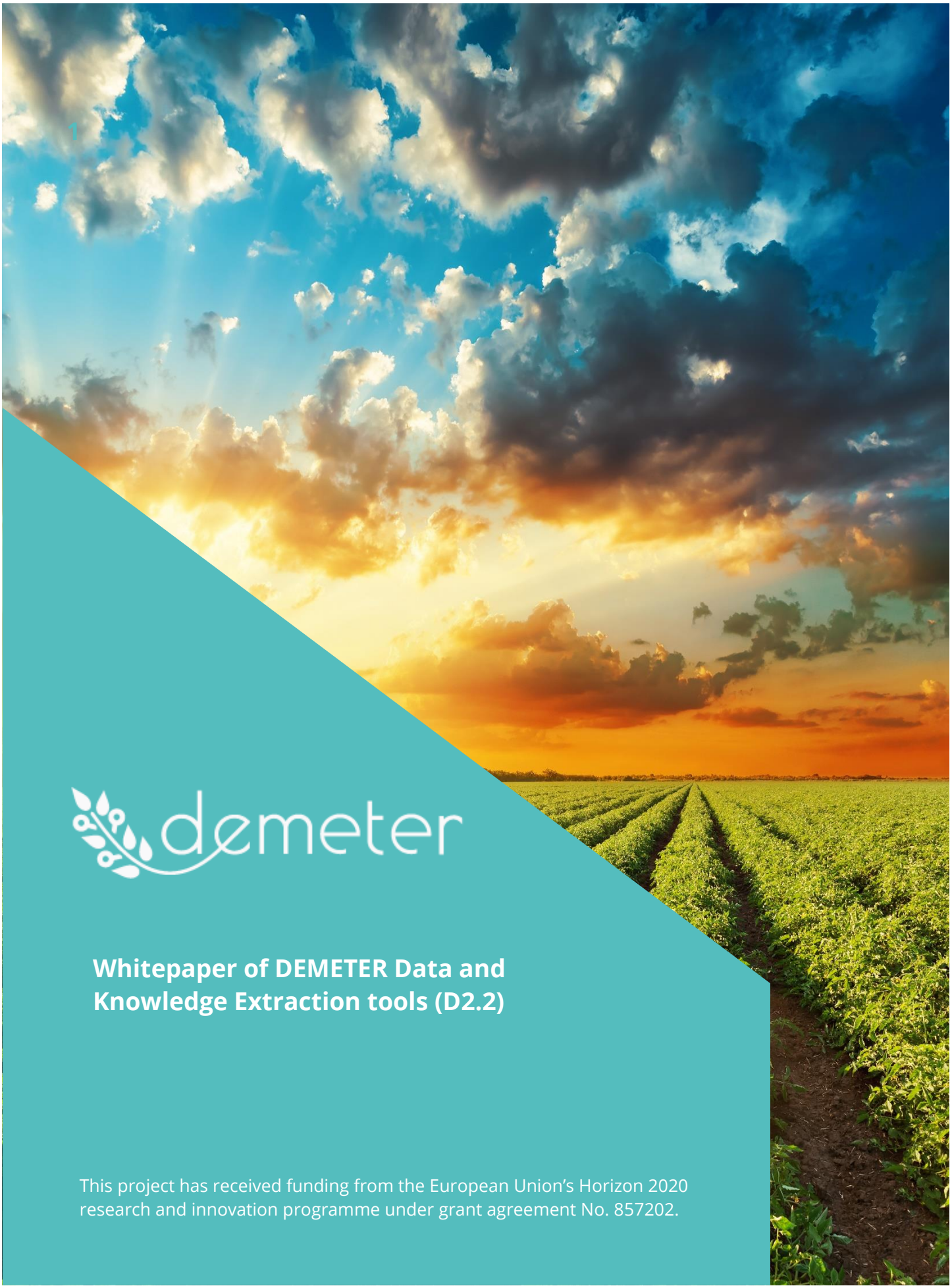

Whitepaper of DEMETER Data and Knowledge Extraction tools (D2.2)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 857202.

