RDF should be automated

• To fit coherently into a broader RDF ”data” context
  – Microdata description and metadata
  – Aggregate data and metadata/tables
  – Classifications, concepts, and ”foundational” metadata holdings

• To increase the absolute number of RDF triples on the Web!
discovery use case

• Which studies are connected with a specific universe consisting of the 3 dimensions: time, country, and population?
• What questions with a specific question text are contained in the study questionnaire?
• What questions are connected with a concept with a specific label?
• What questions are combined with a variable with an associated universe consisting of the 3 dimensions time, country, and population?
• What concepts are linked to particular variables or questions?
• What representation does a specific variable have?
• What codes and what categories are part of this representation?
• What variable label does a variable with a particular variable name have?
• What’s the maximum value of a certain variable?
• What are the absolute and relative frequencies of a specific code?
• What data files contain the entire dataset?
study | universe

ddi:Study a ?study

?studyTitle dc:title

ddi:isMeasureOf

ddi:Universe a ?universe

?universe skos:definition

?universeDefinition

country = 'France' and time = '2005'
and population = 'age: 18-65'
SELECT ?studyTitle
WHERE {
  ?study rdf:type ddi:Study.
  ?study dc:title ?studyTitle.
  ?study ddi:isMeasureOf ?universe.
  FILTER (?universeDefinition ==
      "country = 'France' and time = '2005' and population = 'age: 18-65'"))
}
Study

isMeasureOf

Universe
**instrument | question | concept**

Diagram showing relationships between education, concepts, questions, instruments, variables, and universes. The diagram includes nodes labeled as follows:

- `education`
- `?conceptLabel`
- `skos:prefLabel`
- `ddi:hasConcept`
- `ddi:hasQuestionConcept`
- `?question`
- `ddi:Question`
- `ddi:variable`
- `ddi:hasMeasurementOf`
- `ddi:instrument`
- `ddi:Instrument`
- `?universe`
- `skos:definition`
- `?universeDefinition`

The diagram also contains a question: `What is your highest school degree?` and some additional metadata: `country = 'France' and time = '2005' and population = 'age: 18-65'`.
Study

isMeasureOf

Universe
values | value labels
variable | descriptive statistics
logical dataset | dataset | data file
conceptual model
RDF from codebook and lifecycle

DDI codebook → XSLT transformation → DDI RDF

DDI lifecycle → XSLT transformation → DDI RDF
2 models - one target

DDI Codebook
study description, variables/questions....
"flat" representation

DDI Lifecycle
all of the above + reusability (via refering)....

Mapping
A lot of flattening structures in both cases
Problems with Codebook to RDF

• Missing elements for instrument
  – Create one instrument for the whole study

• Not all elements are identifiable
  – Use other unique labels/names

• Local use of DDI dialects
  – Extensive testing needs to be done, possibly slightly modified xslt for each dialect
Create triples from DDI XML via XSLT

On google code:
http://ddixslt.googlecode.com/svn/trunk/ddi-rdf/

Test it for yourself

Testers and feedback needed

Developed by:
Olof Olsson, Swedish National Data Service, Olof.Olsson@snd.gu.se
Thomas Bosch, GESIS - Leibniz Institute for the Social Sciences, Thomas.Bosch@gesis.org
Benjamin Zapilko, GESIS - Leibniz Institute for the Social Sciences, Benjamin.Zapilko@gesis.org
DDI Codebook

<?xml version="1.0" encoding="UTF-8"?>
<codeBook
   ID="study1-1"
   xmlns="http://www.icpsr.umich.edu/DDI"
   xmlns:qb="http://purl.org/linked-data/cube#"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
>
  <stdyDscr ID="study1">
    <citation>
      <titleStmt>
        <title>title-study1</title>
        <IDNo>study1</IDNo>
      </titleStmt>
      <prodStmt>
        <producer abbr="XYZ" affiliation="affil org">producer1</producer>
        <fundAg abbr="ABC" role="infrastructure">fund org1</fundAg>
      </prodStmt>
    </citation>
    <stdyInfo>
      <subject>
        <keyword vocab="vocab">keyword1</keyword>
        <keyword vocab="vocab">keyword2</keyword>
        <keyword vocab="vocab">keyword3</keyword>
        <topicClass vocab="vocab">subject1</topicClass>
        <topicClass vocab="vocab">subject2</topicClass>
        <topicClass vocab="vocab">subject3</topicClass>
      </subject>
      <abstract>abstract-study1</abstract>
      <sumDscr>
        <timePrd date="2012-01-01" event="start"/>
        <timePrd date="2012-12-31" event="end"/>
        <timePrd date="2012-03-03" event="single"/>
        <timePrd date="2012-06-06"/>
        <nation>narrower location</nation>
        <geogCover>broader location</geogCover>
        <universe>universe</universe>
      </sumDscr>
    </stdyInfo>
  </stdyDscr>
</codeBook>
Generated by http://www.w3.org/RDF/Validator/
Future work on the mapping and DDI-RDF xslt

- New Branches planned
  - DDI Codebook (2.5)
  - DDI Lifecycle (3.1 & 3.2)

http://ddixslt.googlecode.com/
Acknowledgements

• The work on DDI-RDF has been started at
  – the workshop “Semantic Statistics for Social, Behavioural, and Economic Sciences: Leveraging the DDI Model for the Linked Data Web” at Schloss Dagstuhl - Leibniz Center for Informatics, Germany in September 2011
• and has been continued at
  – the follow-up working meeting in the course of the 3rd Annual European DDI Users Group Meeting (EDDI11) in Gothenburg, Sweden.

The authors have been supported by contributions of the participants of both events, Archana Bidargaddi (NSD - Norwegian Social Science Data Services), Franck Cotton (INSEE - Institut National de la Statistique et des Études Économiques, France), Richard Cyganiak (DERI - Digital Enterprise Research Institute, Ireland), Daniel Gilman (BLS - Bureau of Labor Statistics, USA), Marcel Hebing (SOEP - German Socio-Economic Panel Study), Larry Hoyle (University of Kansas, USA), Jannik Jensen (DDA - Danish Data Archive), Stefan Kramer (CISER - Cornell Institute for Social and Economic Research, USA), Amber Leahey (Scholars Portal Project - University of Toronto, Canada), Abdul Rahim (Metadata Technologies Inc., USA), John Shepherdson (UK Data Archive), Dan Smith (Algenta Technologies Inc., USA), Humphrey Southall (Department of Geography, UK Portsmouth University), Wendy Thomas (MPC - Minnesota Population Center, USA), and Johanna Vompras (University Bielefeld Library, Germany).