Case Statement for RDA WG DMP Common Standards

Contents

- WG Charter
- Value Proposition
- Engagement with existing work in the area
- Work Plan
- Adoption Plan
- Initial Membership

***WG Charter**: A concise articulation of what issues the WG will address within an 18 month time frame and what its "deliverables" or outcomes will be.*

The need for establishing this working group was articulated during the 9th plenary meeting in Barcelona during the Active DMPs IG session. The discussion was framed by a white paper by Simms et al. on machine-actionable data management plans (DMPs). The white paper[1] is based on outputs from the IDCC workshop held in Edinburgh in 2017 that gathered almost 50 participants from Africa, America, Australia, and Europe. It describes eight community use cases which articulate consensus about the need for a common standard for machine-actionable DMPs (where machine actionable is defined as "information that is structured in a consistent way so that machines, or computers, can be programmed against the structure"[2])

The specific focus of this working group is on developing common information model and specifying access mechanisms that make DMPs machine-actionable. The outputs of this working group will help in making systems interoperable and will allow for automatic exchange, integration, and validation of information provided in DMPs, for example, by

[1] https://doi.org/10.3897/rio.3.e13086
[2] http://www.ddialliance.org/taxonomy/term/198

checking whether a provided PID links to an existing dataset, if hashes of files match to their provenance traces, or whether a license was specified. The common information models are NOT intended to be prescriptive templates or questionnaires, but to provide re-usable ways of representing machine-actionable information on themes covered by DMPs.

The vision that this working group will work to realise is one where DMPs are developed and maintained in such a way that they are fully integrated into the systems and workflows of the wider research data management environment. To achieve this vision we will **develop a common data model with a core set of elements**. Its modular design will allow customisations and extensions using existing standards and vocabularies to follow best practices developed in various research communities. We will **provide reference implementations** of the data model using popular formats, such as JSON, XML, RDF, etc. This will enable tools and systems involved in processing research data to read and write information to/from DMPs. For example, a workflow engine can add provenance information to the DMP, a file format characterization tool can supplement it with identified file formats, and a repository system can automatically pick suitable content types for submission and later automatically identify applicable preservation strategies.

The deliverables will be publicly available under CC0 license and will consist of models, software, and documentation. The documentation will describe functionality and semantics of terms used, rationale, standard compliant ways for customisation, and requirements for supporting systems to fully utilise the capabilities of the developed model.

The working group will be open to everyone and will involve all stakeholders representing the whole spectrum of entities involved in research data management, such as: researchers, tool providers, infrastructure operators, repository staff and managers, software developers, funders, policy makers, and research facilitators. We will take into account requirements of each group.This will likely speed up and increase adoption of the working group outcomes.

The group will predominantly collaborate online, but will use any possibility to meet in person during RDA plenaries, conferences, workshops, hackathons or other events in which their members participate. All meetings in which decisions are made will be documented and their summaries will be circulated using the RDA website.

The work will be performed iteratively and incrementally following the best practices from system and software engineering. We will evaluate preliminary drafts of the model with community to receive early feedback and to ensure that the developed common model is interoperable and exchangeable across implementations. We will also express existing DMPs using the developed common model and will investigate how to support modification of machine actionable DMPs by various tools involved in data management process, while ensuring that proper provenance and versioning information is stored with. Finally, we will build prototypes to investigate possible system integrations and to evaluate to which degree the information contained in the DMPs can be automatically validated and which actions or alerts depending on a DMP state can be triggered, e.g. by sending notifications to repositories or funder systems.

During our work we will monitor parallel efforts and engage with various research communities to find candidates for pilot studies and to transfer the acquired know-how. Towards the end of the lifetime of the working group we will launch pilot projects in which the model will be customised to suit the needs of the identified interested communities. Pilot studies will use the models to integrate systems and demonstrate how machine-actionable DMPs can work.

We believe that the outcomes delivered by this group will contribute to improving the quality of research data and research reproducibility, while at the same time reducing the administrative burden for researchers and systems administrators.

Value Proposition*: A specific description of who will benefit from the adoption or implementation of the WG outcomes and what tangible impacts should result.*

A common data model for machine-actionable DMPs will enable interoperability of systems and will facilitate automation of data collection and validation processes. The common model and accompanying interfaces and libraries are an essential building block for the infrastructure. Although for some stakeholder groups, the developments will be invisible (and should be) so that the unification and standardisation of a DMP model will bring benefits to all of them.

