Framing the scope of the common data model for machine-actionable Data Management Plans

Tomasz Miksa (tmiksa@sba-research.org)
João Cardoso (joao.m.f.Cardoso@tecnico.ulisboa.pt)
José Borbinha (jlb@tecnico.ulisboa.pt)
1. Introduction

1.1. Motivation (1/2)

- **Technology advances** have led to **ever growing** quantities of **data**.

- Researchers have to **manage**, **preserve** and **publish** their **data efficiently**.

- Data must be **reproduceable** and **reusable**.

- However **data processing** in such large quantities, is a **complex task**.
1. Introduction

1.1. Motivation (2/2)

- The **Data Management Plan (DMP)** is one of the tools available for **Research Data Management (RDM)**.

- The DMP is a **document** describing techniques, methods and policies to be **applied to the data** of a project.

- Some **funding bodies** already **require** that funding **applications** be **accompanied** by a **DMP**.

- However, **researchers do not** have enough **expertise**, and **time**, to **prepare a DMP** and implement it.

- Researchers need **aid** to **plan and manage** data in an **automated and scalable** way.
1. Introduction

1.2. Objective

• The Research Data Alliance (RDA) DMP Common Standards Working Group aims to implement machine-actionable DMPs (maDMPs)

• Improve information exchange in research tools and systems, and embed DMPs in existing workflows.

• Parts of the DMP can be automatically generated and shared.

• To achieve this it is necessary to know existing research data workflows, RDM infrastructure and to have a common data model for maDMPs.
1. Introduction

1.3. Outline

- Fundamental concepts

- DMP Common Standards Working Group

- Consultations

- Proof of Concept Tools

- Looking Ahead
2. Fundamentals

2.1. Research Data Management (RDM)

- **Research Data Management** is one of the approaches to manage, preserve and publish data.

**Definition of RDM:**

“concerns the organization of data, from its entry to the research cycle through to the dissemination and archiving of valuable results. It aims to ensure reliable verification of results and allows for new and innovative research built on existing information.”

2011, Whyte et al.
2. Fundamentals

2.1. What is a Data Management Plan (DMP)

- A **DMP** is a document used to **support data management**.
- That implies detailing its:
  - Identity
  - Organization
  - Documentation
  - Data Quality
  - Preservation
  - Policies
  - Dissemination,
  - Etc...

- Currently most DMPs are **manually created** by researchers using **checklists** and **online tools**, such as **DMP Tool** or **DMP Online**.

https://dmptool.org/
https://dataone.org
https://dmponline.dcc.ac.uk/
2. Fundamentals

2.2. DMP Drawbacks

• In effect, currently a DMP is a mostly **static** document.

• The **information** is often **broad** and **unclear**.

• Ideally a DMP **should be dynamic** in nature, and a **fundamental part** of **research**.

• A **human-readable** description is always **needed**.

• It is however often **considered** a **bureaucratic hassle**.

2. Fundamentals

2.3. Why a Machine-Actionable DMP (maDMP)?

- The maDMP concept was introduced to extend the concept of DMP.

- The maDMP aims to have DMPs be:
  - Machine and Human Readable descriptions
  - Enable Automated Policy Enforcement
  - Sharable
  - Common Data Model
  - Allow for the integration of systems and tools
  - Etc...

Current DMPs

```
<admindata>
  <question>Who is the Principle Investigator?</question>
  <answer>The PI is John Doe from the JDU</answer>
</admindata>
```

maDMPs

```
“dc:creator”:{{
  “foaf:name”:”John Doe”,
  “@id”:”orcid.org/000-1111-2222-3333”,
  “foaf:mbox”:”mailto:jdoe@jdu.edu”,
  “madmp:institution”:”JDU-John_Doe_University”
}}
```

Reuse of existing standards
Use of persistent identifiers
Use of controlled vocabularies
2. Fundamentals

2.3. Why a Machine-Actionable DMP (maDMP)?

- maDMPs can bring **benefits** in **activities** such as:
  - Data Discovery
  - Data Reuse
  - Data Evaluation
  - DMP Creation
  - Data Exchange
  - Data Monitoring
  - Policy Enforcement
  - Workflow Integration
The RDA DMP Common Standards Working Group was created to focus on the standardization of the knowledge contained in a DMP.

- Its objective is to establish a common data model that defines a core set of elements for a DMP.

- The data model is to have a modular design and resort to semantic technologies for representation.

https://www.rd-alliance.org/groups/dmp-common-standards-wg

https://www.rd-alliance.org/groups/active-data-management-plans.html
3. DMP Common Standards Working Group

3.1. How to achieve the objective?

- To achieve this it was first necessary to establish the requirements for a maDMP.

- There is no clear and common definition of an maDMP nor what information should it contain.

- Our solution was to perform user consultations to try to clarify the scope of the common data model.
4. Consultations

4.1. First consultation

- The **first user consultation went broad**, and aimed at answering the following questions:
  - **Who** are the **stakeholders** at each lifecycle stage?
  - **How available information changes** over the lifetime of a DMP?
  - **How need for information changes** over the lifetime of a DMP?

- **User stories** approach

- The **tool** chosen to conduct the user consultation was **GitHub**
4. Consultations

4.1. First consultation

- Stories were classified into **three** generic categories.

- **Colors** were also used to further classification.

