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)

Executive Summary

This document is the final version of the Data Management Plan (DMP) of the DMA lighthouse project. This update was released in project month 36. The DMP describes the data that DMA collected and generated, how it will be exploited, and how it is curated and preserved after the end of the project.

The central vision of the DMA project was to create a basis for an **ecosystem of federated data and service infrastructures** (*Data-Services Ecosystem*) making data from various Austrian data providers accessible and interoperable. The DMA project researched an approach reducing the centralized components to a minimum and emphasising a distributed peer-to-peer architecture. The goal was to give the participating nodes the highest autonomy possible. The metadata catalogue, as a central component, harvests the metadata from the nodes. The responsibility for keeping the metadata up to date is in the hands of the organisation providing the data. The metadata schemas to be applied are defined in the DMA metadata core.

The experimental solution provided by DMA uses a distributed index (the blockchain). The role of the blockchain is to serve as a common transparent trust basis. The distributed ledger is available to all partners and the DMA members decide which information must be shared in that ledger. Agents (i.e. users and organisations), data sets, and services bear a blockchain identifier. Any important event (e.g. registration of a service or data set) can be recorded in the blockchain in an auditable manner. 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.

DMA includes, closed, semi-closed, and open data from Austrian data providers. Data providers issued an agreement on which data will remain closed or semi-closed data, and how the data will be used during and after the project. Closed and semi-closed data was not generally shared during or after the project runtime, but only according to (bi-)lateral terms of services, expressed by access and usage rights in the blockchain. Detailed information about datasets included in DMA was collected in the DMA-Consortium's Data Catalogue.

Risks related to data security, ethical, and legal aspects are discussed in this document. Appropriate measures were defined and implemented.DMA is only responsible for storing and preserving the ingested metadata. The underlying open, closed, and semi-closed data from the data providers will not be stored on DMA. Data providers are responsible for storing, backing up, archiving, and preserving their data according to their own SLAs.

Table of Contents

List of Abbreviations

DMA	Data Market Austria
DMP	Data Management Plan

Introduction

This version of the Data Management Plan (DMP) is the third iteration released in project month 36. The document has been created by AIT, the project partner in charge of the project data management task (T1.4) in consultation with all other project partners. The DMP describes the data that DMA collected and generated, how it was exploited, and how it will be curated and preserved beyond the duration of the project. This DMP complies with H2020 requirements [1].

The consortium partner **AIT** was responsible for implementing the Data Management Plan (hereinafter: DMP) and ensured that it was reviewed and revised during the project runtime. New versions of the DMP were created whenever important changes to the project occurred. With the end of the project, the final version of the DMP is released. DMP updates during the project runtime:

- D1.1: *Initial Data Management Plan* [M6]
- D1.3: *Updated Data Management Plan* [M18]
- D1.5: *Final Data Management Plan* [M36]

After the end of the project it is foreseen to keep the DMA platform prototype running for 6 more months. The hereinafter described activities and responsibilities by project partners will remain in effect for this duration.

1 Data Summary

The central vision of the DMA project was to create a basis for an **ecosystem of federated data and service infrastructures** (*Data-Services Ecosystem*) making data from various Austrian data providers accessible and interoperable. To this end, the DMA platform extracted existing metadata of the datasets to be included in DMA and transformed it into the DMA metadata schema from the various data providers. The dataset profiles were semantically enriched and stored in the DMA metadata catalogue.

The purpose of the DMA reference implementation is to allow end users to process and analyse **open, closed, and semi-closed data from third-party sources**. Owner and access control information is stored on a distributed **blockchain**. Data providers can either set up an own node, which allows them full control and management over access keys for their data sets. An alternative 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 it only stores and crawls metadata. Datasets will never be stored in DMA in cases where 1) it is technically not possible to transfer data (e.g. due to file size); 2) the data provider does not want to transfer its data to the DMA (but rather provides access to it through an API); 3) it does not make sense to transfer the data to DMA (e.g. it is already available open data; SLAs by public providers are in place). The system allows to store/replicate (encrypted) datasets on distributed storage nodes. However, this functionality as well as an encryption system have not been implemented in the current prototype[1]
.