- **Researchers** will benefit from having fewer administrative procedures to follow. Machine-actionable DMPs can facilitate the automatic collection of metadata about experiments. They will accompany experiments from the beginning and will be updated over the course of the project. Consecutive tools used during processing can read and write data from machine-actionable DMPs. As a result, parts of the DMPs can be automatically generated and shared with other collaborators or funders. Furthermore, researchers whose data is reused in other experiments will gain recognition and credit because their data can be located, reused, and cited more easily.
- **Reusing parties** will gain trust and confidence that they can build on others' previous work because of a higher granularity of available information.
- **Funders** and **repositories** will be able to automatically validate DMPs. For example, they will be able to check whether the specified ORCID iD or e-mail are correct, whether the data is available at the specified repository, and whether the data checksums are correct – in other words, whether the information provided in a DMP reflects reality.
- **Infrastructure providers** will get a universal format for exchange of (meta-) data between the systems involved in data processing and data storage. They could also be able to automate processes associated to DMPs, like backup, storage provision, grant access permissions, etc.
- **Society** will be better able to safeguard investment made in research and will gain assurance that scientific findings are trustworthy and reproducible, while the underlying data is available and properly preserved.

Engagement with existing work in the area*: A brief review of related work and plan for engagement with any other activities in the area.*

The need for machine-actionable DMPs is recognized by the community and is being discussed within the Research Data Alliance. Participants of the CERN workshop organized in 2016 identified "encodings for exporting DMPs" as one of the next developments needed[1]. Automation and machine actionability are meant to be key factors enabling deployment of the European Open Science Cloud. A workshop on machine-actionable DMPs organized by the Digital Curation Centre and University of California Curation Center at the California Digital Library at IDCC in Edinburgh in 2017 resulted in a white paper[3] that describes the current state of the art and expresses a need for a common standard for machine-actionable DMPs.

As a result of these ongoing discussions the participants of the 9th plenary meeting in Barcelona during the Active DMPs IG session decided to establish specific working groups that address various identified challenges related to DMPs. The proposed group on DMP common standards will address a high-priority challenge based on the most recent assessments of community needs.

Members of the proposed group are well connected to various community-based initiatives and working groups that address similar topics. The group will monitor and align the efforts with others in this area. We will specifically monitor:

- RDA groups related to DMPs, such as, but not limited to:
    - Active DMPs IG,
    - Research Data Repository Interoperability WG,
    - Reproducibility IG,
    - e-Infrastructure IG,
    - RDA/WDS Certification of Digital Repositories IG,
    - BioSharing Registry: connecting data policies, standards & databases in life sciences WG,

[3] https://doi.org/10.3897/rio.3.e13086

    - Exposing DMPs WG (under review),
- tools, such as, but not limited to:
    - DMPTool,
    - DMPonline,
    - RDM Organiser,
- the DMP fora e.g. Force 11 FAIR DMP or Belmont Forum e-Infrastructures and Data Management Collaborative Research Action
- e-Infrastructure projects e.g. OpenAIRE, EUDAT, European Open Science Cloud (EOSC)
- W3C,
- and others.

Work Plan*: A specific and detailed description of how the WG will operate including:*

· *The form and description of final deliverables of the WG,*

D1. Common data model for machine-actionable DMPs

This deliverable will contain the developed data model and documentation describing semantics of terms used, rationale, and standard compliant ways for customisation of the model.

D2. Reference implementations

Reference implementation of the common data model will provide ready to use models in popular standards such as JSON, XML, RDF, etc. It will also provide example models of DMPs in each format.

D3. Guidelines for adoption of the common data model

Guidelines will be based on lessons learned from the common model development and prototyping. They will describe requirements for supporting systems to fully utilise the capabilities of the common data model.

- *The form and description of milestones and intermediate documents, code or other deliverables that will be developed during the course of the WG’s work,*

M1. Requirements and candidate solutions reviewed (M5)

We will analyse existing DMP tools, as well as tools from domains of digital preservation, reproducible research, open science, and data repositories that cover the full data lifecycle. We will look for mappings to popular DMP creation tools, such as checklists, discuss lessons learned, and identify limits of automation and machine actionability. We will also investigate modelling techniques used in model engineering and linked data domains to identify suitable notation and tools for the common model. Furthermore, we will identify and analyse existing domain-specific standards and evaluate their applicability. Based on this research, we will define requirements for the common model and identify domain-specific models and controlled vocabularies that need to interoperate with the common data model.

M2. Common model specification drafted (M10)

We will design a common data model and example expressions in mainstream representation formats (e.g. JSON). The development will be iterative and based on both real and synthetic examples of DMPs. We will develop prototypes to demonstrate how the model works and what its capabilities are.

M3. Common model refined (M15)

We will develop further extensions to the core model (the model will likely be modular) to evaluate its scalability and customisability. Furthermore, we will test integrations with existing tools and continue evaluation using sample DMPs. Based on these activities we will introduce necessary refinements to the common data model.

