From Requirements to Models

DMP Common Standards WG

Tomasz Miksa, Paul Walk, Peter Neish
Agenda

- Part 1 – Introduction for newcomers
- Part 2 – Status update - open consultations
- Part 3 – Lightning talks from DMP tool owners
- Part 4 – Discussion
- Part 5 – Wrap-up and next steps
Collaborative note taking & attendance

https://goo.gl/crvjh1

https://guidebook.com/g/rdaplenary11
Part 1

Introduction for newcomers
Challenge – why do we need this WG?

- Shortcomings of existing DMPs
  - manually completed, vague, not updated, ...

- Machine-actionable DMPs
  - living documents
  - automate data management
    - collect information from systems
    - trigger actions in systems
  - facilitate validation

- This requires
  - well-defined RDM workflows
  - data management infrastructure
  - common data model

https://doi.org/10.3897/rio.3.e13086
Common data model for machine-actionable DMPs
- to model information from standard DMPs
- NOT a template
- NOT a questionnaire
- modular design
  - core set of elements
  - domain specific extensions

Reference implementations
- ready to use models
  - JSON, XML, RDF, etc.

Guidelines for adoption of the common data model
- requirements for supporting systems
- pilot studies
Example

• Current DMPs – model questionnaires

<administrative_data>
  <question>Who will be the Principle Investigator?</question>
  <answer>The PI will be John Smith from our university.</answer>
</administrative_data>

• Machine-actionable DMPs – model information

"dc:creator":[
  {
    "foaf:name":"John Smith",
    "@id":"orcid.org/0000-1111-2222-3333",
    "foaf:mbox":"mailto:jsmith@tuwien.ac.at",
    "madmp:institution":"AT-Vienna-University-of-Technology"
  }
]
Example

• Currently available – not very useful

<administrative_data>
  <question>Who will be the Principle Investigator?</question>
  <answer>The PI will be John Smith from our university.</answer>
</administrative_data>

Machine-actionable DMP

"dc:creator": [ {
  "foaf:name":"John Smith",
  "@id":"orcid.org/0000-1111-2222-3333",
  "foaf:mbox":"mailto:jsmith@tuwien.ac.at",
  "madmp:institution":"AT-Vienna-University-of-Technology"
} ],

Reuse existing standards, e.g. Dublin Core, PREMIS, etc.
Example

• Currently available – not very useful

<administrative_data>
  <question>Who will be the Principle Investigator?</question>
  <answer>The PI will be John Smith from our university.</answer>
</administrative_data>

• Machine-actionable DMP

"dc:creator": [ {
  "foaf:name": "John Smith",
  "@id": "orcid.org/0000-1111-2222-3333",
  "foaf:mbox": "mailto:jsmith@tuwien.ac.at",
  "madmp:institution": "AT-Vienna-University-of-Technology"
} ]

Use PIDs whenever possible, e.g. ORCID
Example

• Currently available – not very useful

<administrative_data>
  <question>Who will be the Principle Investigator?</question>
  <answer>The PI will be John Smith from our university.</answer>
</administrative_data>

• Machine-actionable DMP

"dc:creator":[ {
  "foaf:name":"John Smith",
  "@id":"orcid.org/0000-1111-2222-3333",
  "foaf:mbox":"mailto:jsmith@tuwien.ac.at",
  "madmp:institution":"AT-Vienna-University-of-Technology"
},

Use controlled vocabularies
Example

• Currently available – not very useful

<administrative_data>
  <question>Who will be the Principle Investigator?</question>
  <answer>The PI will be John Smith from our university.</answer>
</administrative_data>

• Machine-actionable DMP

"dc:creator":[
  {
    "foaf:name":"John Smith",
    "@id":"orcid.org/0000-1111-2222-3333",
    "foaf:mbox":"mailto:jsmith@tuwien.ac.at",
    "madmp:institution":"AT-Vienna-University-of-Technology"
  }
],

Develop own concepts and vocabularies only when needed
Part 2
Open consultations
User story consultation

Goals

- identify stakeholders at each lifecycle stage
  - define which information they provide
  - define which information they expect

As a <stakeholder>, I want <goal> so that <reason>.

