Let's Disco – Publishing person-level data as Linked Data.

Building a Highly-Complex Data Set with Disco, PHDD, XKOS and widely used vocabularies.

XXX[1], XXX[1], and Joachim Wackerow[1]

GESIS – Leibniz Institute for the Social Sciences, Germany
{thomas.bosch,benjamin.zapilko,joachim.wackerow}@gesis.org

Abstract. The DDI-RDF Discovery Vocabulary (Disco) is an RDF Schema vocabulary that supports the discovery of microdata sets and related metadata using RDF technologies in the Web of Linked Data. Disco can be used to discover data sets by searching for specific questions, topics, and geographical coverage. Disco is intended to provide means to describe microdata by essential metadata for the discovery purpose. The DDI (Data Documentation Initiative) is a structured metadata standard related to the observation and measurement of human activity.
For a publication of DDI metadata as Linked Data, widely accepted and adopted RDF vocabularies (e.g. RDF Data Cube, DCAT, SKOS, PROV-O and Physical Data Description) are reused to a large extend. In this tutorial it is shown how Disco is interwoven with these vocabularies. The combined usage of Physical Data Description, Disco, and DCAT enables the creation of data repositories which provide metadata for the description of collections, for data discovery, and for processing of the data.
In different real world use cases, we demonstrate the adoption and inclusion of Disco in existing information systems. Existing DDI instances are stored in form of XML documents which can be transformed into DDI-RDF and therefore exposed as Linked Data. We describe typical constraints on metadata of DDI metadata, as well as how to formulate them by different constraint languages (e.g. OWL 2) and how to actually validate them.

Keywords: DDI-RDF Discovery Vocabulary, Disco, Microdata Metadata, Data Documentation Initiative, Linked Data

1 Call for Participation

The DDI-RDF Discovery Vocabulary (Disco) is an RDF Schema vocabulary that supports the discovery of microdata sets and related metadata using RDF technologies in the Web of Linked Data. It is based on the DDI (Data Documentation Initiative) which is a structured metadata standard related to the observation and measurement of human activity. Since this data is highly complex, widely

accepted and adopted RDF vocabularies (e.g. RDF Data Cube, DCAT, SKOS, PROV-O and Physical Data Description) are reused to a large extend.

In different real world use cases, the presenters of the tutorial demonstrate the adoption and inclusion of Disco in existing information systems. Participants are encouraged to present their own use cases where Disco has been applied or will be used. Together with the presenters, participants will have the possibility to elaborate an RDF representation of their use cases and to formulate typical queries which are necessary to solve use case related problems.

This tutorial addresses Linked Data developers who want to publish Linked Data sets based on complex data like person-level microdata in combination with essential information like provenance, data aggregates and data catalogs.

2 Tutorial Description

The DDI-RDF Discovery Vocabulary (Disco)[1] is an RDF Schema vocabulary that supports the discovery of microdata sets and related metadata using RDF technologies in the Web of Linked Data. Disco can be used to discover data sets by searching for specific questions, topics, and geographical coverage. Disco is intended to provide means to describe microdata by essential metadata for the discovery purpose.

The DDI (Data Documentation Initiative) is a structured metadata standard related to the observation and measurement of human activity. It started out in the mid-1990s as a replacement for traditional archival code books documenting research data, and then branched off to cover the research data life cycle. Over time, as the data landscape has changed, the DDI XML specifications have evolved to add new coverage and functionality to respond to new user requirements. While the Data Documentation Initiative (DDI) is an international metadata standard with origins in the quantitative social sciences, it is increasingly being used by researchers and practitioners in other disciplines. The DDI specifications are also being used to document other data types, such as social media, biomarkers, administrative data, and transaction data. The specification itself is modular and can document and manage different stages of the data lifecycle, such as conceptualization, collection, processing, analysis, distribution, discovery, repurposing, and archiving. This tutorial aims to present the Disco vocabulary and its practical appliance with DDI metadata in detail.

For a publication of DDI metadata as Linked Data, widely accepted and adopted RDF vocabularies (e.g. RDF Data Cube, DCAT, and PHDD) are reused to a large extend. It is shown how Disco is interwoven with these vocabularies. The Data Cube vocabulary is a W3C standard for representing data cubes representing multidimensional aggregate data derived from microdata which is represented by Disco. DCAT is a W3C standard for describing catalogs of data sets. Physical data description (PHDD) represents data (tables) in a rectangular format. The data could be either represented in records with character-separated

[1] current specification is available at https://github.com/linked-statistics/disco-spec

values (CSV) or in records with fixed length. The combined usage of PHDD, Disco, and DCAT enables the creation of data repositories which provide metadata for the description of collections, for data discovery, and for processing of the data.

Existing DDI instances are stored in form of XML documents and validated against XML Schemas[2]. DDI-XML documents can be transformed into DDI-RDF and therefore exposed as Linked Data. For XML, XML Schema is the standard constraint language to validate XML documents. For RDF, however, there are multiple candidates for languages to express constraints on RDF data - but there is no standard language for RDF validation. In 2014, two working groups on RDF validation have been established: the W3C RDF Data Shapes working group[3] and the DCMI RDF Application Profiles working group[4]. These working groups address the formulation and validation of RDF constraints. We will describe typical constraints on metadata of DDI metadata, as well as how to formulate and to validate them. We will show how to formulate these constraints by different constraint languages such as OWL 2.

In 2012, we started the development of Disco in form of a one-week Dagstuhl seminar - another one-week Dagstuhl seminar and two one-week workshops followed. In 2014, we finished the technical review of Disco. We addressed more than ten mailing lists from the Linked Data and the DDI communities and got feedback which has been incorporated into the Disco specification. We are planning to officially publish Disco after three years of development in the first half of the year 2015 after an additional official technical review of the DDI Alliance[5].