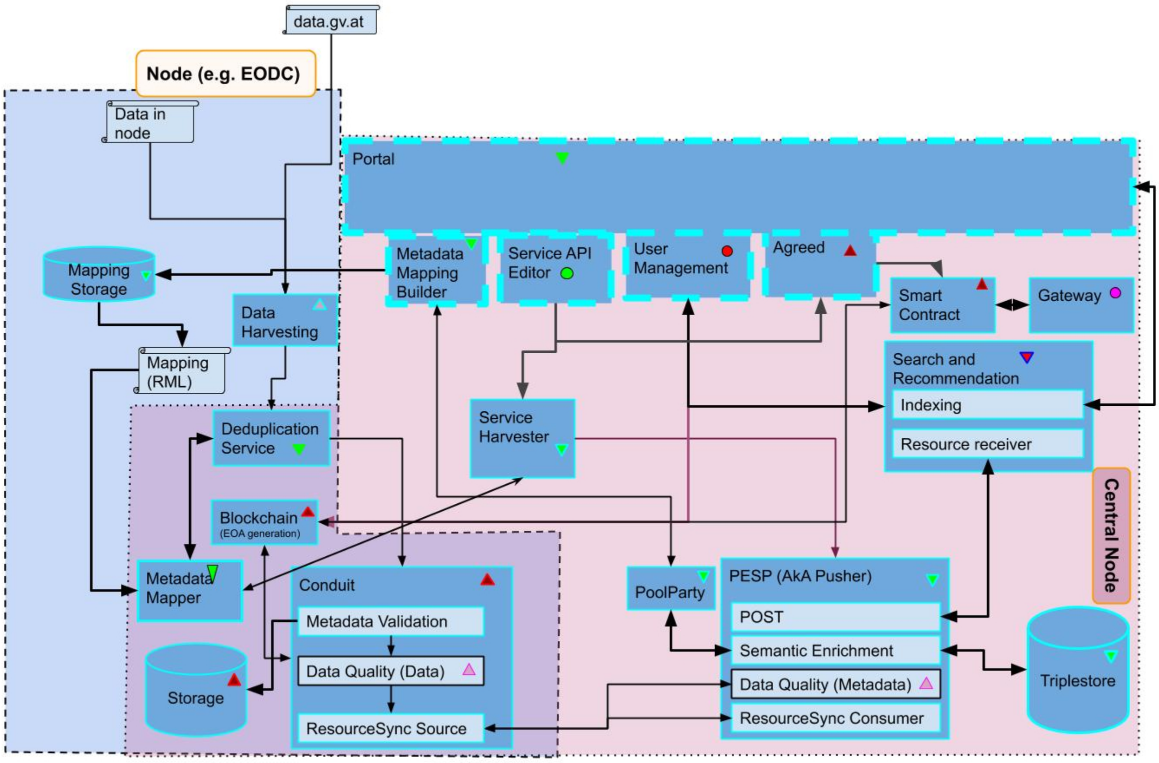

Figure 1 : DMA Components Architecture

Combining data from heterogeneous data sources becomes increasingly challenging the more data owners, licenses, usage rights, terms of service, service level agreements, and regulatory measures such as restricting access to private data are involved. The DMA project uses the Blockchain technology in the following areas:

Unique identification of data assets, services and agents based on blockchain addresses (Ethereum Externally Owned Accounts[2]).

Data asset provenance by capturing important events, such as the creation or modification of a data asset.

Membership voting for managing the membership application process of a

[1] An encryption system can be implemented upon request (e.g. if data providers want to offer such data via DMA). Data providers can provide encrypted data and explanation of how data can be decrypted by the customer.

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

candidate.
Contract conclusion between data or service providers and the DMA customers.

1.1 Data types and origin

The following **data types** are included in the Ecosystem:

A. **Owner and access control information** stored on the distributed blockchain;

B. **Metadata (string)** describing services and datasets.

A list of the datasets included in the Ecosystem for use during the project is provided below. More detailed information is collected in the DMA-Consortium's Data Catalogue [3]. Information about data format, data origin and generation, and the size of data is collected. Metadata of existing open data sources, national and European, were crawled.

1.1.1 List of Datasets

A table listing the proprietary data, of which the aggregated information (metadata) is provided through DMA services, is listed in Table 1. Some of the data can be made openly accessible; other data is restricted or non-public[3] (see Table 1). Upon request, the consortium will consider granting access to some of the restricted data under certain conditions and on a case by case basis (e.g. as was done for the startup calls).

Data type	**Descriptive name**	**Data provider**	**Description**	**Access rights**
Mobility Data	Taxi fleet data	Taxi 40100	**GPS Reports** ZEIT_VON: Datum/Zeit; Begin of time interval ZEIT_BIS: Datum/Zeit; End of time interval MIT_AUSWERTUNGEN: Boolean; With processing (see below) NUR_GPSMESSUNGEN: Boolean; Include vehicles with GPS receiver (should be 1) KEINE_GPSMESSUNGEN: Boolean; Include vehicles without GPS receiver (should be 0) WINKEL_MODUS: Nummer; Angle report mode (see below) MIT_FAHRTRICHTUNG: Boolean; Include GPS direction and velocity MIT_GPSQUALITAET: Boolean; Include GPS quality indicators NUR_STATUS: String; Only include vehicles in specifies state(s)	Non-public

[3] The DCAT standard distinguish between the :restricted and :non-public datasets as follows:
"A restricted dataset is one only available under certain conditions or to certain audiences (such as researchers who sign a waiver). A non-public dataset is one that could never be made available to the public for privacy, security, or other reasons as determined by your agency." The list of access restrictions has to be provided with the metadata. A possible list of access restrictions: registration required (non-discriminatory); authorisation required ("closed data", that only authorized users can access). For the DMA a non-public dataset can be also excluded from the search options and is only available if the data owners offer the dataset directly to the customer.

