DIO
Data Intelligence Offensive
Verein zur Förderung der Datenwirtschaft

Praxiserfahrung Data Stewardship

Erfahrungen im Zuge der FFG Ausschreibung IKT der Zukunft

Arbeitsschwerpunkte DIO

AUFBAU
eines Netzwerks, Funktion als Kooperationsplattform

Aufbau einer (FACH-) COMMUNITY

BERATUNG
und Zusammenarbeit mit Technologiehersteller*innen

ORGANISATION
und Abhaltung spezifischer Events

UNTERSTÜTZUNG
beim Übergang in die Datenwirtschaft,
beim Einsatz relevanter Technologien, bei der Marktbildung, beim sicheren Datenmanagement

LOBBYING
Weiterleitung von Anliegen des DIO-Netzwerks in den öffentlichen Sektor

Der Auslöser – FFG Anforderung DMP

- „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] “

Quelle: [1] Ausschreibung IKTdZ

Der Schock: DMA DMP[1]

DIO
Data Intelligence Offensive

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

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 t
transfer 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 storag
functionality as well as an encryption system have not been implemente
.

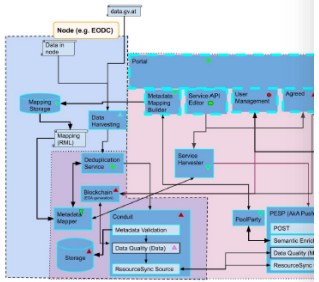


Figure 1 : DMA Components Architecture

Combining data from heterogeneous data sources becomes increasin
data owners, licenses, usage rights, terms of service, service level ag
measures 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[2]).
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 da
customer.

[2] http://www.ethdocs.org/en/latest/contracts-and-transact
transactions.html#externally-owned-accounts-eoas

DMA data catalogue, is stored on distributed re
the catalogue. Documentation of the impleme
GitLab[4].

The project developed a service for ingesting data
data management component (Conduit) is hos
portal. The ingest service is available as an AP
provides guided input process for data descriptio
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 DCA
and extends the schema for DMA use cases. This
international data portals and ensures that
companies.

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

The Data-Catalogue consists of descriptions of d
cluster for a particular topic or company. A Datas
a single source, and available for access or downl

All the metadata fields in the DMA-Metadata C
receiver of data is able to process information ab
must provide information about instances of the

An overview of the DMA-Metadata Core metadata

Data Catalogue

Identifier	Definiton/Description	
Datasets	This property links the Catalo Catalogue	
Main Description	Describes the content of the Data property can be repeated for parallel language versions of the description.	1
Publisher	This property refers to an entity (organisation) 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

[4] https://datamarket.gitlab.io/Documentation/

[5] https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe

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[3]. 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 Portal[7] 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

[6] 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.

[7] https://portal.datamarket.at/#/

Quelle: [1] DMA: https://datamarket.at/wp-content/uploads/2020/04/D1.5.pdf

Die Idee: Data Stewardship

IS A
specialist role that utilizes **data governance** processes, policies etc. for administering an organizations' entire data

IS RESPONSIBLE FOR
ensuring the **quality and fitness for purpose** of the organization's **data assets**, including **metadata and**

HAS OBJECTIVE
the **data quality** of the data assets, datasets, data records and data elements including documentation of metainformation for the data, such as definitions, related rules/governance, physical manifestation, and related data models
FOR **reusing** data-related resources

DOES
identification of the data assets and elements which they will steward, with the ultimate result being standards, controls and data entry, **participation** in the **development and implementation** of data assets, identifying owners/custodian's various responsibilities, aiding with **project requirement data facilitation** and documentation of capture rules.

INTERACTS WITH
business glossary standards analysts (for standards), with data modelers (for standards), with DQ analysts (for controls) and with operational teams (good-quality data going in per business rules) while entering data.

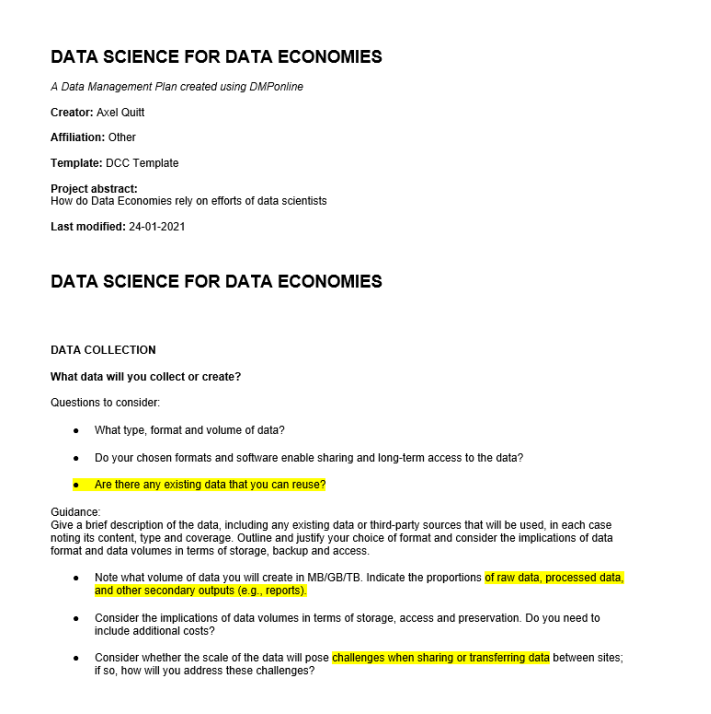
NEEDS
precise purpose, fit for purpose or fitness.

