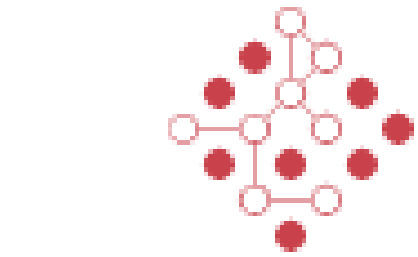

DATA MARKET

AUSTRIA

www.datamarket.at

Data Management Plan

Deliverable number	*D1.3*
Dissemination level	*Public*
Delivery date	*2017-03-31*
Status	*Updated Data Management Plan*
Author(s)	*Michela Vignoli*

FFG

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 second, updated version of the Data Management Plan (DMP) of the DMA lighthouse project. This update was released in project month 18. The DMP describes the data that DMA collects and generates, how it will be exploited, and how it will be curated and preserved. This DMP is a **living document** and will be updated whenever relevant changes occur to the project. One final update of the DMP will be released at the end of the project (month 36). The consortium partner **AIT** is responsible for implementing the DMP and ensures that it is reviewed and revised during the project runtime.

In the DMA project we intend to research on an approach that **reduces the centralized components to a minimum and emphasises a distributed peer-to-peer architecture.** The goal is to give the participating nodes the highest autonomy possible. Metadata of data sets, for example, are stored on the central node and managed on the participating node, exposed using a standardized interface which also informs about metadata updates[1]. 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.

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 events (e.g. registration of a service or data set) can be recorded in the blockchain in a auditable manner.

The **metadata schema** to be applied in the project are defined in the **DMA metadata core**.

The Data-Service Ecosystem will include not only open data, but closed and semi-closed data as well. The data providers will issue 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 will not be 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.**

The experimental solution envisioned by DMA proposes the use of a distributed index (the blockchain). In terms of preservation, DMA will pursue **data replication** with an additional layer of **security through encryption and verification** of data through blockchain transaction analysis (secure multiparty computation).

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

In the current DMP version, an extended **list of the datasets** planned to be included in the ecosystem is provided. More detailed information is collected and will constantly be updated in the Preliminary DMA-Consortium's Data Catalogue.

A list of user stories and sub-elements related to **interoperability of data** is provided.

Risks related to data security, ethical, and legal aspects are addressed and appropriate measures

[1] It is still being investigated to what extent this approach is also applicable to service metadata, i.e. if metadata of services are going to be exposed on the participating nodes and harvested as well, or if they should be registered centrally.

defined.

Table of Contents

Introduction 5

Data Summary 5

Data types and origin 5

List of Datasets 6

FAIR Data 10

Making data findable, including provisions for metadata 10

Data Metadata **Fehler! Textmarke nicht definiert.**

Data Catalogue 11

Dataset 12

Data Distribution 12

Service Metadata **Fehler! Textmarke nicht definiert.**

General Service Properties 13

Making data openly accessible 14

Making data interoperable 15

Increase data re-use (licensing of data) 16

Allocation of Resources 17

Estimated Costs 17

Responsibilities 17

Long Term Preservation 17

Data Security 18

Ethical Aspects 18

References 20

List of Abbreviations

DMA Data Market Austria

DMP Data Management Plan

Introduction

This version of the Data Management Plan (DMP) is the second iteration released in project month 18. 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 collects and generates, how it will be exploited, and how it will be curated and preserved. Regular check points on the status of the data will ensure that the Data Management Plan is implemented. This DMP complies with H2020 requirements [1].

The consortium partner **AIT** is responsible for implementing the Data Management Plan (hereinafter: DMP) and ensures that it is reviewed and revised during the project runtime. New versions of the DMP will be created whenever important changes to the project occur due to inclusion of new datasets, changes in consortium policies or external factors. At project end, the final version of the DMP will be released and the data management of the DMA platform will be handed over to the future platform administrator.

Planned 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]

1 Data Summary

The central vision of the DMA project is 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 will **extract existing metadata** of the datasets to be included in DMA and **transform it into the DMA metadata schema** from the various data providers. The dataset profiles will be **semantically enriched**. The metadata will be stored in the DMA central node. Some of the metadata will be replicated on the individual nodes for backup.

The Data-Services Ecosystem will allow end users to process and analyse **open, closed, and semi-closed data from third-party sources**. Owner and access control information will be stored on a distributed **blockchain**. **Encrypted datasets** can be replicated on distributed storage nodes.

The DMA will act as a uniform data access platform. In some cases, it only stores and crawls metadata. This is the case when 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).