			NUR_FAHRZEUGFLOTTE: Nummer; n/a	
	Energy Transformers Dataset	SIEMENS	Data from energy transformers in the Aspern neighborhood. Number of transformers: 24 Area covered: Seestadt Aspern (full) Available timespan: Jan 2016 – today Time granularity: 2.5 mins Measurements: Current (I), Voltage (V), Phase (cos phi), Active Power (P), Reactive Power (Q) - x 3 (three phase)	Non-public
	Mobility data	T-MOBILE	Number of people (based on extrapolated number of active subscriber) per 500x500m grid cell. Grid Info:Structure:grid_id INT, Grid cell idgrid_geom_wgs84 STRING, WKT string of the grid cell polygon in WGS84Grid Data:Structure:grid_id INT, Grid cell idtime_window_start STRING, Window starting time in minutes CETtime_window_end STRING, Window ending time in minutes CET ,num_people BIGINT, Number of people extrapolated from number of active subscribers Time Duration: 01.09.2017 – 31.09.2017 Window Size: 15min	Restricted
Weather and Climate Data	Climate reference map	ZAMG	interpolated Austrian climate data from 1961-1990, based on measured values: temperature, cloudiness, humidity, precipitation, duration of sunshine, snow depth	Public
	snow data calculated with the SNOWGRID Snow Cover Model Austria	ZAMG	physically-based and spatially distributed snow cover model that is driven with gridded meteorological input data of the integrated nowcasting model INCA using remote sensing and radar data as well as ground observations. Output: snow height and snow water equivalent maps in a spatial resolution of 100 m and a time resolution of 15 minutes in near real-time	Restricted
	Snow chemistry data from Austrian Glaciers	ZAMG	Snow chemistry data from glaciers in the Austrian Sonnblick area: Goldbergkees (ab 1987), Wurtenkees (1983-2012), Kleinfleißkees (2013-). Concentrations of sulphate, nitrate, ammonium, calcium, kalium, potassium, sodium, chloride, pH and conductivity	Restricted
	Daily weather maps from Austria	ZAMG	Daily weather maps from 1865, ongoing (current maps and digitised historic maps)	Non-public
	UV index	ZAMG	Daily UV-index graphs showing the intensity of UV radiation causing sunburn, based on forecasting models by DWD and ZAMG	Non-public
	extreme value weather data in	ZAMG	monthly values of weather conditions in Austrian provincial capitals: day minimum and maximum temperature, day maximum precipitation, day maximum fresh snow, maximum height of snow	Non-public

	Austrian provincial capitals		cover, count of days with snow cover, sum of precipitation, sum of sunshine duration, count of tropical days, count of summer days, count of freezing days, count of ice days	
	measured values of WMO essential TAWES weather stations	ZAMG	current hourly values for essential weather stations according to WMO: temperature, dew point, relative humidity, wind direction, wind speed, gust of wind, precipitation, air pressure at station, air pressure reduced to mean sea level, sunshine duration	Public
	weather forecast data	ZAMG	forecasted weather conditions in numbers and texts for the next week in Austria, per province	Non-public
Earth Observation Data	Sentinel satellite data	ZAMG	national mirror: two-weekday rolling archive of data from European Sentinel satellites: e.g. synthetic aperture radar, land and sea temperature, multispectral data	Non-public
	Earthquakes in Austria and world-wide	ZAMG	datasets to earthquakes registered by Austrian seismological service, including coordinates, focal depth, magnitude and epicentre.	Public
	Seismograms	ZAMG	Seismograms of ground vibrations as registered by the stations of the Austrian seismological service.	Non-public
	live seismic data of the Conrad observatory in Lower Austria	ZAMG	live seismograms of ground vibrations registered by the Conrad observatory in Lower Austria	Non-public
	Daily magnetogram - geomagnetic variation	ZAMG	horizontal magnetic field component H. Below, declination (D) and vertical component (Z) of the local magnetic field	Non-public
	Geomagnetic storms / space weather	ZAMG	Most recent relevant geomagnetic storm from the automatic storm detection module in the Conrad observatory: horizontal magnetic field component H, near real time	Non-public
	Daily gravity variation	ZAMG	earth gravity variation, gravity residuum, air pressure variation measured in the Conrad Observatorium	Non-public
	Sentinel-1 IW data	EODC	**Sentinel-1A IW GRDH** **Sentinel-1B IW GRDH** Copernicus Sentinel-1A/1B Level-1 Ground Range Detected (GRD) high resolution product in interferometric wide swath mode. The product consists of focused SAR data that has been detected, multi-looked and projected to ground range using an Earth ellipsoid model.	Public
	Sentinel-1 EW data	EODC	**Sentinel-1A EW GRDH** **Sentinel-1B EW GRDH**	Public

