Stockholm University

machine actionable Data Management Plans for the future at Stockholm University?

Joakim Philipson - Research Data Analyst

Contact: opendata@su.se

Presentation at RDA-SND maDMP workshop
2020-05-26

What is a Data Management Plan?

Pixabay CC0

Answers questions about What? Where? When? and How? data generated in a research project will be collected / produced / processed/ published / stored

- -> DMPs are *metadata* describing other data and (meta)data
- -> DMPs need a *metadata standard* for that description
- -> DMPs should be FAIR!
- -> DMPs should be machine-actionable (a sine-qua-non for FAIRness)
- -> we need maDMPs!

Stockholm University

Why a Data Management Plan? – what is it good for?- *some use cases*

Pixabay CC0

- **Funders/Curators/Reviewers/Researchers**:

Checking funder requirements are met!

need: easy review / *validation* -> maDMP!

- **Data Managers / Curators / IT-dept.**:

Checking compliance with local data policy!

need: easy review / *validation* -> **maDMP!**

Planning for future data storage!
need: estimates of data vols. -> **maDMP!**

- **Researchers**:

Easing the administrative burden, easy fill-out forms and re-use info!

need: export/ import / linking to re-usable metadata in other systems - **-> maDMP**

Project planning and cooperation tool!
need: sharing / *interoperability* -> **maDMP!**

Secure *Findable* and *Re-usable* data for the future!
need: Backup, standard metadata and checking for **FAIR**! -> **maDMP!**

Stockholm University

How to fulfil the needs? Means to an end.

Pixabay CC0

Funders/Curators/Managers/Reviewers/ Researchers:

needs: easy fill out and review / *validation* -> **maDMP**:

- Auto-completion
- pre-populated template (q&a) / web-forms (from funder/organization)
- data-typing (content constraints)
- enumeration lists / drop-down menus (multiple choice)
  import from data repositories, CRIS systems, ethical vetting etc.

