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

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

Benjamin Zapilko[1], Thomas Bosch[1], and Joachim Wackerow[1]

GESIS – Leibniz Institute for the Social Sciences, Germany
firstname.lastname@gesis.org

Abstract. The DDI-RDF Discovery Vocabulary (Disco) is an RDF Schema vocabulary that enables the representation of person-level data sets and related metadata. It is based on DDI (Data Documentation Initiative) which is a structured metadata standard related to the observation and measurement of human activity. Using Disco, such data sets can be discovered in the Web of Linked Data by using RDF technologies, e.g. searching for specific questions, topics, and geographical coverage.
For a meaningful publication of DDI metadata as Linked Data, additional established RDF vocabularies (e.g. RDF Data Cube, DCAT, SKOS, PROV-O, XKOS, and PHDD) are reused to a large extend. This combined usage enables the creation of data repositories providing metadata of data collections, data sets, relationships between them and related provenance information.
In this tutorial, we show how Disco is interwoven with these vocabularies and how these connections can be used in order to publish highly complex data sets. In different real world use cases, we demonstrate the adoption and inclusion of Disco in existing information systems. Additionally, we present typical metadata constraints 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, Person-level Metadata, Data Documentation Initiative, Linked Data

1 Motivation

The DDI-RDF Discovery Vocabulary (Disco) [1] is an RDF Schema vocabulary which can be utilized to represent person-level data sets and related metadata as Linked Data. It is based on DDI (Data Documentation Initiative) which is a structured metadata standard related to the observation and measurement of human activity. The metadata of such data can get highly complex by e.g. necessary provenance information and relationships to data aggregates, data collections, and the physical data sources. Hence beside Disco, additional widely ac-

cepted and adopted RDF vocabularies (e.g. RDF Data Cube[1], DCAT[2], SKOS[3], and PROV-O[4]) are reused to a large extend [2] in order to enable a full data documentation, which is necessary for discovering such data.

In this tutorial, an introduction to Disco and its inter-connections to these vocabularies is provided. It is presented how these connections can be used in order to publish highly complex data sets. In different real world use cases and by building a data set serving as running example, the presenters of the tutorial demonstrate the adoption and inclusion of Disco in existing information systems. During practical exercises, participants will have the possibility to elaborate an RDF representation of the running example or their own data sets and to formulate typical queries. Participants are encouraged to present their own data sets and use cases where Disco has been applied or will be used. This tutorial addresses Linked Data developers (beginners and experts) who want to publish Linked Data sets based on complex data like person-level data where essential information like provenance, data aggregates and data catalogs is necessary.

Relevance to ISWC 2015. In contrast to other newly developed vocabularies that can be adopted easily 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 developers who may not be familiar with the domain-specific background of person-level or similar data. However, since the number of organizations that aim to publish their data as Linked Data grows and the number of Linked Data researchers that want to use real-world data grows, a tutorial on representing such inter-connected data sets may be a beneficial addition for the audience of ISWC 2015.

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[5]. 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 ISWC conference as well in order to address the the Linked Data community.

2 Tutorial Description

The DDI-RDF Discovery Vocabulary (Disco)[1, 2] is an RDF Schema vocabulary that enables the representation of person-level data sets and related metadata as Linked Data. It is based on DDI (Data Documentation Initiative)[6], a structured metadata standard related to the observation and measurement of human activity. While DDI is an international metadata standard with origins in the

[1] http://www.w3.org/TR/vocab-data-cube/
[2] http://www.w3.org/TR/vocab-dcat/
[3] http://www.w3.org/TR/skos-reference/
[4] http://www.w3.org/TR/prov-o/
[5] http://www.eddi-conferences.eu/ocs/index.php/eddi/eddi14/schedConf/program
[6] http://www.ddialliance.org/

quantitative social sciences, it is increasingly being used by researchers and practitioners in other disciplines. It is 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, its relationship to other necessary RDF vocabularies, and its validation.