			Copernicus Sentinel-1A/1B Level-1 Ground Range Detected (GRD) high resolution product in extended wide swath mode. The product consists of focused SAR data that has been detected, multi-looked and projected to ground range using an Earth ellipsoid model.	
	Sentinel-1 SLC data	EODC	**Sentinel-1A IW SLC** **Sentinel-1B IW SLC** Copernicus Sentinel-1A/1B Level-1 Single Look Complex (SLC) products consist of focused SAR data geo-referenced using orbit and attitude data from the satellite and provided in zero-Doppler slant-range geometry. The products include a single look in each dimension using the full transmit signal bandwidth and consist of complex samples preserving the phase information. Data is only available for a predefined region.	Public
	Sentinel-2 MSI data	EODC	**Sentinel-2A L1 MSI** Copernicus Sentinel-2A Level 1 MultiSpectral Instrument products consist of 13 spectral bands, representing a different central wavelength of the observation. Products are a compilation of elementary granules of fixed size, along with a single orbit. A granule is the minimum indivisible partition of a product (containing all possible spectral bands). Granules, also called tiles, are 100x100 km2 ortho-images in UTM/WGS84 projection.	Public
	Sentinel-3 SLSTR data	EODC	**Sentinel-3A SLSTR RBT** The Sea and Land Surface Temperature Radiometer (SLSTR) is a dual scan temperature radiometer onboard of the Copernicus Sentinel-3 operational mission. The products consist of top of the atmosphere (TOA) Radiances and Brightness Temperature.	Public
	Sentinel-3 OLCI data	EODC	**Sentinel-3A OLCI EFR** **Sentinel-3A OLCI ERR** The Ocean and Land Colour Imager (OLCI) instrument an optical instrument with five camera modules onboard of the Copernicus Sentinel-3 operational mission. The products consist of calibrated, ortho-geolocated and spatially re-sampled Top Of Atmosphere (TOA) radiances for the 21 OLCI spectral bands. EFR stands for "full resolution" product ERR stands for "reduced resolution" product	Public
Financial and Legal Data	Company data (basic info -)	COMPASS	**COMPASS Company data** Commercial Register Number Company Status (active, Insolvency,....) Company Name	Restricted

			Company Address Legal Form Commercial Court UID-Number Phone/Fax Former Company Names Company URLs Company E-Mail Longitude/Latiude	
	Economical in-depth Information	COMPASS	**COMPASS Economical in-depth information** Extended company profile as closed data only, liable to fees. Available e.g.: Ersteintrag, Letzteintrag, Sitz, Korrespondenzsprache, Suchworte, OENACE, Hauptbranche, Bankverbindung, Umsatz, Bilanzsumme, EGT, Cash-Flow, Beschäftigte, Eckdaten zur Bilanzeinreichung, Bilanzstichtag, Kapital, Marken, Import/Export, Niederlassungen, Wirtschaftlicher Eigentümer, Eigentümer, Management, Beteiligungen, Eckdaten zu Rechtstatsachen, balance sheets optional.	Non-public

Table 1 : Proprietary datasets

A table listing the **(Linked) Open Data**, of which the aggregated information (metadata) is provided through DMA services, is provided in Table 2.

Descriptive name	**Data source**
Open Data from data.gv.at, opendataportal.at	data.gv.at, opendataportal.at It will be checked if additional data from gip.gv.at, basemap.at, openstreetmap.org is needed for the DMA pilots.
Linked Open Data from linkeddata.gv.at	linkeddata.gv.at
Several taxonomies & code lists & ontologies that could be used by the Ecosystem	e.g. EuroVoc or GEMET or Esco

Table 2 : (Linked) Open Data

2 FAIR Data

2.1 Making data findable, including provisions for metadata

DMA delivered a platform prototype for commercialization of data and services. To make both data and services discoverable, metadata for both is provided. The Data-Services Ecosystem extracts existing metadata of the datasets from the various data providers and transforms it into the DMA metadata schema. The **central DMA node**, which is hosted in the T-Systems cloud, includes the platform and all core services such as user management. DMA metadata, which is harvested by the

DMA data catalogue, is stored on distributed repositories. DMA service metadata are ingested in the catalogue. Documentation of the implemented DMA services is available on the project's GitLab[4].

The project developed a service for ingesting datasets, which itself invokes additional services. The **data management component (Conduit)** is hosted by Catalysts and is fully integrated in the DMA portal. The ingest service is available as an API as well as a GUI for dataset owners. The GUI provides guided input process for data description and publication. Additional services after ingest include metadata validation (formation, size, validation of ownership, etc.), and semantic enrichment of the metadata.

The metadata schema applied in the project is described in the **DMA Metadata Core**, which is based on DCAT. Metadata is converted on-the-fly to the DMA DCAT standard by means of a mapper.

2.1.1 Data Metadata

The DMA metadata catalogue is based on the DCAT- Application Profiles for data portals in Europe[5] and extends the schema for DMA use cases. This standardization enables future cooperation with international data portals and ensures that the DMA is easily accessible for cooperating companies.

The DMA schema is, as mentioned, similar to the DCAT-AP schemas and consists of the Data-Catalogue, Data-Dataset and Service-Metadata entities. The Data-Dataset are additionally separated into Data-Dataset-distribution entities.