  -> *inherent validation*

Managers/Reviewers/Researchers/Collaborators

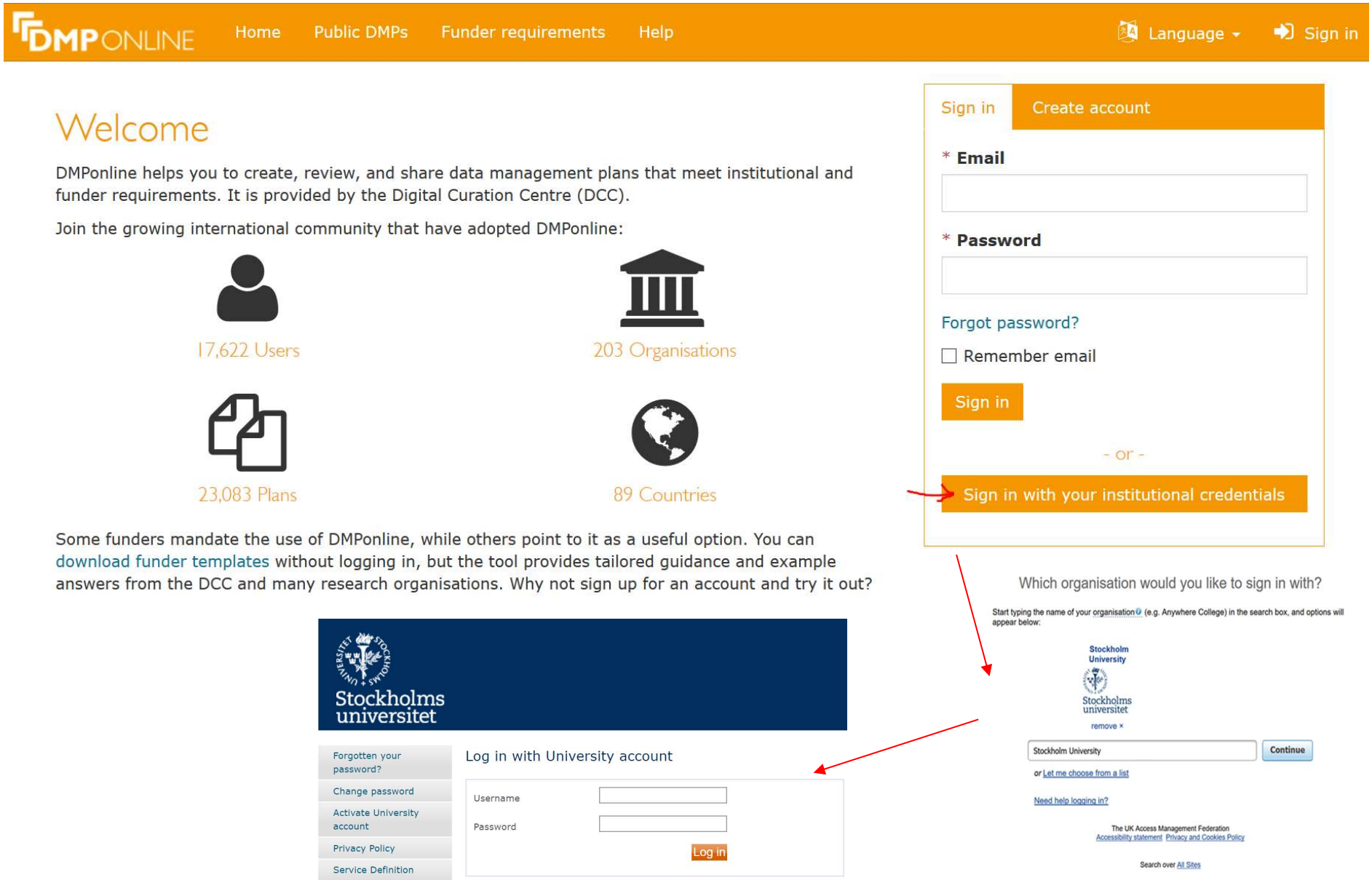
needs: interoperability, press submit ->publish digital object/report

- accessible schemas
- online validation service
- metadata schemas compatible with repositories for data publication
- APIs for data portability

-> *external validation*

Stockholm University

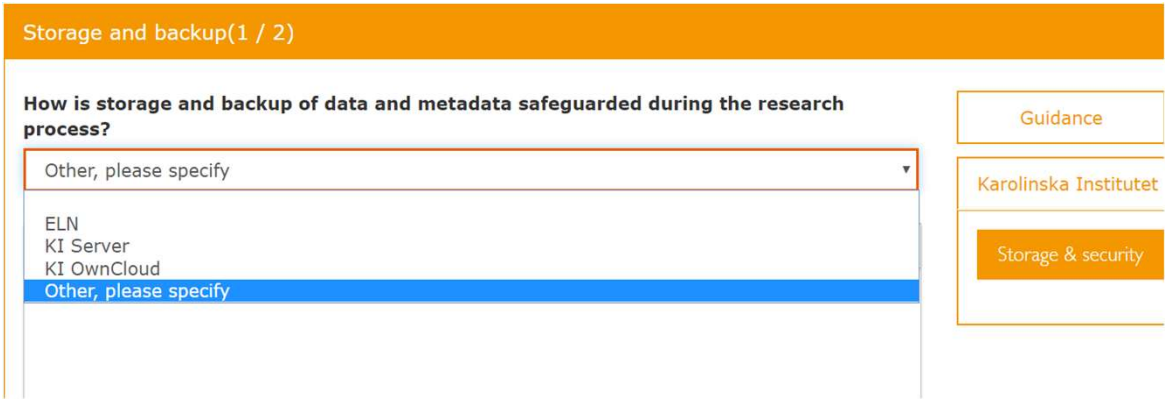
DMP Online – login from https://dmp.su.se



Examples – what can be done?

enumeration lists / drop-down menus (multiple choice) in DMP Online (Karolinska Institute):

Data volumes



Storage & backup options

Stockholm University

Global corporate capital flows and biosphere resilience

Project Details | Plan overview | Write Plan | Share | Download

expand all | collapse all

13/13 answered

Data Collection (2 / 2)

What data will you collect or create?

