

Workshop

Deep Dive into Data

für Antragsteller*innen bei der IKT der Zukunft Ausschreibung 2020

Fachinput für Paneldiskussion Data Management Plan
Axel Quitt, DIO Data Steward / Consultant

Inhalt

- DMPOnline als Werkzeug für DMP-Erstellung
- Vertraulichkeit
- Best Practice
- DMP Grundstruktur
- Praxistipp

Verpflichtender Datenmanagement-Plan

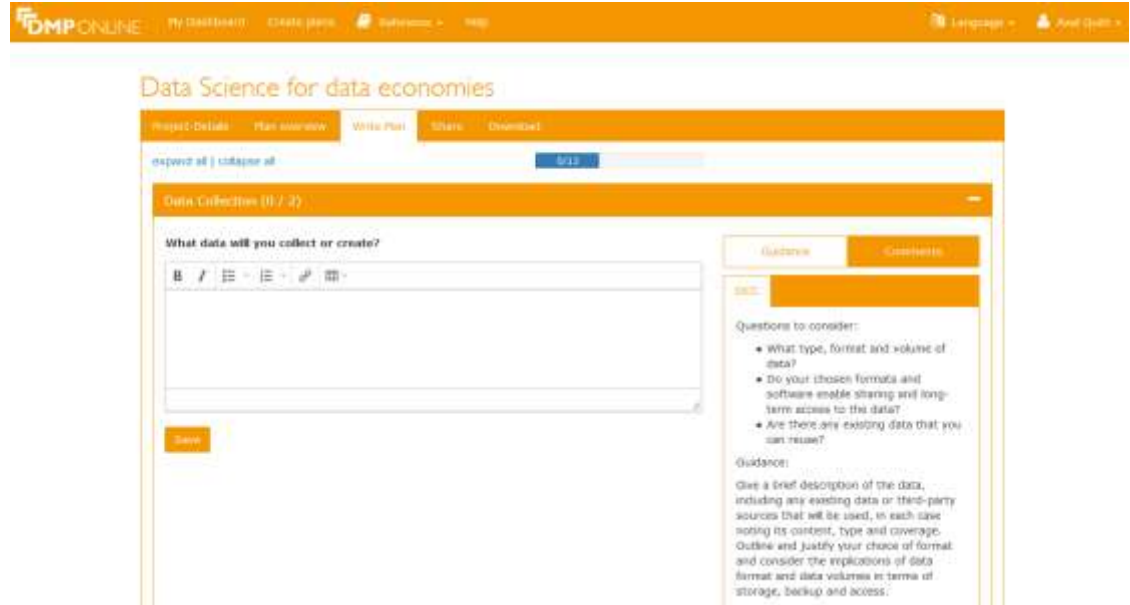
„Antragsteller*innen bei diesem Schwerpunkt sind verpflichtet, einen Datenmanagementplan (DMP) als Annex zur Projektbeschreibung vorzulegen.

Ein Datenmanagement-Plan beschreibt,

- welche Daten im Projekt gesammelt, erarbeitet oder generiert werden,
- wie mit diesen Daten im Projekt umgegangen wird,
- welche Methoden und Standards dabei angewendet werden,
- wie die Daten langfristig gesichert und gepflegt werden, und
- ob es geplant ist, Datensätze Dritten zugänglich zu machen und ihnen die Nachnutzung der Daten zu ermöglichen (sog. „Open Access zu Forschungsdaten“ oder auch in „Datenkreisen“ – ...). ¹

Das Online-Tool DMP Online

- <https://dmponline.dcc.ac.uk/plans>
- Online Formular Service
- Data Management-Plan Struktur
- Ausfüllhilfe inkludiert
- Fragen und Richtlinien unterstützen bei DMP Erstellung
- Content ausschließlich Englisch
- Zusätzlich unterstützt es die Zusammenarbeit im Team



Was ist, wenn ich meinen DMP nicht der Cloud anvertrauen möchte?

Man kann seinen DMP direkt im DMP Online Tool ausfüllen, aber man muss nicht.