DIO

Der Input – moderiert u.a. durch Data Steward

Kopiere Fragen ins Formular und Download

Setze Schwer-punkte (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 formatsOpens in a new window or DataONE Best Practices for file formatsOpens 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:

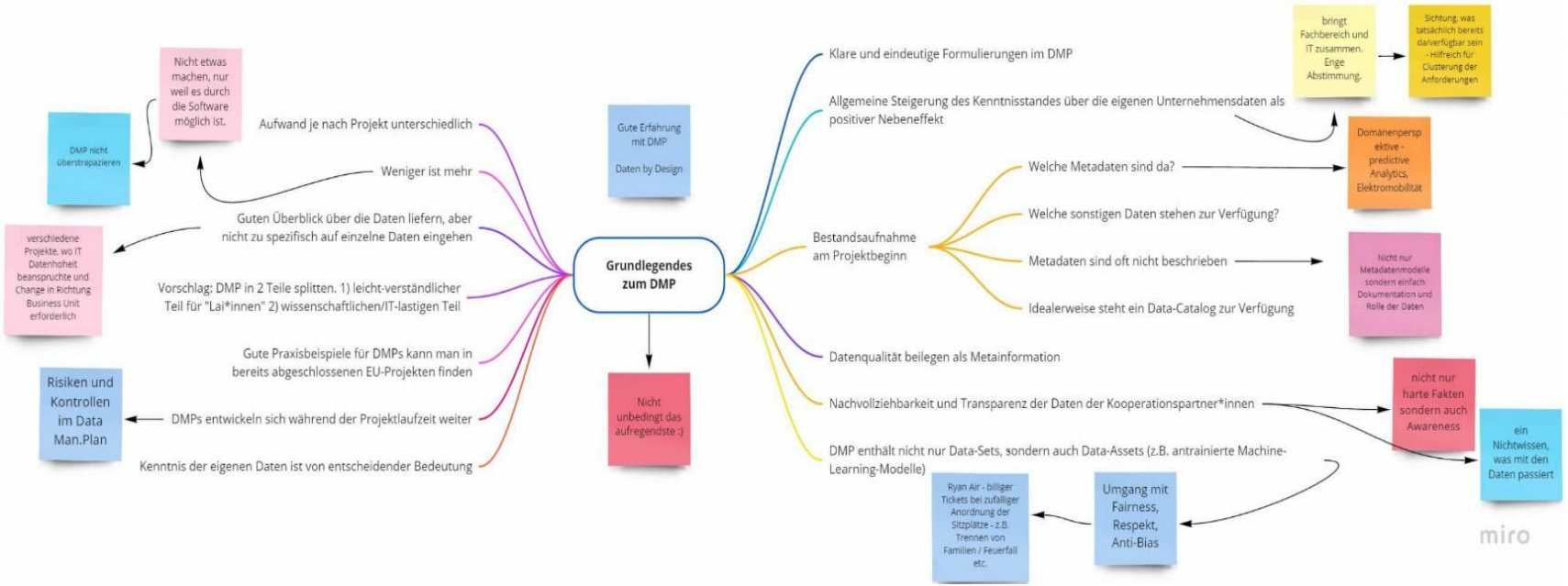
- What standards or methodologies will you use?
- How will you structure and name your folders and files?
- How will you handle versioning?
- What quality assurance processes will you adopt?

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 qualityOpens in a new window.

DIO

Das Ergebnis: eine Wissenslandkarte zum DMP



DIO

Das Wichtigste: Kontinuierliche Zusammenarbeit mit Forschenden

FAQs zum Data Management Plan im Rahmen der IKT der Zukunft Ausschreibung 2020

Untenstehend finden Sie einen Überblick der wichtigsten Fragen aus den bereits abgehaltenen Workshops. Der Beitrag wird laufend mit FAQs aus den kommenden Workshops und dem iterativen Data Coaching und dem individuellen Data Consulting aktualisiert.

− Gibt es eine Ausfüllhilfe für den Data Management-Plan?

Axel Quitt: Das Tool DMPonline bietet eine gute Data Management Plan-Struktur inkl. Ausfüllhilfe. Diverse Fragen und Richtlinien unterstützen bei der DMP-Erstellung. Aber der Content auf DMPonline ist ausschließlich auf Englisch.