As a researcher, I want to inform repository operator on the amount of data in the planning phase, so that they provide information on costs.

https://github.com/RDA-DMP-Common/user-stories/
User story consultation

- https://github.com/RDA-DMP-Common/user-stories/
- 100+ issues defined
- inputs from Europe and Australia
- inputs from individuals and workshops
User story labelling

- https://github.com/RDA-DMP-Common/user-stories/projects/2
- Reviewed by chairs and authors
  - classified
    - in scope - useful for model definition
    - out of scope – often referring to the ecosystem, practices – important but not directly for the common data model
  - labelled

21/03/2018
User story labelling

- [https://github.com/RDA-DMP-Common/user-stories/wiki](https://github.com/RDA-DMP-Common/user-stories/wiki)

- 3 major categories (colours)
  - stakeholders involved
  - project phase
  - subject of information conveyed
    - access control
    - volume
    - financial
    - licensing
    - metadata
    - repository
    - security
    - storage
    - etc.
User story visualisation

- [link](https://goo.gl/znBL3F)
- interactive visualisation - changes on GitHub are visible immediately
- shows relations between stakeholders, phases and information
From user stories to requirements

- https://docs.google.com/document/d/1sWVy0Rqj9fGjs6GyFnBd3fH6XF2088zjK8U-1wLq4c/edit?usp=sharing
- Refactoring of user stories
- Goal: finding overlaps, gaps, duplicates
- Example below

**Metadata**
- taxonomy/classification [14,11]
- Links to metadata of the real data [89, 39]
- Funder information [7]
- Link publications to data [55]
- Authorship [88]
- Multilingual metadata [65]
- Include raw metadata directly in the model [91, 85]
From user stories to requirements

- https://docs.google.com/document/d/1sWVy0Rqj9fGsjs6GyFnBd3fH6XF2088zjK8U-1wLq4c/edit?usp=sharing

- Refactoring of user stories
  - Goal: finding overlaps, gaps, duplicates
  - Example below

Metadata

- taxonomy/classification [14,11]
- Links to metadata of the real data [89, 39]
- Funder information [7]
- Link publications to data [55]
- Authorship [88]
- Multilingual metadata [65]
- Include raw metadata directly in the model [91, 85]
From user stories to requirements

- https://docs.google.com/document/d/1sWVy0Rqj9fGsjs6GyFnBd3fH6XF2088zjkK8U-1wLq4c/edit?usp=sharing

Refactoring of user stories
- Goal: finding overlaps, gaps, duplicates
- Example below

Metadata
  - taxonomy/classification [14,11]
  - Links to metadata of the real data [89, 39]
  - Funder information [7]
  - Link publications to data [55]
  - Authorship [88]
  - Multilingual metadata [65]
  - Include raw metadata directly in the model [91, 85]

short summary of what user stories are about – more specific requirements

21/03/2018
www.rd-alliance.org - @resdatall
From user stories to requirements

▶ https://docs.google.com/document/d/1sWVv0Rqj9fGsjs6GyFnBd3fH6XF2088zjK8U-1wLq4c/edit?usp=sharing

▶ Refactoring of user stories
  ▶ Goal: finding overlaps, gaps, duplicates
  ▶ Example below

▶ Metadata
  ▶ taxonomy/classification [14,11]
  ▶ Links to metadata of the real data [89, 39]
  ▶ Funder information [7]
  ▶ Link publications to data [55]
  ▶ Authorship [88]
  ▶ Multilingual metadata [65]
  ▶ Include raw metadata directly in the model [91, 85]
Requirements grouping

- Similar requirements exist under different labels
- Example
  - Information on the author of the DMP is relevant for
    - Administrative activities
    - Reuse
- We split requirements and grouped them using five categories
  - Administrative, Roles and Responsibilities
  - Data
  - Infrastructure
  - Security, Privacy and Access Control
  - Policies, legal and ethical aspects
Requirements grouping example (Data)

DATA

Format
- Format [80, 12, 99, 62, 67, 54, 80]

Volume
- Data size estimate [5, 77, 80, 100]
  - For specific type of data [62]
- Data size real [54]

Provenance [54]

Metadata
- taxonomy/classification [14, 11]
- Links to metadata of the real data [89, 39]
- Link publications to data [55]
- Authorship [88]
- Multilingual metadata [65]
- Include raw metadata directly in the model [91, 85]

Reuse
- Links to (meta-)data location [89, 90, 56, 39, 60]

Repository [42]
- Persistent identifier for data [92]
- Link publications to data [55, 88]
- Link to License/Contract allowing data usage/storing [56]

Note: we did not move all requirements falling under a specific label, but only a subset that is relevant in this context – in the given example, relevant for data description. Other requirements for Reuse were put into other categories.
Next steps

1\textsuperscript{st} consultation (user stories) went broad

- helped us define the scope of the maDMPs
  - what information should a maDMP contain?
  - who provides and uses this information?