DEMETER Data and Knowledge Extraction tools

1 Abstract

DEMETER project aims to lead the Digital Transformation of the European Agri-food sector based on the rapid adoption of advanced technologies, such as Internet of Things, Artificial Intelligence, Big Data, Decision Support, Benchmarking, Earth Observation, etc. in order to increase the performance in various aspects of farming operations and to ensure the sustainability and viability of the agricultural sector for a longer term. By adopting these technologies and facilitating an interoperable data driven platform providing decision support and control systems form the Agri sector which can not only empower the farmer but also allows them to harness more efficiently their own data and knowledge as well as those shared with others, thereby constantly focusing on mixing human knowledge and expertise with digital information systems. DEMETER also focuses on interoperability as the main digital enabler by connecting farmers and advisors with providers of ICT solutions and machinery. To address these objectives and to promote the targeted technological, business and socio-economic impacts, DEMETER project delivers a Reference Architecture (RA) which is suitable to address the specific challenges in the agri-food domain. This architecture being the main technological backbone of the DEMETER project facilitates the collection, processing of the data generated by the various DEMETER pilots and provides an integrated view over all these different heterogeneous datasets in order to extract knowledge as well as in decision making purposes for the farmers and other stakeholders. The deliverable “D2.2 DEMETER Data and knowledge extraction tools'' focused mainly over the tasks 2.2 (Data Management and Integration), 2.3 (Targeted data fusion, analytics and knowledge extraction) and 2.4 (Data Protection, Privacy, Traceability and Governance Management) from WP2. As a summary of the deliverable this document gives a brief overview of the various tools and techniques and state of the art used in the process of the knowledge extraction from the input data sources and also the technological aspect of the data security of these extracted datasets.

2 Technological summary

The results presented in Deliverable "D2.2 DEMETER Data and knowledge extraction tools" realizes the implementation of the Data & Knowledge (DK) enablers, which are part of the advanced enablers in DEMETER. In particular, the DK enablers include the following facilities:

- Data collection & preparation in order to collect, curate and prepare the input data
- Integration & linking of Data to provide an integrated view over different heterogeneous data sources
- Data fusion to fuse the data collected and analysed
- Data management to support the users' specific preferences
- Data analytics & knowledge extraction for further processing of the fused and integrated data.

Within the DEMETER project the above-mentioned facilities are mapped into three enablers: i) Data Preparation & Integration Enabler ii) Data Management Enabler iii) Data Analytics & Knowledge Extraction Enabler which are presented in the deliverable. The deliverable also covered the data security aspects of the generated datasets over various DEMETER enabled pilots. The data security facilities also involve some Core and Advanced Enablers which aim for example to ensure secure transfer of sensitive data or to prevent access to unauthorized entities. The implementation of all these different enablers are controlled by one of the most important elements of the DEMETER project known as the DEMETER Agricultural Information Model (AIM). Based on this model the semantic interoperability between different systems and data models can be carried out. AIM provides the common vocabulary that is used to model and exchange data between different components thereby providing an integrated view over different and heterogeneous data sources. Therefore we can say that it is a core element of the Data Preparation & Integration Enabler, that is used to transform data into AIM format or translate queries using AIM terms. Similarly, the data generated by the other enablers, (e.g., knowledge extracted, data quality information) can also be represented using the AIM model. The deliverable D2.2 includes an analysis of the state of the art. The analysis includes a description of different approaches, methods and techniques relevant to the implementation of the DK enablers and about the various tools, services and applications that may be re-used during the implementation process. The deliverable also describes the requirement of implementation of all types of enablers for different categories of actions such as:

- Data Integration including semantic Interoperability and integration requirements

- Data Management, including CRUD operation (Create, Read/Retrieve, Update, Delete/Destroy)
- Data storage
- Data synchronization and translation to/from various data access methods via query languages
- Data discovery and data aggregation
- Data Quality & Fusion
- Data Analytics and various Machine Learning techniques
- Data Security & Privacy

The deliverable includes, for each enabler, a description of how it is framed with respect to the DEMETER Reference Architecture (RA), the approach of the specific enabler regarding the action it addresses and a detailed presentation of its design including UML diagrams. Additionally, the deliverable includes a presentation of the initial release of the DK enablers and the work that needs to be followed up for the second and final release of these tools.

3 Description of the DEMETER components/enablers

3.1 Data Management Components

In the deliverable a detailed summary of the data management components or enablers including an introduction to the design/approach was presented, that covers all aspects of multi-tenancy, availability, scalability and QoS of the data management components. DEMETER's Data Management (DM) integrates on one hand the architectural practices and techniques with the tools necessary to access and deliver data to meet the data consumption requirement for all application processes. DEMETER enablers for data management relies on the data provided by the pilots relating to climate, weather, soil, water quality and crop status. These enablers for data management respond to Pilots who intend to store their resources in the DEMETER Enabler HUB. The DEMETER data management Enablers will:

- Allow storage, access, processing and delivery of data taking into account all the specifications of the DEMETER Pilot sites
- make use of a complete set of components consisting of different technologies allowing the heterogeneity and complexity of the project.
- Support the availability of data to various DEMETER front-end applications or enablers independently by including mechanisms to isolate the workload requirements and control various end user parameters.
- Support a varied range of input data formats by using common technologies capable of standardizing them according to standard protocols and common information model or AIM.

