Facilitating FAIRness with a Stockholm University machine-actionable DMP Online template

Joakim Philipson - Research Data Analyst

(Stockholm University, RDM team)

Focus on Open Science 2021-06-18

National requirement

- As from 2019, all grantees from Swedish Research Council must have a **data management plan** if the research generates research data.

- *Swedish Research Council* (**VR**) & Association of Swedish Higher Education Institutions (**SUHF**)
  collaborated on partially reworked version of **Science Europe**'s Core Requirements for Data Management Plans –
  now a template in the **DMP Online** tool – *but only free text fields*

- **Stockholm University** offers DMP Online tool as default to all our researchers at dmp.su.se via SSO local login

- Other DMP tools: ARGOS, DSW, DMPTool , easyDMP, RDMO -
  … all more or less converging to
  RDA Common Standard to produce more *machine-actionable, maDMPs*

Stockholm University
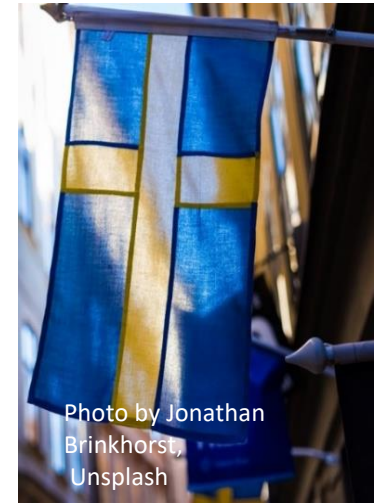
Multiple purpose of SU maDMP template

1. *Easing the administrative burden on researchers* to fill in the DMP by means of preset multiple-choice answer options, drop-down menus, radio buttons etc., including extensive local *Guidance*

2. *Facilitating* semi-automated *review* and *evaluation of FAIRness,* while retrieving needed information for RDM administration

3. *Conforming to the RDA DMP Common Standard* for machine-actionable DMPs by validation against *maDMP-schema-1.1*

Stockholm University

Features of SU maDMP template

- *Preserves structure* (sections I-VI) *and original questions from* Science Europe – *Swedish Research Council* (**VR**) in headings, but with more specific questions and answers (checkboxes, drop-down menus, radio buttons: yes/no/unknown)
- Using DMP Online (Roadmap) API (v0) to get output (as JSON) of entire DMPs (questions and answers) for further processing

- Workflow:
  1. APIv0 curl GET *https://dmp.su.se/plans/72835* ->

```
$ curl -G -k "https://dmponline.dcc.ac.uk/api/v0/plans?plan=72835" -H
"Authorization:Token token=XYZ123456789sTuV012345" > DMP72835.json
```

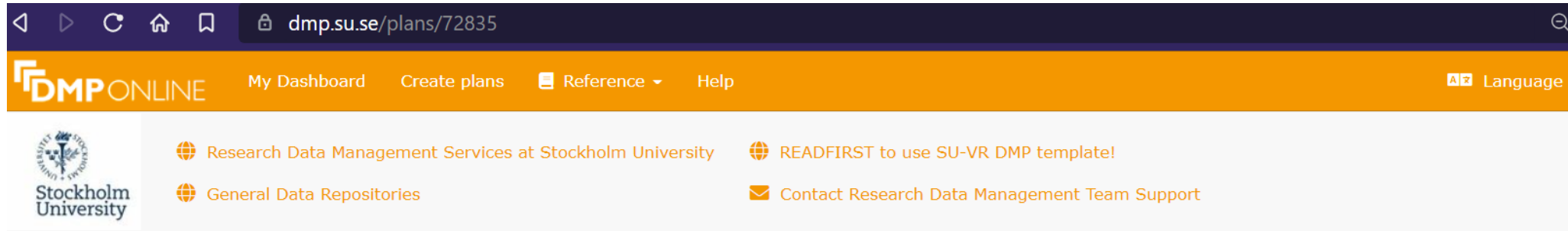 2. Oxygen: *dmp72385.json**2**dmp72385.xml* conversion ->
  3. *dmp72385.xml**2**dmp72385**transform**.xml* with XSLT
       [allowing FAIR-evaluation] ->
  4. *dmp72385transform.**xml**2**dmp72385transform.**json** ->
  5. *dmp72385transform.**json** validated with RDA maDMP

Test example DMP of SU-VR template

Test example DMP of SU-VR template

Demo DMP example of SU-VR template: section 0

Testing SU maDMP model for Climate Data: Sharing the seas.

| Project Details | Contributors | Plan overview | Initial DMP | Full DMP (undevelopped) | Share | Download |

expand all | collapse all   33/33 answered

0: Note on personal data! (1 / 1)   —

Please sign the following declaration before proceeding with your DMP!

