Infrastructure data and service providers registration (M30)

November, 2017

Deliverable Code: D8.7

Version: V2.0 – Final

Dissemination level: Public

This deliverable presents the data and service providers of OpenMinTeD.

H2020-EINFRA-2014-2015 / H2020-EINFRA-2014-2
Topic: EINFRA-1-2014
Managing, preserving and computing with big research data
Research & Innovation action
Grant Agreement 654021

Document Description

D8.7 - Infrastructure data and service providers registration (M30)

WP8 – Operation and Maintenance	
WP participating organizations: ARC, OU, UNIMAN, UKP-TUDA, USFD, INRA, Frontiers, GESIS	
Contractual Delivery Date: 30/11/2017	**Actual Delivery Date:** 11/05/2018
Nature: Report	**Version:** 2.0
Public Deliverable	

Preparation slip

	Name	Organization	Date
From	Dimitrios Galanis Katerina Gkirtzou	ARC	1/12/2017
Edited by	Dimitrios Galanis	ARC	01/05/2017
Reviewed by	Richard Eckhart de Castilho Petr Knoth	UKP-TUDA OU	13/03/2018
Approved by	Androniki Pavlidou	ARC	11/05/2018
For delivery	Mike Hatzopoulos	ARC	15/05/2018

Document change record

Issue	Item	Reason for Change	Author	Organization
V0.1	Draft version	Initial version sent for comments	Dimitrios Galanis Katerina GkirtzouPenny Labropoulou	ARC

V1.0	First version	Incorporating reviewers' comments	Dimitrios Galanis	ARC
V2.0	Final version	Updating content and incorporating last comments	Dimitrios Galanis	ARC

Table of Contents

Disclaimer

This document contains description of the OpenMinTeD project findings, work and products. Certain parts of it might be under partner Intellectual Property Right (IPR) rules so, prior to using its content please contact the consortium head for approval.

In case you believe that this document harms in any way IPR held by you as a person or as a representative of an entity, please do notify us immediately.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of the OpenMinTeD consortium and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 28 Member States of the Union. It is based on the European Communities and the member states cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors. (http://europa.eu/)

OpenMinTeD is a project funded by the European Union (Grant Agreement No 654021).

Publishable Summary

This goal of this report is to present all content (data) and service providers of OpenMinTeD in project Month 30. We also briefly describe how they are registered.

1. *Data and service providers*

D8.7 – "Infrastructure data and service providers registration (M30)" report is available online (as it is specified in the Grant Agreement of OpenMinTeD project) in the following link.

https://docs.google.com/document/d/18xagn_fILqQD-bXKNlztlORUmVv7wCyakVJ1Gx8fr-s/edit?usp=sharing

The information of the Google document is the following due to M30 of the project.

1.1 Data (content) providers

Scholarly and scientific content, which is the main type of data targeted by OpenMinTeD, comes from a wide bulk of stakeholders, e.g. institutional and discipline repositories, academic journals, scientific publishers, etc.

Data providers who are interested in making their content available for TDM through the OpenMinTeD platform must follow the relevant OpenMinTeD guidelines1. There are two main requirements for this: making their metadata available in the corpus construction mechanism and providing a direct link to the data resources.

Providers use their own metadata schemas for the description of publications. In order to register their content in the OpenMinTeD platform, they must implement a connector interface, the definition of which is available at GitHub2. Each implemented connector can easily be integrated into the platform; when this happens, the registration of a provider is completed, and its content and metadata can be deployed. Each connector offers the following functionalities3:

- Performs mapping from the OpenMinTeD metadata schema (OMTD-SHARE metadata schema[4]) to the external provider's schema, allowing the connector to return metadata in a common form.
- Provides search functionality by using the proprietary search API of the data provider and returning the results in a common format.
- Provides access to the full text (e.g. JATS XML, PDF) of the publications, allowing the construction of new corpora following the criteria set by the user query.

[1] https://guidelines.openminted.eu/guidelines_for_providers_of_publications/

[2] https://github.com/openminted/content-connector-api

[3] See also Section 4.5 - "Content Connector" of D6.1 - "Platform Architectural Specification" for more information, http://openminted.eu/wp-content/uploads/2017/01/D6.1-Platform-Architectural-Specification.pdf

[4] The schema is available at https://github.com/openminted/omtd-share-schema and its full documentation at: https://openminted.github.io/releases/omtd-share/

In order to multiply its impact, OpenMinTeD relies on existing infrastructures. In the case of data providers, this principle has led to the collaboration with two main aggregators of scholarly content. Thus, in month 30 of the project (as in month 24), two content providers are registered in the platform:

- **OpenAIRE**[5]: OpenAIRE is an aggregator with outreach to many different Open Access repositories and journals. The respective implemented content connector for this provider is available at the respective GitHub repository[6]. It has fully been integrated to OpenMinTeD platform and thoroughly tested.
- **CORE**[7]: CORE is an aggregator of content stored in Open Access repositories and journals. The respective connector is implemented and available at GitHub[8]. The connector has fully been integrated to the platform.

1.2 Service providers

In OpenMinTeD, under the term "service providers" we include the organizations, researchers and software developers that make available text & data mining software, either fully packaged as an end-user application or as separate components that can be re-used in the creation of new applications.

The offered components must comply with the **interoperability specifications** of the platform; more information is available in the respective deliverables, i.e., D5.2 - "Interoperability Standards and Specifications Report"[9] and D5.5 - "Platform Interoperability Guidelines", as well as in their updated versions, i.e. D5.3, D5.4 and D5.6 respectively.