Das DMP Online Tool bietet die Möglichkeiten, dass ein Team gemeinsam am DMP arbeitet und es von dieser Plattform direkt publiziert wird.

Will man höchste Vertraulichkeit, kann man die Struktur auch als Word-Datei herunterladen (sogar mit Fragen und Richtlinien) und „Private“ auf eigener Infrastruktur arbeiten

Manage collaborators

Invite specific people to read, edit, or administer your plan. Invitees will receive an email notification that they have access to this plan.

Email address	Permissions	
stefan.gindl@researchstudio.at	Editor	Remove
axel.qutt@enginyra.com	Owner	

Invite collaborators

* Email

* Permissions

- Co-owner
- Editor
- Read only



Data Science for data economies

Project Details | Plan overview | **Word Plan** | Status | Download

Download settings

Optional Plan Components
project details overview
question text and section headings
unanswered questions

Format: **pdf** ↓ **oder .docx**

PDF formatting

Font

Face: Arial, Helvetica, Sans Serif | Size (pt): 10

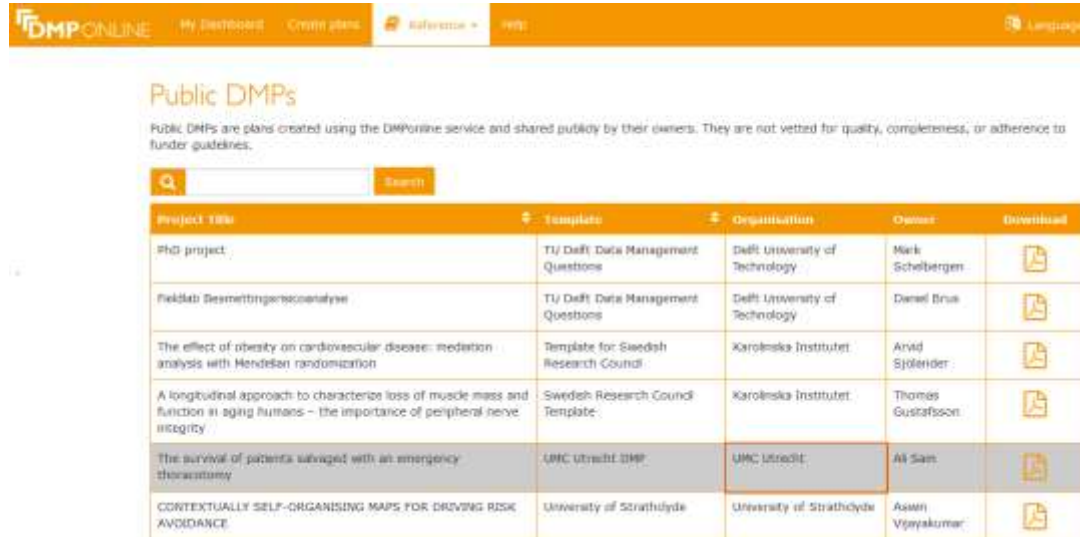
Download Plan

DIO







Bietet DMPOnline Best Practice Beispiele?

Verfügbar ist ein Verzeichnis von vielen bereits publizierten Data Management Plänen

Eine Bewertung über deren Qualität oder Vollständigkeit wird NICHT angeführt



The screenshot shows the DMPOnline website interface. At the top, there is a navigation bar with the DMPOnline logo, links for 'My Dashboard', 'Create plans', 'Reference', and 'Help', and a 'Language' dropdown menu. Below the navigation bar, the page title is 'Public DMPs'. A sub-header explains: 'Public DMPs are plans created using the DMPonline service and shared publicly by their owners. They are not vetted for quality, completeness, or adherence to funder guidelines.' Below this is a search bar with a magnifying glass icon and a 'Search' button. The main content is a table with the following columns: Project title, Template, Organisation, Owner, and Download. The table lists six public DMPs.