For a publication of DDI metadata or similar metadata as Linked Data, additional established RDF vocabularies (e.g. RDF Data Cube, DCAT, SKOS, PROV-O, XKOS, and PHDD) are reused to a large extend. It is shown how Disco is interwoven with these vocabularies. These connections enable the representation of various relevant aspects, e.g. aggregate data derived from person-level data (micro data) by RDF Data Cube, PROV-O and Disco, data collections and catalogs (together with DCAT), and data (tables) in a rectangular format by using PHDD. XKOS can be used to represent statistical classifications.

When publishing data sets, their validation is a necessary requirement. While XML Schema is the standard constraint language to validate XML documents, there is no standard language for RDF validation. However, there are multiple candidates for languages to express constraints on RDF data. In 2014, two working groups on RDF validation have been established addressing these issues: the W3C RDF Data Shapes working group[7] and the DCMI RDF Application Profiles working group[8]. We will describe typical metadata constraints of Disco data sets as well as how to formulate them using different constraint languages such as OWL 2 and to validate them.

The development of Disco started in 2012 at a Dagstuhl seminar followed by additional seminars and workshops. In 2014, we finished the technical review of Disco after contacting multiple mailing lists from the Linked Data and the DDI communities. We are planning to officially publish Disco after three years of development in the first half of the year 2015.

2.1 Objectives of the Tutorial and Learning Objectives.

The main objective of this tutorial is to introduce Disco and to provide assistence for creating Linked Data sets using Disco. Person-level or similar metadata is not restricted to a particular domain and may be an interesting data source for Linked Data developers. Because of the complexity that comes along with such data, 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 verse, we show how these other vocabualries and data represented with such can benefit from the interplay with Disco. Another objective of the tutorial is to share our experience in developing a RDF vocabulary out of an existing and complex domain-specific data model like DDI. We

[7] http://www.w3.org/2014/rds/charter
[8] http://wiki.dublincore.org/index.php/RDF-Application-Profiles

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. The learning objectives of this tutorial are the following: (1) introducing Disco, PHDD, and XKOS for modeling complex person-level data, (2) representing the interplay of Disco data with existing and established vocabularies like RDF Data Cube, DCAT, and PROV-O, and (3) the validation of Disco.

2.2 Tutorial Format and Presentation Style.

The scope of the tutorial covers the introduction of the Disco vocabulary, its connected vocabularies PHDD and XKOS as well as their interplay with other established vocabularies. Additionally, the validation of Disco data sets will be covered. Slides of a previous tutorial[9] will be reused and extended. We will also refer to the current Disco specification[10]. Material for the practical exercises will be provided by the presenters (see Section 2.6). 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 a 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 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 RDF Data Cube, DCAT and PROV-O plays a major role in this context, the tutorial is not restricted to participants who are interested in person-level or similar data.

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

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

5 Tutoring Team

Thomas Bosch 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).

- affiliation: Gesis - Leibniz Institute for the Social Sciences
- email address: thomas.bosch@gesis.org
- homepage: http://boschthomas.blogspot.com/; www.gesis.org

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 modelling and publishing social science relevant data as Linked Data with the focus on data modelling and integration of heterogeneous data sources and information types. He presented his work at several international conferences.

- affiliation: Gesis - Leibniz Institute for the Social Sciences
- email address: benjamin.zapilko@gesis.org
- homepage: www.gesis.org
- primary contact person

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.

- affiliation: Gesis - Leibniz Institute for the Social Sciences
- email address: joachim.wackerow@gesis.org
- homepage: www.gesis.org

References

1. Thomas Bosch, Richard Cyganiak, Joachim Wackerow and Benjamin Zapilko. Leveraging the DDI Model for Linked Statistical Data in the Social, Behavioural, and Economic Sciences. Proceedings of the 2012 International Conference on Dublin Core and Metadata Applications
2. Thomas Bosch, Benjamin Zapilko, Joachim Wackerow and Arofan Gregory. Towards the discovery of person-level data: reuse of vocabularies and related use cases. Proceedings of the International Workshop on Semantic Statistics (SemStats 2013) collocated with the 12th International Semantic Web Conference (ISWC-2013)