Interoperability for Cross-Domain Research: Machine-Actionability and Scalability

Draft Workshop Description

The workshop will be held from 29 August to 2 September 2022 (participants should arrive on 28 August), at the Schloss Dagstuhl Leibniz Center for Informatics in Wadern, Germany.

A significant challenge facing the wide-scale implementation of the FAIR principles for data stewardship is the ready availability of metadata in a processable format, with sufficient context to support accurate and informed reuse and harmonisation of that data. This is an ambitious undertaking, and some parts of this challenge are more easily met than others. This workshop will use a set of use cases to inform how a framework of domain-agnostic standards and models can be usefully employed to permit the communication of the needed metadata, and how such metadata can be collected. Integral to this approach is the idea that both for collection and use, full advantage must be taken of the existing and emerging capabilities of artificial intelligence and machine-actionable metadata harvesting.

The focus of this workshop is three-fold:
1. Identifying the set of data models and standards which can be shared across domains for purposes of supporting machine actionable use and AI methods.
2. Exploring strategies for the practical harvesting and dissemination of the metadata, based on the identified standards and existing technology approaches.
3. Defining the contextual "package" of information which is needed to accurately share, reuse, and harmonise data.

These areas are interconnected, reflecting exploratory work from earlier Dagstuhl workshops over the past several years, and discussions in other fora. The scope spans a variety of concerns in the data sharing space in order to explore a range of current and emerging topics.

The standards and models to be considered will include, but will not be limited to, many which were considered in the 2021 Dagstuhl session Interoperability for Cross-Domain Research: Use Cases for Metadata Standards, which covered the FDOF, DDI-CDI, Schema.org, DCAT, DataCube, SDMX, PROV-O, ODRL, and DPV. More recent work around observable properties, based on the I-ADOPT work in RDA will also be addressed. Relevant cross-domain areas of practice such as DRUM (Digital Representation of Units of Measurement) will also be considered. The relationship of these standards to use-case-specific domain standards and vocabularies will be addressed, in order to explore how the interchangeable expression of metadata can be practically realised at scale.

Real-world use cases will be used to test the ideas put forward in the workshop, and demonstrate their practicability. The final selection of use cases has not yet been finalized, and will depend to some extent on who is available to attend. Currently, the case studies being considered include:

(1) Primary and Reference Data - Integration of reference data and primary data (for example, the European Social Survey case with environmental data; Smart Energy Research Laboratory; projects using geolocated social data, government statistics being integrated across ministries for use as an integrated resource, etc.). The temporal and geographical matches between data streams can be very important here, along with practical approaches to making data useful from the perspective of research questions and potential policy uses.

(2) Sensitive Data/Micro- and Macro-Data - Reuse of microdata across institutional boundaries often conflicts with the need to ensure data confidentiality. Data is often held in disparate systems, complicating access. The linking of aggregate data with the supporting microdata most useful for scientific research is also inhibited by the same barriers. Public health data - especially as regards the recent COVID epidemic - is an example where the microdata themselves require integration across institutions, and feed upward into highly visible and high-demand data such as the SDG indicators. Navigating the links between data at different levels while protecting confidentiality is a difficult challenge which will benefit from agreed approaches and standards.

(3) Oceans and Disasters/Geography and Phenomena Terminology - There has been a lot of work done in the UN agencies (for example in relation to the Ocean Data Information System) and in other domains around harmonizing semantics, and we have also seen some relevant work coming out of the fields of disaster risk reduction, geophysics and environmental monitoring data. Investigation of the effects of climate change is a key element in avoiding disasters and mitigating their impact. This area remains challenging, but it is very important to have a more general approach to combining population data with hard science data in the context of climate change. The idea is to integrate the approaches from the Oceans project and elsewhere into the broader guidance for interoperability.

(4) Describing Physical Samples - For many sciences (including biodiversity, crystallography, nanomaterials and geochemistry) the description and characterization of physical samples and specimens is central to the integration and reuse of data. Some progress has been made toward sufficient digital description of samples within specific domains. These approaches can potentially be used as the basis for a more generalized way of describing physical samples to support data sharing across domains. Connecting descriptions of samples, their environment, and their connection to other quantitative data, are all topics of interest.

The workshop will produce recommendations based on the proposed standards and models and on their application to the use cases considered. Proposed next steps for the use cases will be documented, to illustrate the basis on which the recommendations are formed. Any extensions or changes to the standards and models considered will be documented for communication to the relevant groups which maintain them. The intent of the workshop is to provide concrete input into the formulation of a core interoperability framework for FAIR data-sharing taking into account requirements emerging from user-driven communities, large scale infrastructures and significant domain organisations. Although space and logistics at Dagstuhl mean that not all the WorldFAIR project partners and case studies can be included, this event will certainly feed into the work of that project.