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

## Building domain-specific Data Sets 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 metadata standard related to the observation and measurement of human activity. Using Disco, such data sets which are often used in different domains like the social, ecomical and behavioural sciences can be published as Linked Data. However, for a meaningful and complete publication of these data sets, additional established RDF vocabularies (e.g. RDF Data Cube, DCAT, SKOS, PROV-O, XKOS, and PHDD) are reused to a large extend. This combination allows for expressing information on metadata of data collections, single data sets, relationships between them, different versions of 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 complex and domain-specific data sets. In different real world use cases, we demonstrate the adoption and inclusion of Disco in existing information systems.

**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 the metadata standard DDI (Data Documentation Initiative) which enables the representation of the observation and measurement of human activity and which is often used in domains like the social, ecomical and behavioural sciences. 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 accepted and adopted RDF vocabularies (e.g. RDF Data Cube[1], DCAT[2],

[1] http://www.w3.org/TR/vocab-data-cube/
[2] http://www.w3.org/TR/vocab-dcat/

SKOS[3], and PROV-O[4]) are reused to a large extend [2] in order to enable a full data documentation, which is necessary to represent such data in full extent. In this tutorial, an introduction to Disco and its inter-connections to these vocabularies is given. It is presented how these connections can be used in order to publish complex domain-specific 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. Participants are encouraged to present their own data sets and use cases where Disco may be applied.

**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 interconnected domain-specific data sets may be a beneficial addition for the audience of ISWC 2015. This tutorial addresses Linked Data developers (beginners and experts) who want to publish Linked Data sets based on person-level or similar data as well as researchers who want to use such data sets for their research.

**Previous Versions or Related Tutorials.** Benjamin Zapilko and Thomas Bosch 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 the Linked Data community at the other hand - the idea was to offer this tutorial at the ISWC conference as well in order to address the Linked Data community in particular.

## 2 Detailed 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 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

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

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 domain-specific 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 (also known as micro data) by the RDF Data Cube vocabulary, 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.

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 and researchers. 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 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 a RDF vocabulary out of an existing and complex domain-specific data model like DDI. 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. The learning objectives of this tutorial are the following: (1) introducing Disco, PHDD, and XKOS for modeling complex person-level data, and (2) representing the interplay of Disco data with existing and established vocabularies like RDF Data Cube, DCAT, and PROV-O.

### 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. Slides of a previous tutorial[7] will be reused and extended. We will also refer to the current Disco specification[8]. 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 running example. There will be practical exercises where the participants are asked to model 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 Audience

The intended audience of this tutorial is Linked Data developers - beginners as well as experts - who are interested in modeling and publishing complex domain-specific and highly inter-connected data sets as Linked Data. Additionally, we address researchers who want to use such data sets for, e.g., running reasoning or matching algorithms. 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. Experiences from our last tutorial have shown that the size of a school class (15-30 persons) seems to be a reasonable number of participants. However, depending on the number of participants we are able to adjust the proportion of the practical session in order to avoid losing too much time by conducting the exercises.

## 4 Presenters

**Benjamin Zapilko** is a postdoctoral researcher and lead of the team Data Linking at the GESIS department Knowledge Technologies for the Social Sciences (WTS). He holds a degree in computer science and finished his doctoral research on publishing and matching of Linked Open Data in the application area of the social sciences at Mannheim University in early 2015. His research interests focus on the domain-specific application of semantic technologies and Linked Open Data. He presented his work at several international conferences and workshops. Benjamin Zapilko is member of the RDF Vocabularies Working Group of the DDI Alliance and co-organizer of the Library Track at the Ontology Alignment Evaluation Initiative (OAEI) since 2012. Together with Thomas Bosch, he held a tutorial about Disco at the EDDI14 – 6th Annual European DDI User Conference[9] in 2014.

– primary contact person

[7] 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

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

[9] http://www.eddi-conferences.eu/ocs/index.php/eddi/eddi14

– affiliation: GESIS - Leibniz Institute for the Social Sciences
– email address: benjamin.zapilko@gesis.org
– homepage: http://www.gesis.org/

**Thomas Bosch** studied information systems, holds a degree in Computer Science (M.Sc.), and is currently a PhD student in Computer Science. His research topic is all about RDF validation. He investigates how to validate any RDF constraints (formulated by arbitrary constraint languages) and how to express them generically. He presented his work at international conferences and workshops (e.g. WWW, ISWC, DC, and Dagstuhl seminars). Thomas Bosch is member of the RDF Vocabularies Working Group[10] of the DDI Alliance, part of the editorial board of the DCMI RDF Application Profiles Task Group[11], editor of the specifications DDI-RDF Discovery Vocabulary (Disco)[12] and Physical Data Description (PHDD)[13], and contributes to the W3C RDF Data Shapes Working Group[14].

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

**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: http://www.gesis.org/

## 5 Requirements

Besides a beamer and wifi access, there are no additional requirements for the tutorial. In case wifi access may not be available or interrupted, the intended hands on sessions will be presented on a locally installed triple store.

[10] http://www.ddialliance.org/alliance/working-groups#RDF
[11] http://wiki.dublincore.org/index.php/RDF-Application-Profiles
[12] http://rdf-vocabulary.ddialliance.org/discovery.html
[13] A vocabulary to represent data in a rectangular format in RDF; https://github.com/linked-statistics/physical-data-description
[14] http://www.w3.org/2014/data-shapes/charter

## 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)