The Data-Catalogue consists of descriptions of datasets and provides an overview of the datasets cluster for a particular topic or company. A Dataset is a collection of data, published or curated by a single source, and available for access or download in one or more formats.

All the metadata fields in the DMA-Metadata Core are mandatory classes. This ensures that a receiver of data is able to process information about instances of the class and the provider of data must provide information about instances of the class.

An overview of the DMA-Metadata Core metadata description is given in tables 3 to 6.

Data Catalogue

Identifier	Definiton/Description	Amount
Datasets	This property links the Catalogue with a dataset that is part of the Catalogue	N
Main Description	Describes the content of the Data Catalogue (in a free text field); This 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

Language	This property refers to a language used in the textual metadata describing titles, descriptions, etc. of the Datasets in the Catalogue. This property can be repeated if the metadata is provided in multiple languages.	N
Tags	Tags of the Data Catalogue, defined in a Thesaurus Autocomplete, fixed Vocabulary. At least one TAG or UGT hast to be provided	N
Billing	Provides some information about the billing system used by this catalogue	1
Access rights	This property refers to information that indicates whether the Catalogue is open data, has access restrictions or is not public. A controlled vocabulary with three members (:public, :restricted, :non-public) will be created and maintained by the Publications Office of the EU.	N
User generated Tags	This property contains a keyword or tag describing the Dataset. Free chosen. At least one TAG or UGT hast to be provided	N
Price Model	Provides some information about the pricing system used by this catalogue; this price model is NOT valid for the single datasets included in the catalogue	N

Table 3 : Draft of DMA's Data Catalogue metadata core

Dataset

Identifier	Definiton/Description	Amount
Title	This property contains a name given to the Dataset. This property can be repeated for parallel language versions of the description.	N
Description	This property contains a name given to the Dataset. This property can be repeated for parallel language versions of the description.	1
Publisher	This property refers to an entity (organisation) responsible for making the dataset available.	1
Language	This property refers to a language used in the textual metadata describing titles, descriptions, etc. of the Dataset. This property can be repeated if the metadata is provided in multiple languages.	N
Tags	This property contains a keyword or tag describing the Dataset. Selection from DMA knowledge graph (Thesaurus = controlled vocabulary) only.At least one TAG or UGT hast to be provided	N
User generated Tags	This property contains a keyword or tag describing the Dataset. Free chosen. At least one TAG or UGT hast to be provided	N
Contact point	This property contains contact information that can be used for sending comments about the Dataset.	1
Dataset Distribution	This property links the Dataset to an available Distribution.	N
Theme	This property refers to a category of the Dataset. A Dataset may be associated with multiple themes.	N
Publisher	This property refers to an entity (organisation) responsible for making the Dataset available.	1

Version	This property contains a version number or other version designation of the Dataset.	1
Unique Identifier	This property contains the main identifier for the Dataset, e.g. the URI or other unique identifier in the context of the Catalogue.	1
Access Rights	This property refers to information that indicates whether the Dataset is open data, has access restrictions or is not public. A controlled vocabulary with three members (:public, :restricted, :non-public) will be created and maintained by the Publications Office of the EU.	1

Table 4 : Draft of DMA's Dataset metadata core

Data Distribution

Identifier	**Definiton/Description**	**Amount**
Access URL	This property contains a URL that gives access to a Distribution of the Dataset. The resource at the access URL may contain information about how to get the Dataset.	N
Format	This property defines the formats in which the dataset is available	1
License	Defines the Licence of the Dataset	1
Service Level Definition	A set of SLDs guaranteed by the provider	N
Service Level Agreement	This property refers to the official commitment that prevails between a service provider and a client. Particular aspects of the service – quality, availability, responsibilities – are agreed between the service provider and the service user. SLA includes the SLDs	1
Price Model	Provides some information about the price model used by this dataset (Note: DMA price models have yet to be defined; this property has been introduced to enable different prices for various distributions.)	1
Description	This property contains a free-text account of the Distribution. This property can be repeated for parallel language versions of the description.	N

Table 5 : Draft of DMA's Data Distribution metadata core

2.1.2 Service Metadata

A software service is a tool that is capable of taking the input in a specified format and providing a specified output. Service metadata is needed to make the services discoverable and to include them in the recommender system. [2]

General Service Properties

Identifier	**Definiton/Description**	**Amount**
Description	A description of the DMA service. This property can be repeated for parallel language versions of the description.	1
Publisher	ID of the service owner within the DMA	1