"As a SU user of DMP Online I am aware that, except for information about project members such as PI and contact person, *no personal data* - i.e. any information relating to an identified or identifiable natural person (a '*data subject*', that is a *living* person), should be registered *in this DMP*. If necessary, to avoid this, any actual data registered here (even only as examples) will be completely anonymised before registration."

Q1: I have read and understood the above declaration and hereby certify that this DMP contains no personal data except for information about project members such as PI and contact person.

◉ Yes
○ No

Save

Answered 3 months ago by joakim.philipson@su.se

Comments

Add comments to share with collaborators

B *I* ≔ ▾ ≔ ▾ 🔗 ⊞ ▾

Save

Stockholm University

Testing SU maDMP model for Climate Data: Sharing the seas.

Demo DMP example of SU-VR template: I-Q1

Project Details | Contributors | Plan overview | Initial DMP | Full DMP (undevelopped) | Share | Download

expand all | collapse all

33/33 answered

0: Note on personal data! (1 / 1) +

I: Description of data – reuse of existing data and/or production of new data (7 / 7) −

Original (VR) questions:

- How will data be collected, created or **reused**?
- What types of data will be created and/or collected, in terms of data format and amount/volume of data?

A question about expected amount/volume of data is answered below in the section on Storage and backup so is not needed here.

The questions below are meant to give answers conforming to the RDA-DMP-Common-Standard maDMP-schema-1.1.json, describing
1. Data Quality Assurance measures: file formats, file naming conventions, data cleaning etc. 2. Dataset Id(s), 3. Dataset Id Type, 4. Dataset Description.

Each dataset within a project should ideally have separate answers to questions 2-4.
However, at this level in the initial phase, a dataset is defined here only at a conceptual level, so if detailed identifiers of individual datasets or files are still lacking, there is the possibility of regarding the entire project as one dataset, and give the identifier, identifier type and description of that as answers to questions Q2-Q4.
An answer to question Q1 (text), could also cover all datasets within a project.

Questions Q6-Q9 concern at this point only the sources or access points of **Reused datasets** that have already definite identifiers, locations and formats ("distribution" in the vocabulary of the DCAT standard).

Q1: Data Quality Assurance measures (FAIR data): please check multiple options that will apply to assure quality and integrity of data collected, created or reused.

☑ 1. Non-proprietary file formats (e.g. .csv, .txt, .json)
☑ 2. Sustainable file formats (e.g. .pdf; .csv; .txt)
☐ 3. Software specific file formats (e.g. Matlab - .mat; Stata - .dta)
☑ 4. Proprietary file formats (.doc ; .xls)
☐ 5. Datestamped file names (e.g. '20201011StockholmMeanTemperatures.csv')
☑ 6. Descriptive file names (e.g. '20201011StockholmMeanTemperatures.csv')
☑ 7. File names with only character set [0-9A-Za-z-_.], no spaces.
☐ 8. Other measures (please specify!)

Additional Information

B | I | ☰ ▾ | ☰ ▾ | 🔗 | ⊞ ▾

[Save]

Answered 3 months ago by joakim.philipson@su.se

| Guidance | Comments |

SU

Options 1 and 2 for *file formats* will make your datasets more Interoperable and Re-usable, and so will score higher on FAIR data.
(For sustainable and preferred file formats, please consult further: Sustainable digital file formats from Library of Congress, and / or the SND file format guide .)
Options 5, 6 and 7 for *file names* will make your data files more Findable (and thus also contribute to your FAIR data score).

