# DaMaLOS - First Workshop on Data and Research Objects Management for Linked Open Science

## Co-located at the International Semantic Web Conference ISWC 2020

Leyla Jael Garcia Castro [1] [0000-0003-3986-0510], Sören Auer [2,3] [0000-0002-0698-2864], Maria-Esther Vidal [2,3] [0000-0003-1160-8727], Dietrich Rebholz-Schuhmann [1,4] [0000-0002-1018-0370]

[1] ZB MED Information Centre for Life Sciences, Gleueler Str 60, 50931 Cologne, Germany
[2] Leibniz Information Center for Science and Technology, Welfengarten 1 B, 30167 Hannover, Germany
[3] L3S Research Center, Leibniz University Hannover, Germany
[4] University of Cologne, Albertus-Magnus-Platz, 50923 Cologne, Germany

**Abstract.** Research data is core to data-driven research as it complements scientific publications and expose results from experimental work. Most research activities follow the research data cycle, where data is continuously used, modified and produced, transitioning from one research group to another. For this cycle to prosper, we require Research Data Management plans supporting the findable, accessible, interoperable and reusable (FAIR) principles. Despite playing an important role, data on its own is not sufficient to establish Open Science nor Linked Open Science, i.e., Open Science plus Linked Open Data (LOD) principles. To get the full picture, in addition to data, we also need other digital objects playing a role in research, e.g., software and workflows. In this workshop we will explore what is required for Research Data (and other digital objects) Management Plans to effectively instantiate Linked Open Science, including effective support for LOD, automation by, e.g., machine/deep learning approaches, and innovations to include supporting data elements such as the software used to produce/consume it or the tutorials showcasing usage and fostering further developments.

**Keywords:** Data management, research objects, linked data, open science.

## Editorial Note

The first workshop on Research Data Management for Linked Open Science (DaMaLOS 2020) took place in the virtual space on 2nd November 2020. It has been organized by Leyla Jael Garcia Castro and Dietrich Rebholz-Schuhman, both from ZB MED - Information Centre for Life Sciences (Cologne, Germany), and Sören Auer and María-Esther Vidal from the Leibniz Information Centre for Science and Technology, University Library (Hanover, Germany).

Over recent years the scientific community, in particular in natural sciences and life sciences, has been undergoing a transformation where data science and big data enables new findings and new ways to plan and perform experiments. Data generation, data management, sharing of FAIR data, reuse of data and embedding of data into other existing information types enter into routine work of the scientists, and data and information science has to provide the means to improve the impact from sharing data. This workshop enabled exchange of ideas and technologies, which help to integrate interoperable data into existing data and deliver the benefits into the scientific community.

**Keynote**

In the keynote talk (Carole Goble, "The swings and roundabouts of a decade of fun and games with Research Objects", The University of Manchester) explored the current status on how to engage the research community into manageable sharing of FAIR data as an opportunity. This requires to understand that digital objects (in science) have to be categorized into different types (models, data, pipelines, lab protocols, SOPs, others); however, they could be subsumed under the label "Research Objects" (RO). Such ROs would package the digital object with metadata using standards and standardized code and identifiers leading to FAIR Research Objects and to the concept of the RO-Crate to meet the needs of the users and the research community, certainly using semantic technologies. Future will tell, where to find the right balance between data and metadata to be efficient.

**Invited talk**

The invited talk (Yaser Jaradeh, Open Research Knowledge Graph, TIB) was focused to the use of the Open Research Knowledge Graph (ORKG) to move from scientific publications to an explicit representation of scientific facts in a graph infrastructure. This provides benefits through the reuse of claims from the publication, e.g., statistical results, the exploitation of tabular results and the integration of the metadata, e.g., reference information, for the scientific community. Several examples demonstrate, how the full body of information initiatives new research.

**First session: Data Management Plans**

The first session of the workshop was dealing with "Data management plans" (chaired by Sören Auer, TIB). The first talk ("ASIO: a Research Management System based on Semantic technologies", Jose Emilio Labra Gayo) demonstrated, how in the ASIO infrastructure and the Hercules project metadata and semantic resources enable data management in such a way that functional objects, e.g., Java objects can be derived. This approach shows that data is not only available for statistical analysis, but also for algorithmic interoperability and transformation.

The next talk ("Data Management Plans and Linked Open Data: exploiting machine actionable data management plans through Open Science Graphs", Elli Papadopoulou) gave an overview, how solutions for data management (e.g., Argos,

"argos.openaire.eu") support the process of data management and lead to the integration of the data into the Open Science Graph. This approach altogether leads to the efficient provision of data into the existing repositories, i.e., into the OpenAIRE Research Graph.

Then ("Towards semantic representation of machine-actionable Data Management Plans", João Cardoso) laid out the initiatives by the Research Data Alliance (RDA) for Data Management Plans (DMP) and overall the RDA DMP Common Standards (DCS) Working Group (https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard). One important outcome is the DMP Common Standard Ontology (DCSO), which supports the encoding the metadata related to the origins of the data and the corresponding identifiers. Altogether, such standards are the driver for machine actionable DMPs.

Finally ("Research Object Crates and Machine-actionable Data Management Plans, Maroua Jaoua) explored how to generate RO-Crates from Machine-actionable Data Management Plans (maDMPs) and vice versa, and what difficulties would occur. This approach was tested on existing data objects and lead to an assessment, how many properties could or could not be mapped between the different approaches, where a larger number of properties from maDMPs and a smaller number of properties from RO-Crates.

**Second session: Beyond data and Life Sciences**

The second session had the topic "Data, Life Sciences and beyond" (chaired by Dietrich Rebholz-Schuhmann, ZB MED). In the first talk ("FAIR Data Management to Access Patient Data", Núria Queralt Rosinach) an extensive approach for the integration of clinical anonymised data with external data has been demonstrated. The work done include the use of semantic technologies and the automatic collection and linking of data from different resources, e.g., clinical measurements collected from the laboratories and RNA-Seq data. The final semantic model has been demonstrated on the integration of cytokine measurements for Covid-19 patients.

The second talk ("Towards in-situ knowledge acquisition for research data provenance from electronic lab notebooks", Max Schröder) used BFO-compliant ontologies and in particular Prov-O (provenance ontology) for data integration of data from the electronic laboratory notebook. Instances of classes represent particular entities, i.e., the representatives of the experimental results. The outcome demonstrates the complex task of bringing the data resources together.

In the next talk ("Understanding Semantic Search on Scientific Repositories: Steps towards Meaningful Findability", Thiago Gottardi) an overview on published approaches for semantic retrieval and integrated searches has been given. The overview distinguishes different research areas, shows the development over time (in recent years), identifies the use of different kind of meta-data in the approaches chosen and distinguishes the types of approaches anyways. Altogether, as expected, metadata is crucial for semantic retrieval.

The final talk ("Software as a first-class citizen in research", Leyla Garcia) introduced the Software Management Plan as a means to apply FAIR principles to software similar to the way, how DMPs support reuse of data. In contrast to data, software is

more volatile and reuse comes in different flavors, in particular software can be connected via workflows, i.e., metadata is key for the software interoperability. For reusability, software dependencies and licenses have to be considered, and for reproducibility the versioning, software repositories (e.g., Git) and testing are crucial.

**Final words**

The workshop was well perceived (about 38-42 participants) and the final discussion raised the comment and concern that only an efficient infrastructure, i.e., for DMPs, data reuse and metadata generation, gives the benefits that scientists would expect. Exploring the use cases with the scientific community is paramount anyways.