- ensure easy and efficient configurability of the components, guaranteeing common technologies that comply with the guidelines and design in activities such as provisioning and delivery of data management software.

The data management components are realized by three component blocks called Brokerage Service Environment (BSE) and DEMETER enabler Hub (DEH) and Resource Registry. These blocks or enablers are part of the data management process allowing the acquisition of data, transport and storage. Any of the DEMETER Enhanced Entity (DEE) (e.g any platform, thing, service or application) is made available through the DEMETER Enabler Hub for the purpose of its availability for other services. On the other hand, the instances of the DEEs annotated with their metadata (e.g. characteristics, ownership information, location-based restrictions etc.) are transported from the BSE to the DEH and eventually stored in the Resource Registry. In data management all these three components interact with each other while registering a resource and its metadata in the DEH. The BSE and DEH Enablers use a SaaS (Software as a Service) approach to expose interoperability APIs as RESTful Services in order to manage the data from DEE.

3.2 Data Preparation and Integration Components

Data preparation and integration components summarises a DEMETER specific approach for data integration using Linked Data as a federated layer. Additionally the description includes the design, present stage of development work regarding the implementation of data preparation & integration pipelines for transformation of heterogeneous data into Linked Data (includes description of the tools and programs used in the pipelines and on-going work in the implementation of an enabler API).

Firstly the data preparation stage involves the process of collecting, cleaning or validating and transforming raw data prior to processing and analysis and thereafter storing of the data. This step often involves reformatting data and standardizing data formats, making corrections to data, removing outliers and combining data sets to enrich data. Once the data has been prepared, data integration is the process used to combine data from various sources into meaningful and valuable information. It allows achieving a unified view of the data. The DEMETER approach for data preparation and integration is based on Linked Data which is a continued approach from various projects from the agrifood sector. In those projects extensive use of Linked Data as a federated layer to support large scale harmonization and integration of data from a large scale of heterogeneous sources. This action has been realized through the implementation of instantiations of a 'Generic Pipeline for the Publication and Integration of Linked Data". The main goal of these pipeline instances is to define and deploy (semi-) automatic processes to carry out the necessary steps to transform and publish different input datasets as Linked Data, i.e. automatic

processes for data preparation and integration using Linked Data involving many different data processing tools and programs. These tools facilitate the transformation of input data into RDF format (using mapping specifications to process the input datasets) or the translation of queries to/from SPARQL, plus their linking and integration with other known datasets out there. The main instantiations of the generic pipeline based on the source of the input data are:

- Pipeline for geospatial data (i.e., shapefiles) using some underlying ontology to prepare RML mapping specifications, and specific tools (e.g Geotriples) to transform into Linked Data.
- Pipeline for (semi-)structured input data (e.g. csv, json). The input data is prepared and processed and thereafter mapping specifications (in RML) are generated in a semi-automated process using domain specific modelling vocabularies.
- Pipeline for relational databases for translation of data from relational databases as Linked Data on the fly (i.e. data stays at the source and a virtual semantic layer is created on top).
- Pipeline for hybrid services translation involving data sources of hybrid origin. The structured metadata can be made available as Linked Data using SPARQL compliant endpoints to request the non-sparql backend on the fly.

As part of the DEMETER project , these pipeline instantiations are to be adapted/extended, new pipelines are created and a DEMETER API is in the process of implementation. This DEMETER-enabled API allows reuse and automatization of the functionalities in the pipelines. The DEMETER API is planned to give a common access point for all the tools/services with different interfaces and protocols (CLI, API) used in the pipeline instatiations to prepare and integrate data. The API will also provide access to data based on the AIM model, and will use SPARQL queries to showcase some basic use case scenarios.

3.3 Data quality components

The data quality is also a vital aspect in the process of data management and integration in DEMETER. The deliverable describes the data quality components including descriptions of the quality requirements, the design and on-going work regarding the implementation of a specific data quality enabler API. The enabler is planned to provide facilities like quality assessment of Linked Data, tabular data and the aspects of data provenance.