Title	The name identifier of the DMA service. This property can be repeated for parallel language versions of the description.	1
Unique Identifier	Unique Id of the DMA Service	1
Language	This property refers to a language used in the textual metadata describing titles, descriptions, etc. of the Datasets in the Catalogue. This property can be repeated if the metadata is provided in multiple languages.	N
Tags	Tags of the Data Catalog	N
Contact Point	This property contains contact information that can be used for sending comments about the Dataset.	1
License	Term of use	1
Category	Application domain in which the service is located	N
Theme	Data type on which the DMA service is built upon	N
Price Model	Provides some information about the pricing system used by this service	N
Documentation	References to further documentation of the DMA service. This property can be repeated for parallel language versions of the description.	N
Tags	This property contains a keyword or tag describing the Service. Selection from DMA knowledge graph (Thesaurus = controlled vocabulary) only	N
Created	Date and time of DMA service creation (automatically generated by DMA)	1
Version	This property contains a version number or other version designation of the Service	1
Service Type	The type of service a customer can expect (Note: Service types have yet to be defined; there will be a set of service types later on.)	1
Quality of experience rating	The overall acceptability of the DMA service as perceived subjectively by the end-user	N
User generated Tags	This property contains a keyword or tag describing the Dataset. Free chosen	N

Table 6 : Draft of DMA's Service metadata core

2.2 Making data (openly) accessible

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[6]. 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/#/

- Resource provisioning

US3: Data services management

- create a dataservice (similar to dataset)
- create metadata (service outside DMA - not directly available or outside available)
- service inside DMA
- dataservice management (edit, delete)
- machine task: metadata indexing

US4: Search & browse on DMA for logged in user (data, services,other)

- recommendations of data & services (on top of orga / user profile)

US5: Work Space Management

- create workspace
- select datasets

US6: Monitoring

- Raw data logging & Access monitoring
- for DMA (capacity planning, bus models, …)
- for DMA users (statistics on dataset usage, tariffing, etc)

US7: Data Acquisition

- data.gv.at and opendataportal.at
- europeandataportal.eu and other data acquisition

A **Data Clustering and Slicing prototype**[8] was implemented. The service is a non-core service letting users download subsets of datasets based on naturally-occurring clusters. This prototype identifies clusters of geographic coordinates in CSV data sets. The data clusterer component identifies naturally occurring clusters in CSV datasets, while the data slicer component lets data customers download subsets of given datasets based on the identified clusters. The objective of the data clusterer and slicer service is that in case of very large datasets users would not have to purchase, download and process the entire dataset, instead be able to purchase subsets of it, only the ones that are relevant to them (thus saving them money, time and computing resources).

2.4 Data re-use & licensing of data

The DMA project implemented **blockchain technology for provenance**. The blockchain is used to model actors (data providers, service providers, consumers), objects (datasets, services), and transactions (data creation, transformation, and enrichment; service execution) in a blockchain transaction graph. The necessary entities and transactions to define dataset provenance (data lineage) are modelled and stored it in the DMA blockchain, and the service for reading and writing to this blockchain is implemented. The DMA blockchain serves as decentralized offer and contract registry and enables dataset ownership, authenticity and trust via asymmetric encryption/signatures. This in turn provides a transparent, tamper-proof record of all datasets handled by Data Market Austria infrastructure.

Smart contracts based on Ethereum platform are used to allow concluding contracts between

[8] https://slicendice-slicer.datamarket.at/poc/slicer.php

data providers, service providers, and DMA customers. The smart contracts are created through the **DMA contract management component (Agreed)**. Data providers can create assets (products) and offers based on these assets. Contracts are independently created as multiple contracts can be issued for the same offer.