2\textsuperscript{nd} consultation will go deep

- how do we model specific requirements
  - which specific fields are needed?
  - which models exist?
Consultation 2 – ‘going deep’

- https://goo.gl/DRieP4

- 5 documents to collect requirements, models, specific fields, etc.
  - Administrative, Roles and Responsibilities
  - Data
  - Infrastructure
  - Security, Privacy and Access Control
  - Policies, legal and ethical aspects
Consultation 2 – ‘going deep’

Goal: reach out to experts in each category to learn
  ▶ which concrete information (specific fields)
  ▶ in what form they expect

Next
  ▶ compare collected models and fields
  ▶ select best fitting
  ▶ design the architecture of the model (core model with extensions vs flat model, serializations, etc.)

Timeline:
  ▶ end of May 2018
Part 3

Lightning talks
<table>
<thead>
<tr>
<th>Slot #</th>
<th>DMP software</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>Data Stewardship Wizard</td>
<td>in-person</td>
</tr>
<tr>
<td>#2</td>
<td>DMP Service (OpenAIRE)</td>
<td>in-person</td>
</tr>
<tr>
<td>#3</td>
<td>RDMOrganiser</td>
<td>in-person</td>
</tr>
<tr>
<td>#4</td>
<td>DMPRoadmap (covers DMPonline and DMPtool from DCC and UC3)</td>
<td>in-person</td>
</tr>
<tr>
<td>#5</td>
<td>ReDBox (QCIF)</td>
<td>remote</td>
</tr>
<tr>
<td>#6</td>
<td>UQ Research Data Manager (UQRDM)</td>
<td>remote</td>
</tr>
</tbody>
</table>
The incredibly shrinking DMP?
Hierarchical

3/2 Is special care needed to get the raw data ready for processing?

3/6 Data exchange format has been settled before?

3/7 Will the raw data be transported on physical media, or via the network?

3/10 Is data integrity guaranteed during this stage?

3/11 Is data security guaranteed in this phase?

Is capacity planning needed? For the measurement? For the data transport?
EXTENSIBLE
When is the raw data archived?

- As soon as it comes in, in chunks
- As soon as it has all arrived, in one session
- All at once with the results at project end
<table>
<thead>
<tr>
<th>Slot #</th>
<th>DMP software</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>Data Stewardship Wizard</td>
<td>in-person</td>
</tr>
<tr>
<td>#2</td>
<td>DMP Service (OpenAIRE)</td>
<td>in-person</td>
</tr>
<tr>
<td>#3</td>
<td>RDMOrganiser</td>
<td>in-person</td>
</tr>
<tr>
<td>#4</td>
<td>DMPRoadmap (covers DMPonline and DMPtool from DCC and UC3)</td>
<td>in-person</td>
</tr>
<tr>
<td>#5</td>
<td>ReDBox (QCIF)</td>
<td>remote</td>
</tr>
<tr>
<td>#6</td>
<td>UQ Research Data Manager (UQRDM)</td>
<td>remote</td>
</tr>
</tbody>
</table>
Research Data Management Organiser

rdmorganiser.github.io

@rdmorganiser
Why RDMO?

Organizer instead of plan

- Organize the data management over the whole lifetime of a project and beyond
- Engage all stakeholders

Local instead of central

- Full customization of the content
  - disciplinary context
  - institutional environment
- Easily deployable for universities or infrastructure providers
- Embed into existing infrastructure

Image credit: M.Minderhoud, Agenda, CC BY-SA 3.0
RDMO Architecture

Input
- Questions
  - Catalogs
  - Sections
  - Subsections
  - Question sets
  - Questions

Domain
- Attributes
- Entities

Options
- Option sets
- Options

Conditions

Output
- Views
- Tasks

User
- Projects
- Snapshots
- Values

Datamodel
Attribute: project/dataset/ipr/copyrights

Condition: project/legal-aspects/ipr/yesno == 1

Option set

Option: work of art

Question: Does copyright law apply to this dataset?

Question set

Entity: project/dataset/ipr

Attribute: project/legal-aspects/ipr/yesno

Question: Does the project use data that is protected by intellectual property rights?