Project title	Template	Organisation	Owner	Download
PhD project	TU Delft Data Management Questions	Delft University of Technology	Mink Schelbergen	
Fieldlab Bestmättningsreosonalyse	TU Delft Data Management Questions	Delft University of Technology	Daniel Brus	
The effect of obesity on cardiovascular disease: mediation analysis with Mendelian randomization	Template for Swedish Research Council	Karolinska Institutet	Arvid Ståhlender	
A longitudinal approach to characterize loss of muscle mass and function in aging humans – the importance of peripheral nerve integrity	Swedish Research Council Template	Karolinska Institutet	Thomas Gustafsson	
The survival of patients salvaged with an emergency thoracotomy	UMC Utrecht DMP	UMC Utrecht	Ali Sam	
CONTEXTUALLY SELF-ORGANISING MAPS FOR DRIVING RISK AVOIDANCE	University of Strathclyde	University of Strathclyde	Aswin Vijayakumar	

Grundstruktur eines Data Management Plans in DMPOnline

The screenshot displays the DMPOnline interface. At the top, there is a navigation bar with the logo 'DMP ONLINE' and links for 'My Dashboard', 'Create plans', 'Reference', and ' Hilfe'. On the right side of the bar, there are options for 'Language' and the user 'Axel Quitt'. Below the navigation bar, the main content area is titled 'Data Science for data economies'. A sub-navigation bar contains tabs for 'Project Details', 'Plan overview', 'Write Plan', 'Teilen', and 'Download'. The 'Write Plan' tab is active. Below this, there are controls for 'expand all | collapse all' and a blue button indicating '13/13 answered'. The main content is a list of seven categories, each with a progress indicator and a plus sign for expansion:

- Data Collection (2 / 2)
- Documentation and Metadata (1 / 1)
- Ethics and Legal Compliance (2 / 2)
- Storage and Backup (2 / 2)
- Selection and Preservation (2 / 2)
- Data Sharing (2 / 2)
- Responsibilities and Resources (2 / 2)

Praxistipp

- Kopiere Fragen ins Formular und Download
- Setze Schwerpunkte (z.B. hier in Gelb)

DATA SCIENCE FOR DATA ECONOMIES

A Data Management Plan created using DMPOnline

Creator: Axel Guft

Affiliation: Other

Template: DCC Template

Project abstract:

How do Data Economies rely on efforts of data scientists

Last modified: 24-01-2021

DATA SCIENCE FOR DATA ECONOMIES

DATA COLLECTION

What data will you collect or create?

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 at 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.

- Note what volume of data you will create in MB/GB/TB. Indicate the proportion of **raw data, processed data, and other secondary outputs (e.g., reports)**.
- Consider the implications of data volumes in terms of storage, access and preservation. Do you need to **isolate additional costs?**
- Consider whether the scale of the data will pose **challenges when sharing or transferring data** between sites; if so, how will you address these challenges?

- See UK Data Service guidance on [recommended formats/Open in a new window](#) or DataONE Best Practices for [file formats/Open in a new window](#)
- Give a summary of the data you will collect or create, noting the content, coverage and data type, e.g., labular data, survey data, experimental measurements, models, software, audiovisual data, physical samples, etc.
- Consider how your data could complement and integrate with existing data, or whether there are any existing data or methods that you could reuse.
- Indicate which data are of long-term value and should be shared and/or preserved.
- If purchasing or reusing existing data, explain how issues such as copyright and IPR have been addressed. You should aim to minimise any restrictions on the reuse (and subsequent sharing) of third-party data.

How will the data be collected or created?

Questions to consider:

- What standards or methodologies will you use?
- How will you structure and name your folders and files?
- How will you handle versioning?
- **What quality assurance processes will you adopt?**

Guidance:

Outline how the data will be collected/created and which community data standards (if any) will be used. Consider how the data will be organised during the project, mentioning for example naming conventions, version control and folder structures. Explain how the consistency and quality of data collection will be controlled and documented. This may include processes such as calibration, repeat samples or measurements, standardised data capture or recording, data entry validation, peer review of data or representation with controlled vocabularies.