I will not collect data but will use existing data and third-party sources.

The main publicly available sources of data are:

- Geolocalised data from the Brazilian Institute of spatial studies about land use and forest cover in the Amazon biome. Shapefiles. http://www.inpe.br/ Creative Commons Attribution-ShareAlike (CC BY-SA) 3.0
- Economic statistics from the Brazilian Institute of Geography and Statistics, municipal and state level. Spreadsheets. https://www.ibge.gov.br/ CC BY-SA
- Database created in a previous project about flows of foreign capital to companies in Brazil, from the Central Bank of Brazil, 2000-2011. CSV. https://doi.org/10.17045/sthlmuni.5857716 CC BY-NC 4.0
- Data about flows of foreign capital to companies in Brazil, from the Central Bank of Brazil, 2011-2019. CSV. https://dadosabertos.bcb.gov.br CC BY-SA
- Data about loans and subsidies extended by the Brazilian Development Bank. https://www.bndes.gov.br CC BY-SA
- Data from the research project Trase about exports of different commodities from Brazil. CSV. https://trase.earth/ CC BY 4.0
- Forest cover and forest concessions data from Global Forest Watch. Shapefiles. http://data.globalforestwatch.org/ CC BY 4.0
- Financial statements, annual reports, proxy reports and sustainability reports published by various companies. PDF. Various licenses.
- United States Geological Survey Yearbooks on minerals in Brazil. PDF. https://www.usgs.gov/ US public domain.

Guidance | Comments

DCC

Questions to consider:

- What type, format and volume of data?
- Do your chosen formats and software enable sharing and long-term access to the data?
- Are there any existing data that you can reuse?

Guidance:

Give a brief description of the data, including any existing data or third-party sources that will be used, in each case noting its content, type and coverage. Outline and justify your choice of format and consider the implications of data format and data volumes in terms of storage, backup and access.

https://doi.org/10.17045/sthlmuni.5857716.v4

Stockholm University

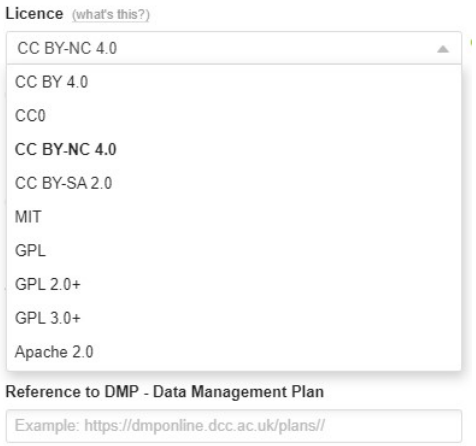
DMP *Global_corporate_capital_flows…*:

Ethics and Legal Compliance (2 / 2)

How will you manage copyright and Intellectual Property Rights (IPR) issues?

Please refer to the list of data sources under the section 'Data Collection', which includes a discussion of copyright issues.

Licensing help via repos (e.g. su.figshare.com)



EUDAT licensing wizard: https://ufal.github.io/public-license-selector/

Intellectual Property Rights

- State who will own the copyright and IPR of any existing data as well as new data that you will generate. For multi-partner projects, IPR ownership should be covered in the consortium agreement.
- Outline any restrictions needed on data sharing, e.g., to protect proprietary or patentable data.
- Explain how the data will be licensed for reuse. See the DCC guide on How to license research data and EUDAT's data and software licensing wizard.