Question set
Examples for attributes and entities

- project/coordination/name
- project/costs/
- project/dataset/
- project/dataset/description
- project/dataset/data_security/
- project/dataset/ipr/
- project/dataset/metadata/
- project/dataset/pids/
- project/dataset/preservation/
- project/dataset/sensitive_data/
- project/dataset/sharing/
- project/dataset/size/
- project/dataset/storage/
- project/dataset/versioning_tool/
- project/funder/
- project/legal_aspects/
- project/partner/
- project/preservation/
- project/schedule/
- project/support/
<catalog xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:rdmo="https://rdmorganiser.github.io/terms/questions/rdmo">
  <dc:comment/>
  <order>1</order>
  <title lang="en">RDMO</title>
  <title lang="de">RDMO</title>
  <sections>
    <section>
      <dc:comment/>
      <order>0</order>
      <title lang="en">General</title>
      <title lang="de">Allgemein</title>
      <subsections>
        <subsection>
          <dc:comment/>
          <order>0</order>
        ...
        ...
      </subsections>
    </section>
  </sections>
</catalog>
Programmable JSON API

curl -X GET -H 'Authorization: Token oojoh3phaighaebiNeiyeeCeiY3Peuv2eitoojoh' \
https://rdmo.aip.de/api/v1/projects/values/?attribute__path=project/dataset/size/volume

[
  {
    "id":10061,
    "project":"https://rdmo.aip.de/api/v1/projects/projects/69/",
    "attribute":"https://rdmo.aip.de/api/v1/domain/attributes/262/",
    "set_index":0,
    "collection_index":0,
    "text":"
    "option":null,
    "created":"2017-05-29T14:50:20.009917Z",
    "updated":"2017-05-29T14:50:20.009924Z"
  },
  ...
]
Resources

Website: rdmorganiser.github.io
GitHub organisation: github.com/rdmorganiser
RDMO source code: github.com/rdmorganiser/rdmo
RDMO questionnaire: github.com/rdmorganiser/rdmo-catalog
Documentation: rdmo.readthedocs.io
Demo instance: rdmo.aip.de
Mailinglist: rdmo@listserv.dfn.de
Twitter: @rdmorganiser
Slack: rdmo.slack.com
GitHub issues: github.com/rdmorganiser/rdmo/issues
Supplemental material cloud.aip.de/index.php/s/dmRTnb1lqS3Es
<table>
<thead>
<tr>
<th>Slot #</th>
<th>DMP software</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>Data Stewardship Wizard</td>
<td>in-person</td>
</tr>
<tr>
<td>#2</td>
<td>DMP Service (OpenAIRE)</td>
<td>in-person</td>
</tr>
<tr>
<td>#3</td>
<td>RDMOrganiser</td>
<td>in-person</td>
</tr>
<tr>
<td>#4</td>
<td>DMPRoadmap (covers DMPonline and DMPtool from DCC and UC3)</td>
<td>in-person</td>
</tr>
<tr>
<td>#5</td>
<td>ReDBox (QCIF)</td>
<td>remote</td>
</tr>
<tr>
<td>#6</td>
<td>UQ Research Data Manager (UQRDM)</td>
<td>remote</td>
</tr>
</tbody>
</table>
DMPRoadmap Data Model

th Jimmy Angelakos
RDA 11  Plenary – Berlin 21/03/2018
DMPRoadmap users:
- DMPonline (DCC, UK)
- DMPTool (UC3, USA)
- DMP Assistant (Portage, Canada)
- DMP OPIDoR (France)
- DMPTuuli (Finland)
- DMPMelbourne (U of Melbourne)
- and more: https://github.com/DMPRoadmap/roadmap/wiki/Local-installations-inventory
Not as complicated as it looks!
roadmap

“Main axis” closeup
roadmap

Questions closeup

```
question_formats --1→ questions

questions --1→ answers

answers --n→ users

notes --1→ answers

question_options --n→ answers
```
Part 4

Discussion
Part 5

Wrap-up and next steps
Contributions needed!

Champions to lead requirements and model collection in each area:

- Administrative, Roles and Responsibilities
- Data
- Infrastructure
- Security, Privacy and Access Control
- Policies, legal and ethical aspects
Timeline

Our main focus: 12\textsuperscript{th} Plenary Meeting in Botswana

5-8 November 2018
Working Groups Complementary Activity

- **Q1 2018**: Exposing DMPs
- **Q3 2019**: DMP Common Standards

- **Q4 2017**: DMP Common Standards
- **Q2 2019**: Reference models

**Current practices**

**Processes**

**Requirements**

**Structures**
Staying in touch!

► Sign up to the group
  ► https://www.rd-alliance.org/groups/dmp-common-standards-wg

► Participate in the consultation

► Contact group chairs