- Outline how the data will be collected and processed. This should cover relevant standards or methods, quality assurance and data organisation.
- Indicate how the data will be organised during the project, mentioning, e.g., naming conventions, version control and folder structures. Consistent, well-ordered research data will be easier to find, understand and reuse.
- Explain how the consistency and quality of data collection will be controlled and documented. This may include processes such as calibration, repeat samples or measurements, standardised data capture, data entry validation, peer review of data or representation with controlled vocabularies.
- See the DataOne Best Practices for [data quality/Open in a new window](#)

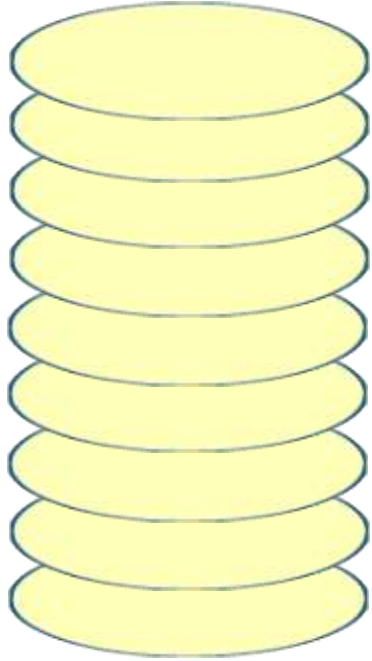


Weitere Ressourcen

Moderationsinput: Interessante Fragen die für Data Management Plan interessant sein können

- Welche Daten können wiederverwendet werden → im Sinne bereits existierender Daten, die Eingang in die Forschungsarbeit finden und im 2. Sinne – welche werden nach der Forschungsarbeit existieren und wiederverwendet werden können.
- Wie findet man diese, und unter welchen rechtlichen Rahmenbedingungen können sie genutzt werden → Link zu Moderation im Rechtsteil
- Herausforderungen beim Datenaustausch (Menge, Formate, Tools, Vorarbeiten) und Möglichkeiten sie zu überwinden → möglichst frühzeitig überlegen (Praxistipp)
- Wie verhandelt man, um Übereinstimmung/Einvernehmen zur Datenverwendung zu erzielen → Tipps und Erfahrungen von den Teilnehmern
- Betriebswirtschaftliche Betrachtungen, v.a. Kosten für Datenakquisition, Vorverarbeitung etc. und Stakeholderbetrachtungen gem. Datenmanagement-Layern.

Ebenen des Datenmanagements



- Use Case Vision und Anforderungsdefinition
- Datenwert- / -preisbemessungsmethoden
- Stakeholder-Analyse national / international
- Umfeld-Analyse IST-Stand national / international – bereits bestehende DK
- Rechtliche Rahmenbedingungen für Use Case Umsetzung und Datenverwendung
- Datenmanagement – Datenaufbereitung, -anreicherung, -qualität
- Servicearchitektur / Technische Schnittstellen / Funktionalität
- Dokumentation Datenkreis – Managementebenen
- Infrastrukturarchitektur / Betriebsprozesse

Beispiel: Datamanagement-Plan des Data Market Austria (DMA) Forschungsprojekts¹



**DATA MARKET
AUSTRIA**

www.datamarket.at

Data Management Plan

Deliverable number: D1.5
Dissemination level: Public
Delivery date: 2020-04-30
Status: Final Data Management Plan

Author(s): Alois Wipflinger, Sven Gschürtz, Steffen Ratt

The Data Market Austria Project has received funding from the programme "ICT of the Future" of the Austrian Research Promotion Agency (FFG) and the Austrian Ministry for Transport, Innovation and Technology (Project 859404).



Table of Contents

- Introduction
- Data Summary
 - Data types and origin
 - List of Datasets
- FAIR Data
 - Making data Findable, including provisions for metadata
- Data Catalogue
 - Data Catalogue
 - Dataset
 - Data Distribution
 - Service Metadata
 - General Service Properties
- Making data (openly) accessible
- Making data interoperable
- Data re-use & Licensing of data
- Allocation of Resources
- Estimated Costs
- Responsibilities
- Long Term Preservation
- Data Security
- Legal and ethical Aspects
 - 5.1 Data Protection
 - 5.2 Measures to ensure ethical and legal standards
 - 5.3 Privacy and trust
 - 5.4 Survey and data collection in Task 5.4
- References

to make use of the central node hosted by T-Systems for administering access keys.