Stockholm University

Demo DMP example of SU-VR template: II-Q1

Data repositories curated by SU RDM-team

https://su.figshare.com

Discover research from Stockholm University
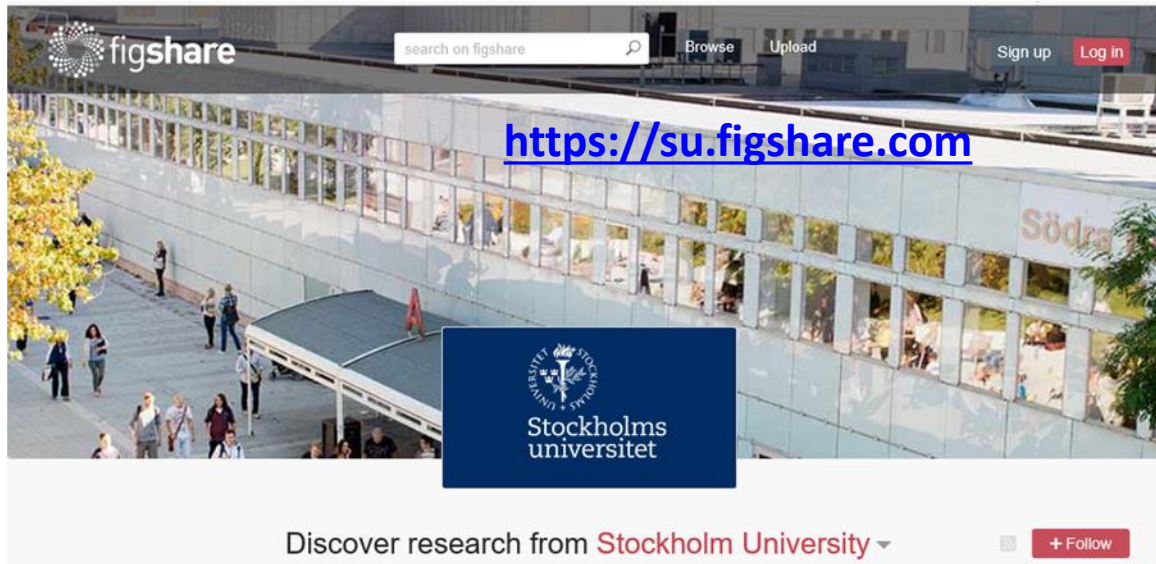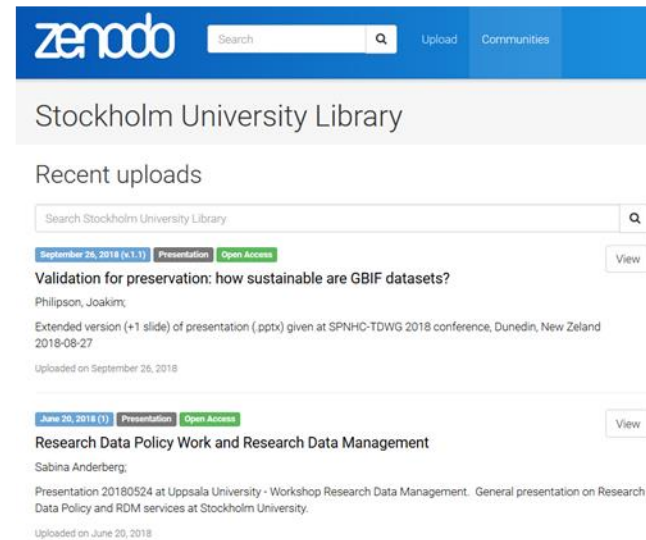
Open data for: Eklund, R., & Wiens, S. (2018). Visual Awareness Negati...
Stefan Wiens   28/02/2018

Supplement for Performing contrast analysis in factorial designs: From...
Stefan Wiens   02/05/2017

Open data for Larsson, M., Tirado, C., & Wiens, S. (2017). A Meta-Anal...
Stefan Wiens   12/05/2017
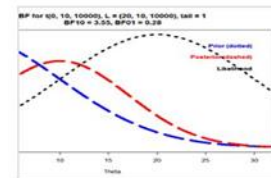
Aladins Bayes Factor in R
Stefan Wiens   05/06/2017

https://zenodo.org/communities /stockholmuniversitylibrary

- **Bolin Centre DB**

- **Dataverse**

- **SND - https://snd.gu.se**

✓ **Create standard metadata** (for DMP q & a)
✓ Pre-order **DOI** for your dataset (or private link)
✓ **Curation** and counseling from SU
✓ *Harvest and transform for* **archival** *(su.figshare, so far)*
✓ *One* **storage** */ backup option during project (embargo)?*

Testing SU maDMP model for Climate Data: Sharing the seas.

Demo DMP example of SU-VR template: II-Q3

II: Documentation and data quality (3 / 3) —

Q3: Which of the following data quality documentation and safeguard measures, if any, would you consider applying to your dataset?
(Multiple options possible. For options 4. *Pre-registration*, 6. *Supplementary documentation* or 9. *Other*, please specify to the extent possible in the comment area, e.g. by giving the URLs of particular services you intend to use for certain data quality measures.)

☐ 1. Data clean(s)ing
☑ 2. File format and software description
☑ 3. Integrity check of data files
☑ 4. Pre-registration
☐ 5. Repeated measurements
☐ 6. Suppl. documentation (e.g. codebook / variable description, ELN, questionnaire, stimuli)
☑ 7. Validation of data input
☑ 8. Registered reports
☐ 9. Other (please specify!)

