Workshop: Interoperability of Metadata Standards in Cross-Domain Science, Health, and Social Science Applications

Schloss Dagstuhl – Leibniz Center for Informatics, October 1-5, 2018 in Wadern, Germany

Background

Standards are a vital tool enabling integration and semantic linking of data within and between disciplines. However, standards tend to get developed and adopted within disciplines or application domains with little consideration of cross-discipline requirements and technologies, so data integration can often only be easily achieved within and between closely allied fields. Addressing global scientific challenges that depend on cross-discipline integration remains difficult. The challenge is to make cross-discipline data integration a routine aspect of data-driven science.

Metadata support data discovery, selection, access and use, and are critical for data integration. Data from different sources/domains should be described in a way that cross-discipline discovery can detect and access the relevant data collections, and so that transformations and analyses can be automated. The use of cross-discipline data should become efficient, scalable and reproducible, enabling discipline-neutral data processing and analysis tools to be applied. Furthermore it would be possible to apply (meta-)data mining approaches and reasoning. In sum, new opportunities of insights and realization will develop.

A CODATA initiative on interdisciplinary data integration[1] is seeking to explore these challenges and opportunities in relation to three specific case studies in interdisciplinary research into infectious disease outbreaks, disaster risk and resilient cities. These case studies provide a concrete focus for exploring the potential of interoperability and data integration through metadata alignment.

[1] http://dataintegration.codata.org/

Focus of the Workshop

The workshop will build on a platform provided particularly by the following activities: (i) two previous workshops on DDI and interoperability with other specifications[2], (ii) work to extend and refine DCAT by the W3C Dataset Exchange Working Group (DXWG)[3], and (iii) the three detailed case studies and pilots from the CODATA initiative mentioned above. Metadata activities in the Research Data Alliance provide additional background and context.

There are several different areas where metadata comes into play:

- Description of studies or data collections for discovery purposes.
- Descriptions of provenance and scientific context for purposes.
- Description of data variables or dimensions for analysis purposes.
- Description of data transformation steps for recording purposes (possibly also for reusing the transformation steps on similar data).
- Controlled vocabularies to ensure standardized and agreed concepts (in relation to variables, collections, measurements, techniques and procedures etc.).

The capability to express discoverable and structured metadata must be automatic and achieved as far as possible using tools that are familiar and in common use.

Topics for Discussion and Possible Outcomes

Areas of exploration and discussion will identify and describe following:

- Common rules for metadata specifications
- Advantages and limitations of generic approaches
- Techniques for profiling or specializing generic standards for specific applications
- Best practices for setting up domain-specific data/metadata for cross-domain use
- Controlled Vocabularies, domain-independent and useful domain-specific ones
- Contact points/overlaps of specifications, crosswalks and transformations
- Identification of gaps. Possible workarounds, possible areas for future specifications

The output of the workshop will likely be reports and working documents on one or more of these topics.

[2] https://ddi-alliance.atlassian.net/wiki/spaces/DDI4/pages/39911463/Dagstuhl+Sprint+October+2016+Week+Two, https://ddi-alliance.atlassian.net/wiki/spaces/DDI4/pages/7864406/Dagstuhl+Sprint+October+2015

[3] https://www.w3.org/2017/dxwg/wiki/Main_Page

Metadata Specifications

The core objective of the workshop will be to investigate and advance alignment between the cross-disciplinary and domain-specific metadata standards, and to bridge from standards focusing on collection-level to variable-level metadata.

Metadata standards that may be considered include[4]:

- Study- or collection-level: DCAT, Dublin Core, ISO 19115-1, DDI 4
- Variable and dimension level
  - Microdata: DDI 4, W3C SSN, FHIR-HL7, CDISC, EML, SensorML, GSIM
  - Aggregate data: W3C DataCube, ISO 19123, Frictionless data
- Provenance: W3C PROV-O, ISO 19115-2
- Workflows/data transformation: DDI 4

Data transformations to prepare data for analysis may be described in machine-actionable form. DDI 4 uses some patterns of BPMN to achieve this, and CSV on the Web addresses transformation of tabular data into semantic form.