M4. Dissemination and pilot studies (M18)

We will formulate guidelines for the adoption of the common model and release final documentation of the developed model and reference implementations. We will disseminate the results of our work through mailing lists, participation in conferences, as well as social media. We will launch pilot studies that implement the working group outcomes. We will facilitate and encourage crowd-sourced descriptions of implementations beyond the direct activities of the working group.

- *A description of the WG's mode and frequency of operation (e.g. on-line and/or on-site, how frequently will the group meet, etc.),*

The group will predominantly collaborate online, but will use any opportunity to meet in person during RDA plenaries, conferences, workshops, hackathons or other events in which their members participate. All meetings in which decisions are made will be documented and their summaries will be circulated using the RDA website.

The group will have regular monthly calls to report on progress and discuss open issues. We will also use GitHub to host developed models and source code. We will use issue tracking mechanisms to discuss enhancements, bugs, and other issues. Important updates, such as reaching a milestone, will be communicated through the RDA website.

- *A description of how the WG plans to develop consensus, address conflicts, stay on track and within scope, and move forward during operation, and*

Group consensus will be achieved primarily through mailing list discussions, where opposing views will be openly discussed and debated amongst members of the group. If consensus cannot be achieved in this manner, the group co-chairs will make the final decision on how to proceed.

The co-chairs will keep the working group on track by setting milestones and reviewing progress relative to these targets. Similarly, scope will be maintained by tying milestones to specific dates, and ensuring that group work does not fall outside the bounds of the milestones or the scope of the working group.

- *A description of the WG's planned approach to broader community engagement and participation.*

The working group case statement will be disseminated to mailing lists in communities of practice related to research data and repositories (e.g. ICSU World Data System) in an effort to cast a wide net and attract a diverse, multi-disciplinary membership. Group activities, where appropriate, will also be published to related mailing lists and online forums to encourage broad community participation.

Adoption Plan*: A specific plan for adoption or implementation of the WG outcomes within the organizations and institutions represented by WG members, as well as plans for adoption more broadly within the community. Such adoption or implementation should start within the 18 month timeframe before the WG is complete.*

Representatives of various stakeholders groups who are prominent in the area of DMPs have already joined this working group, including:

- DMPRoadmap
- DMPonline / Digital Curation Centre
- DMPTool / California Digital Library
- ELIXIR data stewardship wizard
- RDM Organiser
- Islandora
- Phaidra
- Open Science Framework
- Data Intensive Research Initiative of South Africa (DIRISA)
- Belmont Forum e-Infrastructure and Data Management
- DSA-WDS Core Trustworthy Data Repositories
- DMP OPIDoR
- INESC-ID
- INIA - Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria
- EDINA

- Data Archiving and Networked Services (DANS)

These representatives have agreed to consider implementing the standards recommended by the working group in their respective tools. Some of them have already committed to active participation in the group and plan to adopt the outputs. We will continue to seek representatives from a variety of research communities to ensure that this working group's deliverables are widely adopted.

Initial Membership*: A specific list of initial members of the WG and a description of initial leadership of the WG.*

Leadership:

- Chair: Tomasz Miksa (SBA Research, Austria)
- Co-chair: Paul Walk (University of Edinburgh, Great Britain)
- Co-chair: Peter Neish (University of Melbourne, Australia)

Members/Interested (based on 9^{th} Plenary volunteer list and subsequent calls):

- Adil Hasan
- Amir Aryani
- Andreas Rauber
- Andrew Janke
- Anna Dabrowski
- Antonio Sánchez-Padial
- Christoph Becker
- Cristina Ribeiro
- Daniel Mietchen
- Dessi Kirilova
- Fernando Aguilar
- Heike Görzig
- Janez Štebe
- Jens Ludwig

- Jérôme Perez
- Joao Aguiar Castro
- João Cardoso
- Jonathan Petters
- Karsten Kryger Hansen
- Lesley Wyborn
- Madison Langseth
- Marie-Christine Jacquemot-Perbal
- Mark Leggott
- Mustapha Mokrane
- Myriam Mertens
- Natalie Meyers
- Nobubele Shozi
- Paolo Budroni
- Peter Doorn
- Peter McQuilton
- Peter Neish
- Raman Ganguly
- Rob Hooft
- Sarah Jones
- Stephanie Simms
- Terry Longstreth
- Timea Biro
- Wim Hugo

---

[1] CERN workshop on Active DMPs: indico.cern.ch/event/520120/attachments/1302179/2036378/CERN-ADMP-iPRES206.pdf