Additional Information

| B | *I* | ☰ ▾ | ☷ ▾ | 🔗 | ▦ ▾ |

https://osf.io/, https://doi.org/10.1016/j.pacfin.2018.11.003

Save

Answered 2 months ago by joakim.philipson@su.se

| Guidance | Comments |

SU

1. *Data clean(s)ing* and consolidation will contribute to secure quality of data overall. (Here are some possible tools to use for this: frictionlessdata.io ; goodtables.io ; openrefine.org)

2. A detailed *description* (preferably as part of the general metadata entered in a repository web form) of *file formats* and *software* that was used to produce the dataset and possibly (in the case of software specific formats) also needed to open or understand/interpret the data will help make your dataset more Re-usable – and thus more FAIR!

3. *Integrity checks* of datafiles are facilitated by data repositories by means of *checksum* algorithms such as MD5, SHA-256 etc.
[https://en.wikipedia.org/wiki/Checksum, https://en.wikipedia.org/wiki/SHA-2]
These may be particularly relevant to apply when data files are moved from one location to another, to check that nothing was lost on the way or that data has not been "corrupted" since a previous upload. Automatic integrity checks, producing a hash check sum, are built- in to certain applications, but there are also tools to download freely for this, such as DROID from National Archives, UK.

4. *Preregistration* "is the practice of

Stockholm University

Testing SU maDMP model for Climate Data: Sharing the seas.

Demo DMP example of SU-VR template IV-Q1

Project Details	Contributors	Plan overview	Initial DMP	Full DMP (undeveloped)	Share	Download

expand all | collapse all

33/33 answered

II: Documentation and data quality (3 / 3) +

III: Storage and backup (3 / 3) +

IV: Legal and ethical aspects (3 / 3) −

Original (VR) questions:

- How is data handling according to legal requirements safeguarded, e.g. in terms of handling of personal data, confidentiality and intellectual property rights?
- How is correct data handling according to ethical aspects safeguarded?

Q1: Will the creation, collection or reuse of dataset(s) in your project entail processing of *personal data*, i.e. any information relating to an identified or identifiable natural person (a '*data subject*', that is a *living* person)?

○ yes
● no
○ unknown

[Save]

Answered 1 minute ago by joakim.philipson@su.se

This answer removes 2 questions from your plan.

Guidance	Comments

SU

The *General Data Protection Regulation* (**GDPR**) provides the following definition (article 4.1):

'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

If you need help to find out whether you will be processing *personal data* in your project, answer "unknown" and the SU-RDM team will be automatically notified and will contact you about your potential need for support.

Stockholm University

IV: Legal and ethical aspects (4 / 4) —

Original (VR) questions:

- How is data handling according to legal requirements safeguarded, e.g. in terms of handling of personal data, confidentiality and intellectual property rights?
- How is correct data handling according to ethical aspects safeguarded?

Q3: Will the creation, collection or reuse of dataset(s) in your project entail any of the following:

- **(a) processing of *special categories of personal data* according to the General Data Protection Regulation (EU 2016/679), i.e. personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation**
- **(b) processing of personal data regarding violations of law that include crimes, judgments in criminal cases, penal law sanctions, or administrative deprivation of liberty**
- **(c) physical interventions on research subjects or deceased persons**
- **(d) methods with the purpose of affecting a research person physically or mentally, or which includes an apparent risk of injuring the research subject either physically or mentally**
- **(e) studies of biological material that has been taken from a living or deceased person, and can be traced to that person**

Or, further, will the creation, collection or reuse of dataset(s) in your project include:

- **(F) Data from animal research**
- **(G) Data on genetic resources and/or traditional knowledge associated with genetic resources**
- **(H) Data that can be used for military purposes or concerning products that can be used for military purposes**
- **(I) Data that are sensitive in some other respect**

(Possible ethical review documentation applying to any of points (a-e) will be asked for separately.) If you answer yes to any of the points (F-I) and there is already relevant documentation or applications, please provide (a) reference(s) to any application(s)/ approval(s)/decision(s)/document(s), if possible by URL(s) / PID(s) such as DOI(s) giving direct access, or registration no. (Swe. *diarienr.*) in the text field below.