The starting point for a quality assessment is the quality requirement, which depends on the context of the use case and its requirements. Based on the requirement, the concept of data quality can be illustrated by a number of quality characteristics (dimensions) of varying importance. The ISO defines these characteristics as a

"category of data quality attributes that bears on data quality" (ISO - ISO/IEC 25012:2008 - Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model 2020). Having this concept in mind, relevant data quality characteristics were identified and next they were further elaborated and break-down into single data quality measurements which fit the use case context and needs. For example, the data quality assessment characteristics or completeness can be measured with a metric ratio of null values that calculates the ratio of the number of null values within a dataset. Furthermore, those data quality metrics can then be instantiated to perform the real measurement, e.g., by coding a python script for a specific data set.

3.4 Data analytics and fusion components

The data analytics and data fusion components are within the DEMETER Data Analytics and Knowledge Extraction Enabler. The difference between analytics and fusion tasks is that analytics tasks extract knowledge from data sources, while fusion tasks extract features from data sources. The activities regarding data fusion and data analytics involve some specific components including the design of the enabler API for data fusion and analytics. Within this enabler the analytics and fusion tasks are already targeted and use-case specific and thus every module addresses one or multiple requirements in the DEMETER pilots. Moreover there are general machine learning (ML) components that can be utilized by all analytics and fusion modules. The ML components aim at centralizing recurring tasks of analytics and fusion modules within the scope of the DEMETER Data Analytics Lifecycle that involves data acquisition and access, as well as Data Preparation and Integration tasks (supported by the DEMETER Data Management, and Data Preparation & Integration Enabler). The Data Analytics and Knowledge Extraction Enabler takes care of the remaining aspects of the lifecycle and aims at offering those capabilities to support upstream Decision Support and Benchmarking as well. The Data Analytics (and Fusion) API will be implemented as a REST-based Web Service using FastAPI, which is a high-performance web framework for building API for Python. The solutions will be deployed as a dockerized container. Few initial analytics modules are proposed as of now to illustrate the nature of DEMETER analytics tasks that aim at extracting new facts from existing data, by employing primarily machine learning and deep learning methods such as: using imagery data for the identification of visual elements through pattern extraction, pattern extraction for counting fruit fly using computer vision template, optimal fertilizer usage using pattern extraction techniques, data analysis for water salinity and plant toxicity in rice fields. In the later stages of the project, more analytics/knowledge extraction tasks are aimed to be implemented as modules, to make them more accessible for other users of the DEMETER Enabler Hub.

3.5 Data security components

Security being a fundamental aspect of the DEMETER project involves the security of the data and more specifically the access to the data, authentication, access to the resources, authorization, privacy or confidentiality. Data protection and regulations such as GDPR have been taken into account in the DEMETER security components. The data security components within the DEMETER security architecture are as follows:

- Authentication is the management of different entities which must be registered in a system in order to establish the permissions for accessing the system. There are different available solutions such as Keyrock (a Generic Enabler from FIWARE), Workspace ONE access (from VMWare) or Microsoft Identity Manager that could make use of different protocols for performing the authentication process. There are some authentication frameworks, which are widely adopted and used, that have become standards such as OAuth2 or OpenID. The Demeter Identity Manager (IdM) Enabler is based on the FIWARE Keyrock GE and will provide the Keyrock's API for authentication based on the OAuth 2.0 protocol.
- Authorization or the access control to the resources of a system where traditional approaches were not designed to deal with heterogeneous scenarios where interoperability is a must. For Demeter the proposal of using Distributed Capability-Based Access Control (DCapBAC) was conceived. DCapBAC has been implemented as a DEMETER Security component to manage resource access control.
- Traceability components are common for all the authentication and authorization components that also provide a logging activity so that traceability is implemented and can be easily followed. Any user/device wishing to access the DEMETER Dashboard or use the DEMETER services will have to login and, once logged and according to their granted permissions, will be able to recover a determined type or amount of information. These operations will be performed thanks to the Identity Management and Authorization components.
- Confidentiality component deals with data protection and is a relevant aspect for the DEMETER project. The prevention of readings by unauthorized users is a fundamental aspect for the robustness of the security system. The Transport Layer Security protocol is an IETF standard designed to increase security in communication between network entities providing both confidentiality and integrity of data.

demeter

CONTACT US

Kevin Doolin, Project Coordinator

Kevin.Doolin@WaltonInstitute.ie

or email **INFO@H2020-DEMETER.EU**

For more information visit **WWW.H2020-DEMETER.EU** and follow us on social media:

H2020DEMETER **H2020DEMETER** **h2020-demeter** **h2020-demeter**