Guidance | Comments

DCC

Questions to consider:

- Who owns the data?
- How will the data be licensed for reuse?
- Are there any restrictions on the reuse of third-party data?
- Will data sharing be postponed / restricted e.g. to publish or seek patents?

Guidance:

State who will own the copyright and IPR of any data that you will collect or create, along with the licence(s) for its use and reuse. For multi-partner projects, IPR ownership may be worth covering in a consortium agreement. Consider any relevant funder, institutional, departmental or group policies on copyright or IPR. Also consider permissions to reuse third-party data and any restrictions needed on data sharing.

Examples – ongoing integration efforts

- **Common metadata standard**(?) – or at least *interoperable /* crosswalks (transforms, maps)

RDA DMP Common Standard, DCAT(?),
Ontology work (hack. issue),

DataCite (UCT hack. issue)

- Integration with CRIS (UCT), InvenioRDM (hack. issues)
- **Metadata schemas compatible** with repositories for data publication

-> import from data repositories

-> export from repositories to maDMP (Dataverse, Figshare – hack. issue)

- **APIs for (meta)data portability**:

export *maDMP* from RDMO, DMP Online/DMP Tool, DSW

Stockholm University

Examples – ongoing validation efforts

- **Inherent validation:**

Data Stewardship Wizard - FAIRness metrics,

Datatyping in easyDMP (hack. Issue)

- **External validation**(?):

SU idea of phased validation schema (Schematron), conditional on funder and organisation (univ.) for self-validation / review, also of compliance with local organisation data policy + online validation service (links to already existing)

Stockholm University

But wait, what *is machine actionable* anyway?

Def. = ***machine actionable*** in foundational paper of FAIR:

- "a continuum of possible states wherein a digital object provides increasingly more detailed information to an autonomously-acting, computational data explorer." (***Wilkinson, M. et al., 2016*** **https://doi.org/10.1038/sdata.2016.18**)

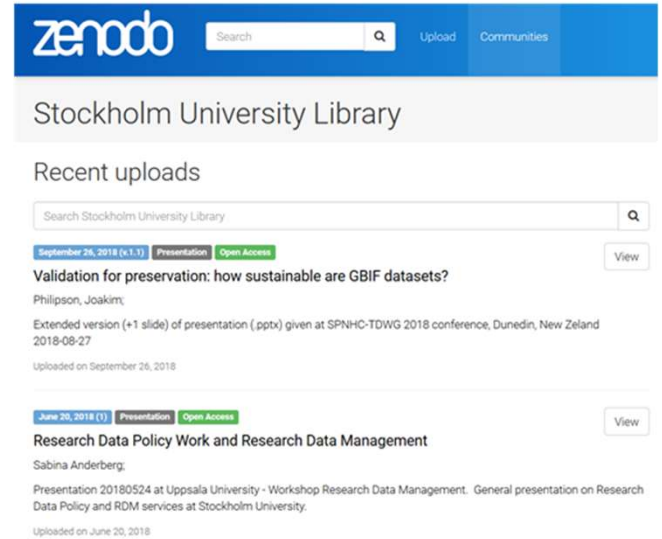
- **file format identification**
- **content constraints (incl. data types and homogeneous input)**
- **complexity of metadata standard**
- **schema validation against accepted metadata standards**
- ***semantic* sustainability of metadata - *interpretability***
- **sustainability of metadata standards and file formats**

The last two points on this list address the obvious ***time aspect*** of machine actionability; ***machine actionable not only now***,

but also in a distant ***future***!

Stockholm University

Data repositories curated by SU RDM-team

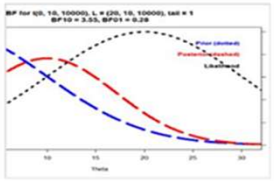


https://zenodo.org/communities/stockholmuniversitylibrary

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

- **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*)*?

Stockholm University

Thanks for your attention!

Contact

opendata@su.se