1.1 Data types and origin

The following **data types** will be included in the Ecosystem:

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

B. **Metadata (string)** describing services and datasets, potentially encrypted;

C. Optionally, **encrypted datasets** replicated on distributed storage nodes.

In the current DMP version an updated **list of the datasets** planned to be included in the Ecosystem is provided. More detailed information is collected and will constantly be updated in the Preliminary DMA-Consortium's Data Catalogue [3]. In particular, information about data format, data origin and generation, and the size of data is collected.

1.1.1 List of Datasets

A table listing the proprietary data, of which the aggregated information (metadata) will be provided through DMA services, is provided in Table 1. Some of the data can be made openly accessible; other data is restricted or non-public[2] (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. In the upcoming months, this will in particular apply to the startup call.

Due to the insolvency of Bouncing Bytes, a new data provider for taxi fleet data, Taxi 40100, has been identified.

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) NUR_FAHRZEUGFLOTTE: Nummer; n/a	Non-public
	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),	Non-public

[2] 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. This is not yet finally decided.

			Active Power (P), Reactive Power (Q) - x 3 (three phase)	
	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 Austrian provincial capitals	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 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	Non-public
	measured values of WMO essential TAWES weather	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

	stations			
	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** 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.	Public
	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	Public

			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.	
	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 Company Address Legal Form Commercial Court UID-Number Phone/Fax Former Company Names Company URLs Company E-Mail Longitude/Latiude	Restricted
	Economical in-depth Information	COMPASS	**COMPASS Economical in-depth information** Extended company profile as closed data only, liable to fees.	Non-public

			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.	

Table 1 : Proprietary datasets

A table listing the **(Linked) Open Data**, of which the aggregated information (metadata) will be 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 will deliver a platform for commercialization of data and services. To make both data and services discoverable, metadata for both will be provided. The Data-Services Ecosystem will extract existing metadata of the datasets to be included in DMA and transform it into the DMA metadata schema from the various data providers. The metadata will be stored in the **central DMA node**.

The project will develop a **service for ingesting datasets**, which itself will invoke additional services. The ingest service will be available as an **API** as well as a **GUI** for dataset owners. The GUI application will provide 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** to be applied in the project are described in the **DMA Metadata Core**, which is based on DCAT (see below). The updated draft metadata core is provided in Tables 3-6.

2.1.1 Data Metadata

The DMA metadata catalogue is based on the DCAT- Application Profiles for data portals in Europe[3]

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

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 be 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 current DMA -Metadata Core V0.1 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
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

The Data-Service Ecosystem will include not only open data, but closed and semi-closed data as well. The data providing project partners will issue an agreement 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 will not be generally shared during or after the project runtime, but only according to the yet to be developed standardised license, whose terms will be expressed and stored in the blockchain.**

The centrally stored metadata of the federated data and services can be queried and accessed via the DMA platform. The user will be redirected to the data providers' platforms to access and download available open datasets. The DMA operator will also have storage available (open cloud infrastructure), where data can be accessed directly on DMA.

In the DMA project we will apply **blockchain technology for data access regulation**, in particular to the closed and semi-closed datasets, where self-executing contracts on the blockchain will be investigated to model even fine-grained data access and data usage arrangements. 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[4]).

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

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.

Traditional access control mechanisms (e.g. central access control policies, password protection) will be considered as fallback solution if the envisaged blockchain solution turns out to be impracticable.

2.3 Making data interoperable

To facilitate data interoperability and discoverability, we decided to provide an overview of the data catalogue on the DMA landing page without registration. The landing page itself 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 will be designed in a manner so that the access to data becomes independent of the type of cloud or infrastructure provider. On the other hand the facettes of the central portal and all other DMA components enforce the rule set on which access to DMA Metadata is granted. 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
  - Resource provisioning
- *US3: Dataservices 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)*

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

    - 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

2.4 Data re-use & licensing of data

If data providers provide data that is **licensed** by third parties, they are responsible for disclosing and specifying the licensing terms. It is currently not foreseen to re-publish already available open datasets on the Ecosystem.

DMA will provide **services for improving the quality of submitted datasets**. These would include 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.

In the DMA project we will apply **blockchain technology for provenance**. The blockchain will be 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 necessary to define dataset provenance (data lineage) will be modelled and stored it in the DMA blockchain and the service for reading and writing to this blockchain implemented. The DMA blockchain will serve as **decentralized dataset registry** and enable dataset ownership, authenticity and trust via asymmetric encryption/signatures. This in turn will provide a transparent, tamper-proof record of all datasets handled by Data Market Austria infrastructure.

In addition, **smart contracts** based on Ethereum platform will be used to allow concluding contracts between data providers, service providers, and consumers.

The traditional **central registry approach** will be considered as fall-back solution if the envisaged blockchain implementation turns out to be impracticable.