Additional relevant standards are likely to be uncovered during the development of the CODATA initiative.

The Dagstuhl Venue

Dagstuhl is an internationally recognized conference and research center for computer science. Dagstuhl provides excellent conditions for intense and long workshops. Participants stay overnight in the institute. The institute pursues its mission of furthering world class research in computer science by facilitating communication and interaction between researchers.

The Dagstuhl institute, Leibniz Center for Informatics, can be understood as remote retreat for scientific meetings. The whole place is designed to support intensive communication for week-long meetings. This way, a special dynamic develops which can be used for intense exchange and discussion. See references below.

Practical Information on DDI Dagstuhl Workshops

The workshops are 5 days in duration. This way, topics can be explored in depth. It is intended that participants stay the whole time. Otherwise the special dynamic of such a workshop can be disturbed.

There is a maximum of 25 people for these workshops. A plenary room and up to three smaller rooms for working groups are provided. Dagstuhl is supporting this kind of events. The participants have individual rooms in the institute for 70 Euro/night/person including full board (subsidized by Dagstuhl). Food, coffee, tea, and water are included in the accommodation rate.

The workshop will be start on Monday October 1st at 09:00 and will end on Friday October 5th at approx. 15:00 or earlier. Participants are strongly advised to arrive at Dagstuhl on Sunday September 30th. Further detailed information is available at the related wiki page[5].

[4] A list of links to these specifications and standards is given at the end of the document

[5] https://ddi-alliance.atlassian.net/wiki/spaces/DDI4/pages/449216513/Practical+Information

Daily Schedule

07:30 - 08:45	Breakfast
09:00 - 10:30	Workshop
10:30 - 10:45	Coffee Break
10:45 - 12:15	Workshop
12:15 - 13:45	Lunch
13:45 - 15:15	Workshop
15:15 - 15:30	Coffee Break
15:30 - 17:00	Workshop
18:00 - 19:00	Dinner
19:00 - 20:00	Possible evening session
Evening	Informal discussion (with drinks on own expense)
20:00	Cheese platter

References to metadata specifications:

- BPMN - Business Process Model and Notation http://www.bpmn.org/
- CDISC - Clinical Data Interchange Standards Consortium https://www.cdisc.org/standards/share
- CSV for the web - Comma-separated-variables for the Web http://www.w3.org/TR/tabular-metadata/
- DCAT - Dataset Catalogue vocabulary https://www.w3.org/TR/vocab-dcat/
- DDI - Data Description Initiative http://www.ddialliance.org/
- EML - Ecological Metadata Language https://knb.ecoinformatics.org/#external//emlparser/docs/eml-2.1.1/index.html
- FHIR - Fast Healthcare Interoperability Resources, HL7 - Health Level Seven International https://www.hl7.org/fhir/, http://www.hl7.org/
- Frictionless Data https://frictionlessdata.io/
- GSIM - General Statistical Information Model https://statswiki.unece.org/display/gsim/Generic+Statistical+Information+Model
- ISO 19115-1 - Geographic Information Metadata https://www.iso.org/standard/53798.html
- ISO 19115-2 Geographic information -- Metadata -- Part 2: Extensions for acquisition and processing https://www.iso.org/standard/67039.html
- ISO 19123 - Geographic Information – Schema for coverage geometry and functions https://www.iso.org/standard/40121.html
- PROV-O - The PROV Ontology https://www.w3.org/TR/prov-o/
- RDF Data Cube Vocabulary https://www.w3.org/TR/vocab-data-cube/
- SSN - Semantic Sensor Network vocabulary https://www.w3.org/TR/vocab-ssn/
- SensorML - Sensor Model Language http://www.opengeospatial.org/standards/sensorml
- TDWG - Biodiversity Information Standards http://www.tdwg.org/