DMA acts as an uniform data access platform, in its current version metadata. Datasets will never be stored in DMA in cases where 1) it is too large (e.g. due to file size); 2) the data provider does not want DMA (but rather provides access to it through an API); 3) it does not relate to DMA (e.g. it is already available open data). ILOs by public providers allows to store/locate (encrypted) datasets on distributed storage functionality as well as an encryption system have not been implemented.

DMA, Data catalogue, is stored or distributed in the catalogue. Documentation of the metadata (GDLP).

The project developed a service for ingesting data management components (Glossary) in a partial. The ingest service is available as an API provides guided input process for data description include metadata validation (format, size, enrichment of the metadata).

The metadata schema applied in the project is based on DCAT. Metadata is converted on-the-fly




Figure 1: DMA Components Architecture

Combining data from heterogeneous data sources (company, research data centers, Science, usage rights; terms of service, service level agreements) such as restricting access to private data are involved. Blockchain technology in the following areas:

- Unique identification of data assets, services and user addresses (Ethereum Identifiable Virtual Accounts).
- Data asset provenance by capturing important events/modification of a data asset.
- Membership voting for managing the membership.

An encryption system can be implemented upon request (e.g. if data provider (SaaS), Data provider can provide encrypted data and explanation of how it was created.

<https://www.ethereum.org/en/learn/contracts-and-tokens/memberships.html>

It is not a primary purpose of the DMA Data-Service Ecosystem to provide open data resources. Its aim is to provide a trading platform for closed, semi-closed, and open data, Data providing project partners issued agreements on how the data provided by them can be used for the duration of the project. Closed and semi-closed data will not be generally shared during or after the project lifetime, but only according to the standardized license, whose terms will be approved and stored in the blockchain.

DMA implemented an infrastructure for ensuring and accessing federated data and services provided via the DMA platform. Applied Blockchain technology for data access regulation, in particular self-executing contracts on the Blockchain for accessing closed and semi-closed datasets were implemented to model even fine-grained data access and data usage arrangements. The authentication gateway is hosted by DMA and is connected with the Blockchain component. The system registers links to closed or semi-closed data resources in the authorization gateway and generates identifiers on the blockchain, which are linked to a smart contract. Once the user is granted access to the data set, he/she is redirected to the data provider's platform to access and download available datasets.

Data access levels demanding the highest degree of legal certainty are those affecting private data or data for which royalties on a per use or per user basis have to be made. The challenges faced here consist of delivery checks have to be made to guarantee that only the beneficiary gains access to the data, the granularity at which data can be accessed, and the legal status according to which a service is delivered, or the access cannot be restricted.

2.3 Making data interoperable

On the beta version of the DMA Portal² an overview of recently created data sets and services is provided. The DMA Portal landing page is the GUI for the central node, which provides the necessary functionality to run the basic processes related and documented as user stories. The central node is designed in a manner that the access to data becomes independent of the type of cloud or infrastructure provider. OpenShift was used as basis for the container deployment.

We break down the use of metadata and standards into various use cases. Only user stories (GUI) and sub-elements related to interoperability of data are listed here:

- GUI: Browse public general
 - gather general information
 - identify relevant metadata & services from catalogue (search & browse)
 - access general documentation etc.
 - the distribution into other functions, independently on which infrastructure they run is defined by building blocks.
- W3C Dataset management (creation & update)
 - Basic data set management for creation and editing
 - Choose data producing method

² For open datasets only unique identifiers are generated, and they are not stored on the blockchain as access rights are not limited and there is no context to request.

³ <https://www.datamarket.at/>

Data Catalogue

Identifier	Definition/Description	
Dataset	This property links the Catalogue Catalogue	1
Main Description	Describes the content of the Dataset, property can be repeated for parallel links access of the description	1
Publisher	This property refers to an entity (organizational) responsible for making the Catalogue available	1
Title	Describes the name of the Data Catalogue	1
Catalogue unique identifier	Unique ID of the Data Catalogue	1

¹ <https://datamarket.at/ffg/15/management/>

² <https://www.aon.com/aon/infocenter/infocenter-application-on-the-data-market-ecosystem>