The **potential re-use** of the various datasets will also be assessed, e.g. if the choice of technology, the formats, the metadata, and the license are suitable to ensure subsequent use. It needs to be stressed that the Ecosystem will include closed and semi-closed data as well, thus there will be a number of datasets which will not be suitable for re-use.

Currently **no embargo periods** are foreseen for semi-closed data to be published on the Ecosystem. Metadata of existing open data sources, national and European, will be crawled.

3 Allocation of Resources

3.1 Estimated Costs

At the current state of the project, it is not possible to make an estimation of costs that will occur for the data management task during the project runtime. A first cost estimation for DMA services in general is currently being developed in the first version of the business plan (D3.3). Currently, no charges for additional storage and backup services are foreseen. If applicable, estimated costs for additional data management services to be integrated in the platform will be specified in the next version of this document.

In a later update of this document it will be specified how the data storage will be managed after the project runtime, as well as foreseen related costs. Also, it will be specified if further costs are expected to arise for the preparation of data archiving or re-use after the project.

Currently, it is foreseen to provide the platforms base services at no charge. 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).

3.2 Responsibilities

The consortium partner **AIT** is responsible for implementing the DMP and ensures that it is reviewed and revised during the project runtime.

Name and contact details of the person responsible on behalf of the beneficiary AIT during the project runtime:

Michela Vignoli
AIT Austrian Institute of Technology GmbH
Digital Safety and Security Department
Donau-City-Straße 1
1220 Vienna
michela.vignoli@ait.ac.at
+43 50550-4216

DMA will only be responsible for storing, preserving, and backing up 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.

3.3 Long Term Preservation

For **long-term logical preservation** the dataset may be normalised to a CSV format. This depends on its input format type. For **long-term bit preservation** the data may be encrypted and/or distributed to other DMA nodes or cloud services. This depends on the requested service level.

The service assigning and resolving **persistent unique identifiers (PID)** will be used and supported further after the end of the project.

4 Data Security

The experimental solution envisaged by DMA proposes the use of a distributed index (the blockchain). In terms of preservation, DMA will pursue **data replication** with an additional layer of **security through encryption and verification** of data through blockchain transaction analysis (secure multiparty computation). Replication addresses the problem of bit preservation (maintaining integrity of bits); DMA will also consider the problem of logical preservation (understanding the meaning of bits) through **data normalisation services** (the migration solution). A second approach towards logical preservation is **emulation**, and the use of service virtualisation implementation (using Docker[5]) is an approach that ensures software reproducibility in a manner very similar to emulation.

As fall-back solution a traditional approach using the **file system with regular backups** will be taken if the envisaged blockchain solution turns out to be impracticable.

The consortium partner AIT will define which **backup and recovery functionalities** will be implemented and carried out. In a future update, this section will include details on these functionalities as well as incident management and disaster recovery.

It is not foreseen to transfer sensitive data to DMA. The technology to be applied for the trusted data transfer mechanism in DMA will be specified in a later version of this document.

Further risks related to data security, illegal procurement or manipulation of data will be addressed and appropriate measures taken. Especially in terms of sensitive or closed/restrictive data strict rules will apply and the access to such data will be secured and regulated. Unauthorised users will not be granted access to the data.

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. The technical design of the DMA should be in line with the underlying tenor of avoiding or reducing data processing to the extent absolutely necessary.

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] https://www.docker.com/

5.2 Measures to ensure ethical and legal standards

Measures to ensure compliance to ethical and legal standards are currently being 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 is being developed by DUK. Research on the intersection of law and technology will show to what extent the technical implementation of legal provisions will be feasible.

Further, the contractual framework the brokers will be embedded in, is investigated by COMPASS.

TDA provided a draft **Code of Coduct ("Netiquette")** taking into account the specific roles of the DMA-operator, Data Market Provider, Data Market Customer, Infrastructure Provider, and Broker. TDA is also assigned to the task to draft guidelines for a facultative certification. This **certification** could concern the DMA, Brokers, Data Market Providers or Infrastructure Providers.

All consortium members will be asked for feedback to the draft contracts. All WP 3-members will be involved in stakeholder-workshops, which shall be held to discuss parameterisation of the **model-data license** with potential providers and customers. This license will be 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 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, which will be selected because of their submissions to the FFG-Call. INITS has the lead in defining processes, which will help the companies for their project work.

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

The second phase of T3.4 involves the construction and dissemination of a questionnaire to elicit the opinions and perspectives of participants involved with the DMA project. To achieve this research milestone, research data will be collected in the form of primary data directly from individual actors and representatives of participating organisations. It is recognized that this stage of the task may result in the collection of either personal data, business secrets, or data that is politically sensitive in nature. Participating survey respondents, and the organisations that they

represent, together with the data they generate, must be handled with care. Stringent measures must be 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 need be 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.

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