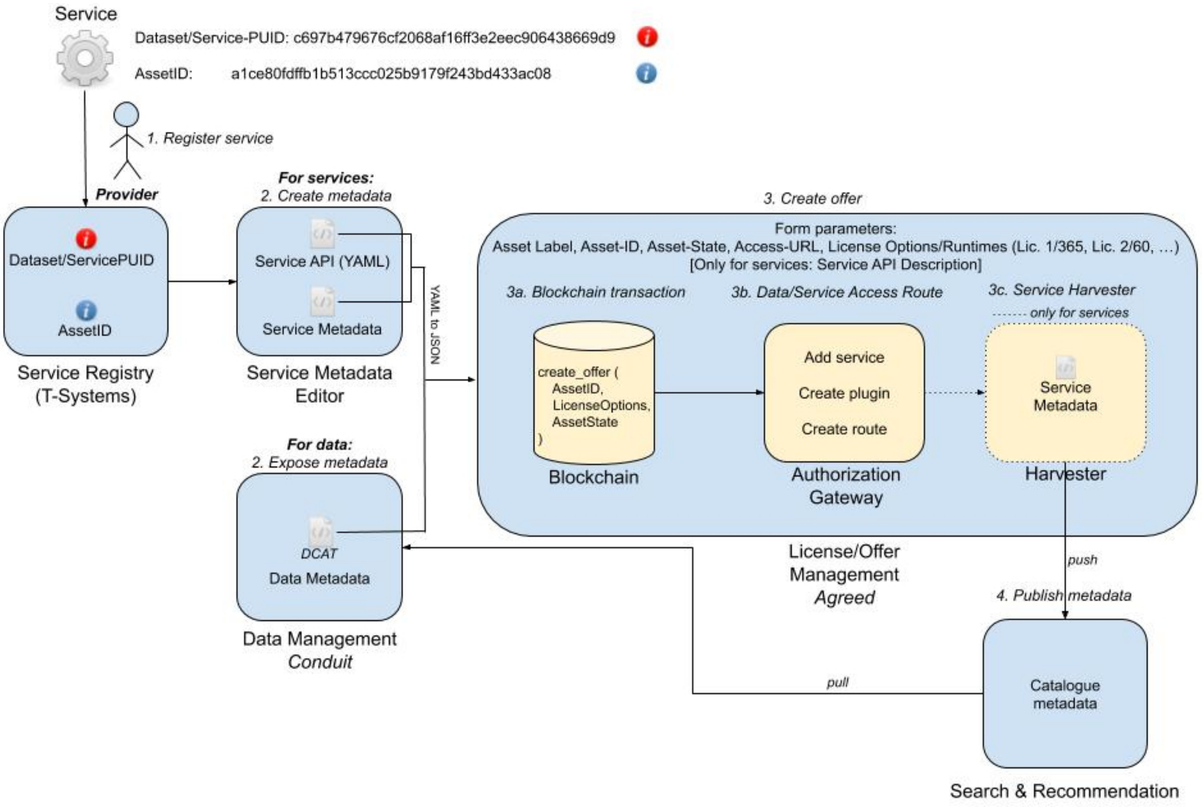

Figure 2: DMA Offer Creation and Metadata Publishing Workflow

Standard licenses for data use and re-use (e.g. Creative Commons licenses) can be defined when datasets are added to DMA. If data providers provide data that is licensed by third parties, they are responsible for disclosing and specifying the licensing terms.

Services for improving the quality of submitted datasets were provided and are available as open source components. However, they are not implemented in the DMA platform prototype. This includes automated tools that identify issues in CSV files, e.g. missing irregularity and encoding issues, and tools for automatic normalisation of data entities like mapping to common date or time representations and numeric formats. These tools will be drawn from a researched repository approved of data cleaning/modification patterns for defined data/content types. An application developed in this task will also provide users with a manual mechanism to perform changes to datasets, including branching and forking functionality.

3 Allocation of Resources

3.1 Estimated Costs

Data Management related costs that have occurred in the DMA project were 1) costs for data management by data providers (data provision and hosting), 2) infrastructure hosting (central instance), and 3) services hosting.

The estimation of costs that occurred for data management at our data providing partners is in total 38.000EUR. This includes both data provision and hosting of the data at ZAMG. EODC, T-MOBILE and COMPASS did not have any costs that can be attributed directly to the project.

For hosting the DMA central instance at T-SYSTEMS, the costs were 3.200EUR per month. For the last fifteen months, this makes an estimated total for infrastructure costs of about 48.000EUR.

For hosting the DMA core services at ZAMG and CATALYSTS we estimated a total of 2.500EUR per month. For the last fifteen months the total estimation of core services hosting costs was about 38.000EUR.

After the end of the project, the DMA central instance and metadata catalogue will be kept up and running for six months. Foreseen costs match the costs which occurred previously in the project. For this additional duration after project end, the costs will be un-funded. No additional costs for storage and backup services for the DMA metadata catalogue are foreseen. No further costs are required for preparing data archiving or re-use after the project.

The platform basic services will be provided at no charge for the duration of six months after the end of the project. This does, however, not preclude pilot service brokers to provide added-value services on a fee-basis. How this will be handled in future will also be an outcome of the business model considerations (D3.3, D3.5).

3.2 Responsibilities

DMA is only responsible for storing and preserving the ingested metadata. The underlying open, closed, and semi-closed data from the data providers will not be stored. The **data providers** will be responsible for storing, backing up, archiving, and preserving their data.

During the project runtime, the consortium partner **AIT** was responsible for implementing, reviewing and revising the DMP. After the end of the project, the DMA platform prototype will be kept running for 6 additional months. During this period and depending on the continuation of DMA after the end of funding through the FFG Lighthouse Project, new responsibilities and roles will be defined. This will be in line with the project's exploitation plan (D3.3, D3.5).

3.3 Long Term Preservation

For long-term logical preservation the collected metadata will be normalised to a CSV format and stored at SWC. No long-term bit preservation measures are foreseen for the collected metadata.

A hashing system is implemented. Each data set has a random identifier. Namespaces, identifier format, and quality services are specified in D5.3. A PID schema was specified in D5.1; however, it was not implemented in DMA. Each data provider is responsible to create unique identifiers as needed.

4 Data Security

The experimental solution implemented by DMA makes use of a distributed index (the blockchain). Transaction data are always fully replicated via the blockchain nodes. Replication addresses the problem of bit preservation (maintaining integrity of bits). In the current beta phase no additional replication or backup solutions are implemented. Data providers are responsible to take care of backup strategies on their own. A second approach towards logical preservation is **emulation**, and the use of service virtualisation implementation (using Docker[9]) is an approach that ensures software reproducibility in a manner very similar to emulation.

5 Legal and ethical Aspects

This section addresses questions related to ethics and legal compliance of the included datasets and define how ethical issues and IPR are managed in the project.