- Overall the **granularity** of the gathered information proved to be too high for meaningful analysis.

https://github.com/RDA-DMP-Common/user-stories
4. Consultations

4.2. Second consultation

• The second user consultation aimed at narrowing down the discussion.
  • Specific fields and standards in use
  • Models for specific requirements
  • The target audience were stakeholders with expertise in research data lifecycle

• There were two exercises:
  • Narrow down the core set of elements from sets of high level requirements


Look forward for the TPDL 2019 in OSLO!!
Maybe the 4th CAS workshop could be held there?
4. Consultations

4.2.1. Second consultation – First Exercise

- Review a series of use cases displaying workflows that could be automated in a maDMP ecosystem.

- There were 9 workflows available for review
4. Consultations

4.2.2. Second consultation – Second exercise

• Participants were assigned with documents containing sets of high level requirements.

• They were asked to review the documents and focus on three points:
  • Further specifying existing requirements
  • Suggest models, vocabularies and other sources that could be reused into the DMP common data model
  • Examples and justification for any suggested field.
5. Proof of Concept Tools

5.1. How to prove the maDMP concept?

- To prove that RDM can be automated, three tools that read/write DMPs were developed.

- Students from the TU Wien course on Digital Preservation developed prototypes of automated DMP Generator Tools.

- The tools had could focus on two scenarios:
  - DMP creation on the onset of a research project
  - DMP creation or update during the course, or at the close of a research project.
5. Proof of Concept Tools

5.2. First scenario
5. Proof of Concept Tools

5.2.1. First scenario – Tool interaction

5. Proof of Concept Tools

5.2.2. First scenario – Created DMP

Data management plan for project **test**

**Author:** Projektass. Dr.techn. Mag. **Tomasz Miksa**

**Contact information:**

- **Email:** tomasz.miksa@tuwien.ac.at
- **Institute:** Forschungsbereich Information und Software Engineering
- **Room:** unknown
- **Website:** http://www.ifs.tuwien.ac.at/~miksa/

**Sample Files:**

- **Name:** Agenda.txt
  - **Type:** input Mime-Type: text/plain; charset=US-ASCII
  - **Size:** 764
  - **Number of similar files:** 37

- **Name:** Workflows for maDMPs.pdf
  - **Type:** input Mime-Type: application/pdf
  - **Size:** 2931191
  - **Number of similar files:** 57712

- **Name:** datalifecycle.png
  - **Type:** input Mime-Type: application/octet-stream
  - **Size:** 98765
  - **Number of similar files:** 1904

**License:**

**Repository:** Access to Research and Communications Annals
http://arca.igc.gulbenkian.pt/

---

### Machine-actionable DMP

```json
{
  "project": "test",
  "author": {
    "firstName": "Tomasz",
    "lastName": "Miksa",
    "preceedingTitles": "Projektass. Dr.techn. Mag.",
    "postpositionedTitles": "",
    "mailEmail": "tomasz.miksa@tuwien.ac.at",
    "employee": {
      "organizationUnit": {
        "value": "Forschungsbereich Information und Software Engineering"
      },
      "roomCode": "null",
      "room": {
        "roomCode": "unknown"
      },
      "website": {
        "website": {
          "value": "http://www.ifs.tuwien.ac.at/miksa/
        }
      },
    },
    "repository": {
      "id": 1989,
      "name": "Access to Research and Communications Annals",
      "run": "http://arca.igc.gulbenkian.pt/",
      "delete": "http://arca.igc.gulbenkian.pt/elestended/request"
    },
    "license": false,
    "files": {
      "name": "Agenda.txt",
      "mimeType": "text/plain; charset=US-ASCII",
      "size": 764,
      "type": "input",
      "number": 37
    },
    "name": "Workflows for maDMPs.pdf",
    "mimeType": "application/pdf",
    "size": 2931191,
    "type": "input",
    "number": 57712
  }
}
```
5. Proof of Concept Tools

5.3. Second scenario

- BASIC INFO
- ADMINISTRATIVE DATA
- GET METADATA
- PRESERVATION
- GENERATE DMP

- GitHub
- OAI-PMH

John Smith

10 YEARS
5. Proof of Concept Tools

5.3.1. Second scenario – Tool interaction

5. Proof of Concept Tools

5.3.2. Second scenario – Created DMP

Correlating Alcohol Consumption and UFO Sightings in the USA

Authors

Ethical Questions

Licenses and Redistribution

Code Preservation

Data Preservation

Access and Security

Data Sharing

Github Repository

Zmende Repository

Data Usage and Project

Responsibility for Data Management

Resources

Human readable version of the DMP

Machine-actionable version of the DMP
6. Looking ahead

- To **turn the knowledge gathered** through both the consultations and **proof of concept tools** into a **common data model** for an maDMP

- A final **third consultation** was performed at the **12th RDA Plenary** in Gaborone in Botswana.
  - Set **classes** and **properties**
  - **DMP components** as **extensions** of the core data model
  - **Guidelines to customize** and **implement** the DMP common data model
  - **Full results still under analysis**

- **First draft** to be presented at **13th RDA Plenary** Meeting in **Philadelphia (2-4 April 2019)**
• **E-ARK4ALL** is a consortium working with the EC’s **Connecting Europe Facility (CEF)** to develop the eArchiving Building Block

• **eArchiving** will provide **core specifications, software, training** and **knowledge** to tackle the challenge of data management and reuse.

Take one of my leaflets to know more!!

http://ec.europa.eu/cefdigital

http://e-ark4all.eu/
THANK YOU!