+ Kann ich der DMPonline-Cloud vertrauen?

+ Wie sicher sind die Daten, die ich z.B. auf DMPonline eintrage? Was geschieht damit und wer hat Zugriff darauf?

+ Gibt es auf DMPonline Beispielpläne?

+ Gibt es Beispiele für Datenlandkarten?

+ Gibt es eine Vorgabe bzgl. der Länge des Data Management Plans?

Hinweise und Beobachtungen zu Data Management vom DIO Data Steward

12. März 2021 | IKT der Zukunft, News

Zum Thema Data Management und Erstellung von DM-Plänen für die laufende Ausschreibung IKT der Zukunft, lassen sich aus den Anfragen und Diskussionen mit Antragsteller*innen einige allgemeine Themen identifizieren zu denen DIO Data Steward/Coach Axel Quitt folgende Empfehlungen veröffentlicht:

Struktur des Datenmanagement-Plans

Wie in den Workshops besprochen, ist kein bestimmtes Template für den DMP vorgegeben. Es ist bereits ersichtlich, dass unterschiedliche Vorlagen benutzt werden. Bei Verwendung jeder Vorlage empfehle ich drei Dinge:

1. AN DEN LESER DENKEN
2. AN DEN LESER DENKEN
3. AN DEN LESER DENKEN

Eine klare Kapitelstruktur, mit erkennbaren Überschriften, Kapitelnummern und Absätzen, eventuell auch absichtlich gesetzten Leerseiten macht das Dokument für den Adressaten wesentlich besser lesbar und verständlich. Bitte auch ein Inhaltsverzeichnis berücksichtigen, sodass man per Link oder Menü zu einzelnen Kapiteln springen kann.

Lesbarkeit des Dokuments

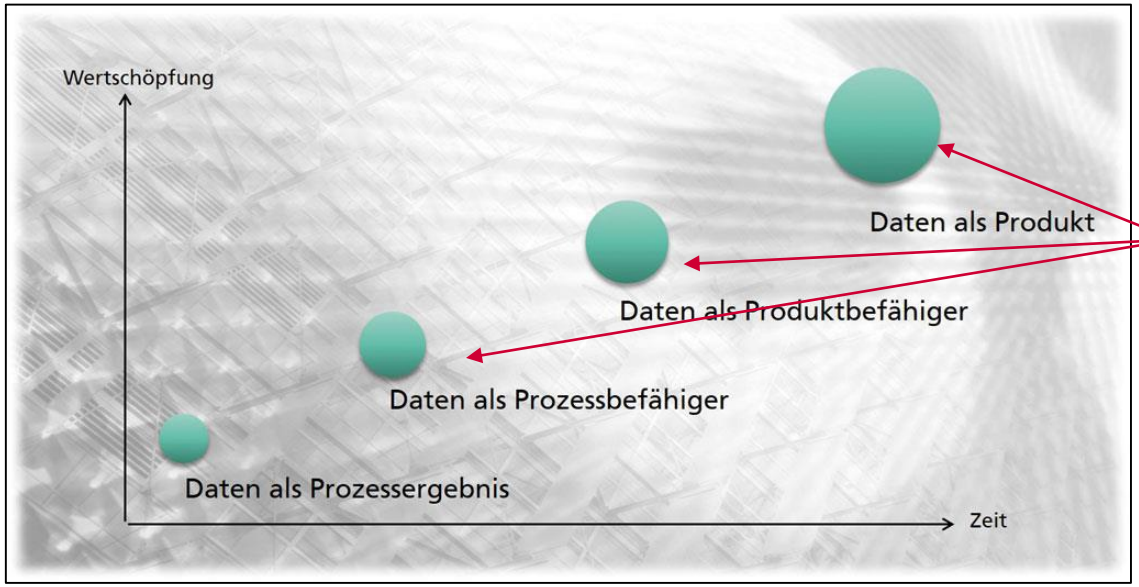
Abkürzungen bitte immer in einem kurzen Absatz erläutern, am besten bei der erstmaligen Verwendung der Abkürzung

Schätzen

Viele Details sind selbstverständlich bei Antragstellung noch nicht bekannt. Wo es möglich ist, sollten aber zumindestens Schätzungen oder Angaben von Größenordnungen im DMP angeführt werden.

Entscheidungen

Fazit: Alte und Neue Anforderungen an Data Stewards



Neue Aufgabenge-biete Data Steward

DIO
Data Intelligence Offensive
Verein zur Förderung der Datenwirtschaft

Diskussion

Data Steward Ausprägungen

- Data object data steward
  - responsible for managing reference data and attributes of one business data entity
- Business data steward
  - responsible for managing critical data, both reference and transactional, created or used by one business function
- Process data steward
  - responsible for managing data across one business process
- System data steward
  - responsible for managing data for at least one IT system

DIO