

WORKSHOP

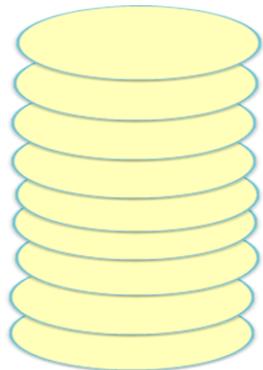
Deep Dive into Data für Antragsteller*innen bei der IKT der Zukunft Ausschreibung 2020

Moderation: Dipl.-Ing. Dr. Stefan Gindl, Senior Data Scientist / RSA FG
Input: Dipl.-Ing. Axel Quitt, Data Steward, Consultant / DIO

Verpflichtender Datenmanagement-Plan (deutsch)

- „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“

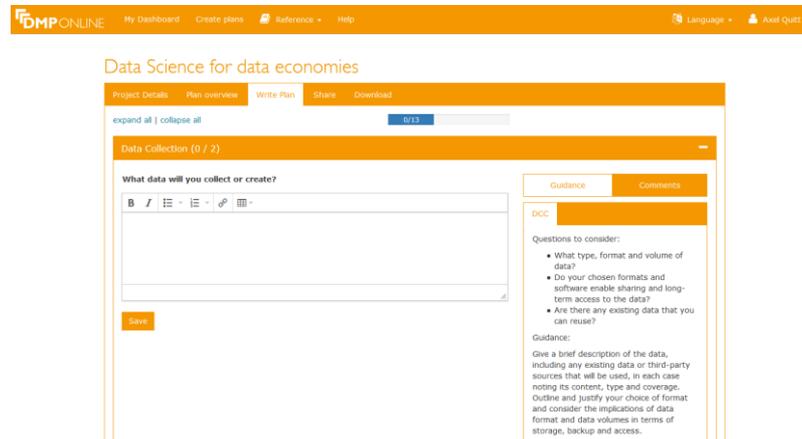
Ebenen des Datenmanagements



- 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

Das Online-Tool DMPonline

- <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 DMPonline Tool ausfüllen, aber man muss nicht.
- Das DMPonline 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.quitt@enginyra.com	Owner	

Invite collaborators

* Email

* Permissions

- Co-owner
 Editor
 Read only

Data Science for data economies

Project Details Plan overview Write Plan Share Download

Download settings

Optional Plan Components

- project details coversheet
 question text and section headings
 unanswered questions

Format

pdf

oder .docx

PDF formatting

Font

Face

Arial, Helvetica, Sans-Serif

Size (pt)

10

Download Plan

Gibt es auf DMPonline Beispielpläne?

- Verfügbar ist ein Verzeichnis von vielen bereits publizierten DMPs
- Eine Bewertung über deren Qualität oder Vollständigkeit wird NICHT angeführt

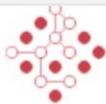


Public DMPs

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.

Project Title	Template	Organisation	Owner	Download
PhD project	TU Delft Data Management Questions	Delft University of Technology	Mark Schelbergen	
Fieldlab Besmettingsrisicoanalyse	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 Sjölander	
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	

Beispiel: Datamanagement-Plan des Data Market Austria (DMA) Forschungsprojekts ¹



DATA MARKET
AUSTRIA

www.datamarket.at

Data Management Plan

Deliverable number	D1.5
Dissemination level	Public
Delivery date	2019-09-30
Status	Final Data Management Plan
Author(s)	Michela Vignoli, Sven Schlarb, Roman Karl

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 855404)



Table of Contents

- Introduction
- Data Summary
 - Data types and origin
 - List of Datasets
- FAIR Data
 - Making data findable, including provisions for metadata
 - Data Metadata
 - 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 3.4
- References

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

DMA acts as a uniform data access platform. In its current version metadata. Datasets will never be stored in DMA in cases where 1) it is transferred data (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 make to DMA (e.g. it is already available open data, SLAs by public providers allows to store/replicate (encrypted) datasets on distributed storage functionality as well as an encryption system have not been implemented

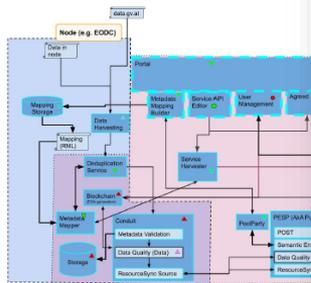


Figure 1 : DMA Components Architecture

Combining data from heterogeneous data sources becomes increasing data owners, licenses, usage rights, terms of service, service level agreements such as restricting access to private data are involved. T Blockchain technology in the following areas:

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

¹ An encryption system can be implemented upon request (e.g. if data provider DMA). Data providers can provide encrypted data and explanation of how it customer.

² <http://www.ethdocs.org/en/latest/contracts-and-transactions.html#externally-owned-accounts-eoa>

DMA data catalogue, is stored on distributed ledger the catalogue. Documentation of the implemented GILab¹.

The project developed a service for ingesting data data management component (ConduIT) is host portal. The ingest service is available as an API provides guided input process for data description include metadata validation (formation, size, enrichment of the metadata).

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

2.1.1 Data Metadata

The DMA metadata catalogue is based on the DC and extends the schema for DMA use cases. This international data portals and ensures that companies.

The DMA schema is, as mentioned, similar to the Catalogue, Data-Dataset and Service-Metadata separated into Data-Dataset-distribution entities.

The Data-Catalogue consists of descriptions of datasets for a particular topic or company. A Dataset a single source, and available for access or download All the metadata fields in the DMA-Metadata Catalogue of data is able to process information must provide information about instances of the

An overview of the DMA-Metadata Core metadata

Data Catalogue

Identifier	Definition/Description
Datasets	This property links the Catalogue Catalogue
Main Description	Describes the content of the Data Catalogue. This property can be repeated for parallel language versions of the description.
Publisher	This property refers to an entity (organisation) responsible for making the Catalogue available.
Title	Describes the name of the Data Catalogue
Catalogue Unique Identifier	Unique Id of the Data Catalogue

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 runtime, but only according to the standardised license, whose terms will be expressed and stored in the blockchain.

DMA implemented an infrastructure for querying 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 authorisation gateway is hosted by ZAMG and is connected with the blockchain component. The system registers links to closed or semi-closed data resources in the authorisation 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 providers' 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 comprise speed 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 repudiated.

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. Open Shift was used as basis for the container deployment.

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

US1: Browse public (portal)

- 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.

US2: Dataset management (creation / upload / ...)

- Basic data set management for creation and editing
- Choose data provisioning method

¹ For open datasets only random identifiers are generated, and they are not stored on the blockchain as access rights are not limited and thus no contract is required.

² <https://portal.datamarket.at/>

¹ <https://datamarket.gitlab.io/Documentation/>

² <https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe>

Grundstruktur eines Data Management Plans in DMPonline

DMP ONLINE My Dashboard Create plans Reference Hilfe Language Axel Quitt

Data Science for data economies

Project Details Plan overview Write Plan Teilen Download

expand all | collapse all 13/13 answered

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

- Kopieren Sie Fragen ins Formular und Download
- Setzen Sie Schwerpunkte (z.B. hier in Gelb)

DATA SCIENCE FOR DATA ECONOMIES

A Data Management Plan created using DMPonline

Creator: Axel Quitt

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 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.

- Note what volume of data you will create in MB/GB/TB. Indicate the proportions **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 include 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](#) **Opens in a new window** or DataONE Best Practices for [file formats](#) **Opens in a new window**
- Give a summary of the data you will collect or create, noting the content, coverage and data type, e.g., tabular 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](#) **Opens in a new window**.

Interessante Fragen die für den DMP 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.