5.1 Data Protection

Ethically questionable material or personal data will not be provided or stored on DMA. Whenever personal data is processed, the compliance with the principles of data protection are to be proven by the controller. These principles encompass, for instance, data minimisation, meaning to only process the data necessary for the pursued purpose. **Privacy by design** indicates to create data processing technically already in favour of strong protection of personal data. To name one technical design decision of the DMA, not to store data centrally but to have the data transfer handled between data provider and customer supports the data minimisation principle.

Through a broad definition of data, possible transaction objects within the DMA should be restricted as little as possible. However, due to the high amount of data protection requirements, personal data as subject matter is not expected to be the main application scenario of future trade in data within the DMA. For the portal pilot phase and the duration of the start-up call, we explicitly exclude the use of any personal data.

5.2 Measures to ensure ethical and legal standards

Measures to ensure compliance to ethical and legal standards have been developed by the WP3 team. With the aims to provide a **standardised model-contract** and to increase legal certainty within the legal relation of the Data Market Provider to the Data Market Customer, a model data license was developed. The implementation through the contract management tool enables Data Market Providers to create licenses which are compliant with the standardised model-contract and using them when creating offers.

Further, the contractual framework the brokers will be embedded in was investigated.

All WP 3-members and other consortium members were involved in stakeholder-workshops, which were held to discuss parameterisation of the **model-data license** with potential providers and customers. This license is based on the assumption that Data Market Providers disclose information like if personal data is concerned or if intellectual property rights of third persons are involved. The DMA cannot verify this information but can provide guidelines meant to serve the

[9] https://www.docker.com/

Data Market Providers as - from the perspective of the DMA - non-binding support. These guidelines are envisaged as one part of the measures, which will be taken to “accompany” the Start-Ups and companies selected because of their submissions to the FFG-Call. INITS defined processes, which supported the companies for their project work.

A **Code of Coduct (“Netiquette”)** taking into account the specific roles of the DMA-operator, Data Market Provider, Data Market Customer, Infrastructure Provider, and Broker provides drafted guidelines for desirable and undesirable interaction on the DMA. Another idea is a facultative **certification** that could concern the DMA, Brokers, Data Market Providers or Infrastructure Providers.

5.3 Privacy and trust

Issues of **privacy and trust** amongst data trading participants have been identified during the first stage of Task 3.4 as being potential significant impediments to the successful uptake of the data trading platform. Solutions to mitigate their negative effects have been proposed based on an in-depth review of scholarly and practitioner literature (refer to D3.1). In particular, a number of key indicators leading to negative levels of privacy and trust were outlined:

an inconsistent level of protection for natural persons and private data;
divergences in the handling and storage of data hampering the free movement of personal data within the internal market;
a lack of knowledge regarding data sharing;
difficulties in determining the trustworthiness of data suppliers;
lack of knowledge of the law leading to potential violations;
and inconsistent levels of protection for members across participating organisations.

5.4 Survey and data collection in Task 3.4

To elicit opinions and perspectives of participants involved with the DMA project, research data was collected in the form of primary data directly from individual actors and representatives of participating organisations. This stage of the task has resulted in the collection of non-sensitive personal data. Participating survey respondents, and the organisations that they represent, together with the data they generate, were handled with care. Stringent measures were undertaken to protect participant anonymity and to preserve data integrity. The following principles and practices to ensure the ethical collection, handling, and storage of data were applied:

1. **Ethical Recruitment of Questionnaire Respondents**: To maximise data quality and participant satisfaction, and to protect fundamental rights and participant dignity, the ethical recruitment of participants on a voluntary, non-discriminatory basis must be practiced.

2. **Informed Consent of Subjects**: Prior to administering the questionnaire, or any activity concerned with primary data collection, it is recommended that written permission be sought from each participant separately, where possible. Participants must be given the option to opt out of the process at any time without consequence.

3. **Adherence to Data Processing Standards**: Project researchers must adhere to recognized national, EU, and international standards when collecting, storing, analyzing, and disseminating

sensitive personal, business, or political data for this task.

4. **Ethical Use and Dissemination of Results**: All forms of collected data and research results must continue to be used and disseminated ethically. Project researchers will continue to be bound to the DMA consortium's agreement to tender for internal review all conference, workshop, and publication proposals prior to their submission.

5. **GDPR Compliance:** All research activities associated with this task carried out in the GDPR post-enforcement period will be conducted in a manner so as to preserve the rights and obligations enshrined in this regulation.

6 References

[1] H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020. Version 3.0, 26 July 2016.

[2] Johann Höchtl et al., D6.1 Service Technology Specification and Development Roadmap. Final version, 31 May 2017. https://datamarket.at/wp-content/uploads/2017/10/DMA_Deliverable_D6.1_FINAL_v01.pdf

[3] DMA 2018, DMA-Consortium's Data Catalogue: https://docs.google.com/spreadsheets/d/1HvPZiCp1xWYC8oeFuDwVf5KqtCOIKZ51nPtvQfKewpQ/edit#gid=1336034758