2.1 Objectives of the Tutorial

The domain-specific origin of Disco increases the difficulties for adopting it by developers who are not familiar with the domain. However, person-level metadata is not restricted to a particular domain and may be an interesting data source for Linked Data developers. Hence, as our main objective, we want to provide assistance for creating such complex Linked Data sets using Disco. Because of the complexity, we especially want to demonstrate how Disco data sets can be connected with information represented using other existing vocabularies that are relevant in that context. Vice versa, we show how these other vocabularies and data represented with such can benefit from the interplay with Disco.

Another objective of the tutorial is to share our experience in developing an RDF vocabulary out of an existing and highly-complex domain-specific data model. We provide insight into the development and engineering process for creating Disco. Finally, since Disco is already being adopted in the DDI community, we aim to increase the use of Disco in the Linked Data community.

---

[2] XML Schemas of current DDI version available at http://www.ddialliance.org/Specification/DDI-Lifecycle/3.2/XMLSchema/

[3] http://www.w3.org/2014/rds/charter

[4] http://wiki.dublincore.org/index.php/RDF-Application-Profiles

[5] http://www.ddialliance.org/

2.2 Relevance to ESWC 2015

In this tutorial, the DDI-RDF Discovery vocabulary (Disco) and its interplay with other existing vocabularies is introduced. In contrast to other newly developed vocabularies that can be adopted easily by Linked Data experts in most cases, the complexity of Disco is higher because of its domain-specific origin and the tight interplay with other existing vocabularies. This makes Disco more difficult to be adopted by Linked Data experts that may not be familiar with the domain-specific background of person-level micro data. However, since the number of organizations that aim to publish their complex data as Linked Data grows and the number of researchers that want to conduct their research with real-world data grows, a tutorial on representing such highly complex and interconnected data sets may be a beneficial addition for the audience of ESWC 2015.

2.3 Scope and Level of Detail of the Material

During the tutorial, we built one real world complex data set which serves as ongoing running example. The whole data set can be exported from the Microdata Information System[6]. In all sessions, there will be practical hands-on exercises where the participants are asked to model and query particular parts of this data set. The infrastructure for these exercises, i.e. a triple store, will be provided for all participants. We will extend slides of a previous tutorial[7] and where appropriate refer to the current Disco specification[8].

2.4 Intended Audiences

The intended audience of this tutorial are Linked Data developers - beginners as well as experts - who are interested in modeling and publishing complex and highly inter-connected data sets as Linked Data. Since the connection with other vocabularies like QB, DCAT and PROV-O plays a major role in this context, the tutorial is not restricted to participants who are interested in person-level micro data.

2.5 Learning Objectives

The learning objectives of this tutorial are the following:

- Introduction of Disco, PHDD, and XKOS for modeling complex person-level micro data
- Representing the interplay of Disco data with existing vocabularies
- Validation of Disco

[6] http://www.gesis.org/missy/editor/

[7] tutorial slides available at: http://de.slideshare.net/boschthomas/201412-lets-disco-eddi-2014 and http://de.slideshare.net/boschthomas/201412-lets-disco-2-eddi-2014

[8] current specification is available at https://github.com/linked-statistics/disco-spec

2.6 Practical Sessions

During the tutorial, one complex data set will be built, which serves as ongoing running example. In all sessions, there will be practical exercises where the participants are asked to model and query particular parts of the example data sets. The infrastructure for these exercises, i.e. a triple store with example data, will be provided for all participants by the presenters.

3 Tutorial Length

We propose a full-day tutorial. In the previous tutorial about Disco, we experienced that a half-day tutorial offers only enough time to cover the most important aspects about Disco and to include few hands-on exercises. Since we now want to expand the focus of the tutorial to the interplay of Disco with other vocabularies and its validation, a full-day seems to be appropriate. Furthermore, we want to offer hands-on exercises again, because they were well received at the last tutorial.

4 Previous Versions or Related Tutorials

Thomas Bosch and Benjamin Zapilko held a half-day tutorial at the EDDI14 – 6th Annual European DDI User Conference in London[9]. The tutorial was well received and the presenters got overall positive feedback. Since the audience of Disco is two-fold - data professionals and the DDI community at the one hand and Linked Data experts at the other hand - the idea was to offer this tutorial at the ESWC conference as well. At the EDDI conference, the first part of the audience - data professionals and the DDI community - was addressed. At the ESWC conference, the second part of our audience - Linked Data experts - will be addressed.

5 Tutoring Team

Thomas Bosch[10] studied information systems, holds a degree in Computer Science (M.Sc.), and is currently a PhD student. His research topic is all about RDF validation. More specifically, he investigates how to validate any RDF constraints (formulated by multiple constraint languages) and how to express them generically. He held multiple presentations at international conferences and workshops (e.g. WWW, ISWC, DC, Dagstuhl seminars).

Benjamin Zapilko holds a degree in Computer Science and has recently finished his PhD thesis about methods for matching social science relevant Linked Data. The defense talk will be at the end of January. Beside his PhD research, he works on modeling and publishing social science relevant data as Linked Data

[9] http://www.eddi-conferences.eu/ocs/index.php/eddi/eddi14/schedConf/program
[10] http://boschthomas.blogspot.com

with the focus on data modeling and integration of heterogeneous data sources and information types. He presented his work at several international conferences.

Joachim Wackerow holds a degree in sociology. He is the vice chair of the DDI Alliance Technical Committee whose charge is to create, maintain, and enhance the DDI specification. He is the chair of the DDI Alliance RDF Vocabularies Working Group whose charge is to develop RDF vocabularies for efficient use of DDI metadata in the context of the Semantic Web/Linked Data. He organizes the annual European DDI user conferences and holds DDI introduction workshops.

References