○ yes
● no
○ unknown
○ yes, but not (a-e)

Additional Information

[B | *I* | ≣ ▾ | ≣ ▾ | 🔗 | ▦ ▾]

| Guidance | **Comments** |

SU

First list (a.-e.): Answering *yes* to any of the **points a.- e.** above means that you may need to apply for ethical review for research taking place *in Sweden*. See information here. For research taking place in other countries, other rules may apply.

For support concerning *ethical review of research involving humans*, please contact etik@fs.su.se.

Second list (F-I):
F. Animal research, and the use of data from animal research, may require ethical review or other approvals. See information concerning animal research here (information in Swedish). Some information in English is provided here.

For guidance concerning what applies to animal research, please contact etik@fs.su.se.

G. The use of **genetic resources and/or traditional knowledge associated with genetic resources** is regulated in the **Nagoya Protocol**, implemented through the **EU Access and Benefit-Sharing (ABS) Regulation**. Information from the *Swedish Environmental Protection Agency* is found here. For guidance concerning matters related to the Nagoya Protocol and The ABS Regulation, please contact info@fs.su.se.

Stockholm University

DMP example of SU-VR template: API output & transform

1. Raw.json

```json
[
  {
    "id": 72835,
    "title": "Testing SU maDMP model for Climate Data: Sharing the seas.",
    "grant_number": null,
    "last_updated": "2021-06-16 14:05:01 UTC",
    "creation_date": "2021-03-04 10:42:54 UTC",
    "test_plan": true,
    "template": {
      "title": "SU-VR template",
      "id": 1456260114
    },
```

2. Convert to .xml

```xml
<?xml version="1.0" encoding="UTF-8"?>
<array>
    <id>72835</id>
    <title>Testing SU maDMP model for Climate Data: Sharing the seas.</title>
    <grant_number>null</grant_number>
    <last_updated>2021-06-16 14:05:01 UTC</last_updated>
    <creation_date>2021-03-04 10:42:54 UTC</creation_date>
    <test_plan>true</test_plan>
    <template>
        <title>SU-VR template</title>
        <id>...
```

3. Transform & FAIR eval.

```xml
<?xml version="1.0" encoding="UTF-8"?>
<dmp xmlns:xlink="http://www.w3.org/1999/xlink"
     xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
     xmlns:mets="http://www.loc.gov/METS/"
     xmlns:local="local-functions"
     OBJID="72835"
     LABEL="Testing SU maDMP model for Climate Data: Sharing the seas.">
    <title>Testing SU maDMP model for Climate Data: Sharing the seas.</title>
```

```xml
        <FAIRscore>
            <sectionI>
                <DataDescription>
                    <FileFormatsUNamesQ1>
                        <Option>7. File names with only character set [0-9A-Za-z-_.], no
                                spaces.</Option>
                        <F-value>1</F-value>
                        <Option>6. Descriptive file names (e.g.
                                '20201011StockholmMeanTemperatures.csv) </Option>
                        <F-value>1</F-value>
                        <Option>4. Proprietary file formats (.doc ; .xls)</Option>
                        <Option>1. Non-proprietary file formats (e.g. .csv, .txt, .json) </Option>
                        <I-value>1</I-value>
                        <R-value>1</R-value>
                        <Option>2. Sustainable file formats (e.g. .pdf; .csv; .txt)</Option>
                        <I-value>1</I-value>
                        <R-value>1</R-value>
                    </FileFormatsUNamesQ1>
```

4. Valid RDA maDMP-schema-1.1.json

```json
{
    "dmp": {
        "xmlns:xlink": "http://www.w3.org/1999/xlink",
        "xmlns:xsi": "http://www.w3.org/2001/XMLSchema-instance",
        "xmlns:mets": "http://www.loc.gov/METS/",
        "xmlns:local": "local-functions",
        "OBJID": 72835,
        "LABEL": "Testing SU maDMP model for Climate Data: Sharing the seas.",
        "title": "Testing SU maDMP model for Climate Data: Sharing the seas.",
        "description": "Using this published dataset as basis example:\r Crona, Beatrice
        "created": "2021-03-04T10:42:54Z",
        "modified": "2021-06-16T14:05:01Z",
        "contact": {
            "mbox": "kl@su.se",
            "name": "Lillepold, Kate",
            "contact_id": {
```

Validate with
URL: /C:/Users/joph9849/Desktop/DMPetPolicy/RDA-DMPvalidation/maDMP-schema-1.1.json
Schema type: JSON
☐ Embedded Schematron rules
Public ID:
Schematron phase:
OK Cancel

UCL-SUpresent\DMP72835transform.j... ■ Validation successful

Stockholm University

Thanks for your attention!

Contact

✉ opendata@su.se