Software can be registered in OpenMinTeD following the procedure described in the respective guidelines[10]; the main requisites are:

1. All the required metadata according to OMTD-SHARE metadata schema are provided and stored in the Registry[11].
2. The actual software resource (e.g. Maven artifact or Docker image) can be downloaded from where it is provided (e.g. Docker Hub) and deployed at OpenMinTeD.

[5] https://www.openaire.eu/

[6] https://github.com/openminted/content-connector-openaire

[7] https://core.ac.uk/

[8] https://github.com/openminted/content-connector-core

[9] http://openminted.eu/wp-content/uploads/2016/12/D5.2-Interoperability-Standards-and-Specifications-Report-v1.2.pdf

[10] https://guidelines.openminted.eu/guidelines_for_providers_of_sw_resources/sharing-software-through-openminted.html

[11] see D6.1 - "Platform Architectural Specification" for more information on the Registry.

In month 30, four organizations have prepared or are in the process of preparing their components and applications for registration, e.g.,

- Making them compliant with the interoperability specifications (e.g adapting output format) if needed.
- Creating metadata descriptions from scratch or by converting and/or enriching existing metadata records and making these compatible with the OMTD-SHARE metadata schema.
- Creating docker images of their software (if needed).

The aforementioned four providers are partners in the OpenMinTeD project:

- **TECHNISCHE UNIVERSITÄT DARMSTADT - UKP Lab (UKP-TUDA)**[12]**:** It offers DKPro Core[13], a repository of UIMA components for NLP: Approximately ~ 200 components are currently available with the respective OMTD-SHARE metadata files.
- **The University of Sheffield (USFD)**[14]**:** It offers a repository of GATE[15] components for NLP. Approximately 50 components are currently available with the respective OMTD-SHARE metadata files.
- **UNIMAN - National Centre for Text Mining (NaCTeM)**[16]**:** It offers a repository of UIMA components for NLP: The components are under preparation
- **Institut National de la Recherche Agronomique (INRA):** It offers a repository of NLP components created with the AlvisNLP/ML framework[17]: The components are under preparation

The components that are ready (DKPro Core, USFD) have been registered to the OMTD workflow execution backend by automatically creating the respective Galaxy XML files. Many of them have been tested successfully as parts of Galaxy workflows with sample input data; an example of a Galaxy workflow is shown in Fig. 1 below.

[12] https://www.ukp.tu-darmstadt.de

[13] https://dkpro.github.io/dkpro-core

[14] http://www.sheffield.ac.uk/

[15] https://gate.ac.uk/

[16] http://www.nactem.ac.uk/

[17] https://github.com/Bibliome/alvisnlp

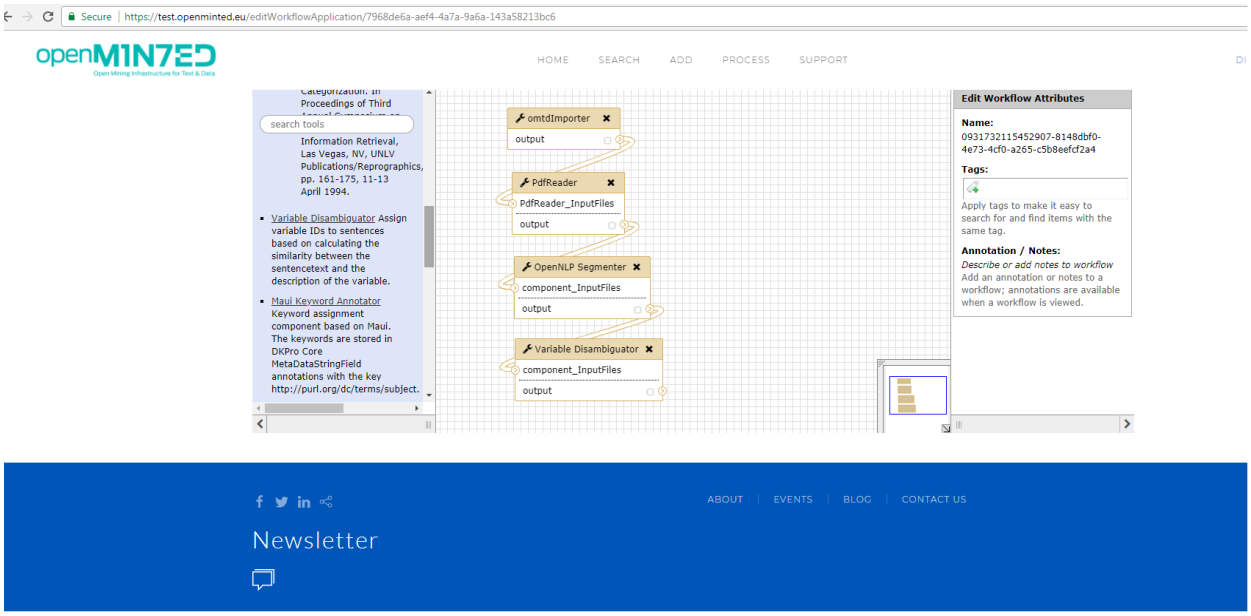

Figure 1: OpenMinTeD Galaxy workflow

Some of these workflows were also registered in the OMTD Registry and were tested with corpora automatically created with the OMTD corpus builder that uses as data sources the OpenAIRE and CORE aggregators.

In addition, in the context of WP9 a set of applications[18] are being prepared (by OMTD partners) to be registered in the platform via one of the following options:

- using the workflow editor and creating the respective workflows from the required components.
- by creating ready-to-use end-user applications.

For example

- Two text mining components developed by ARC are made compatible with the OMTD specs:
  - madIS funding mining extractor: This component mines publications' full texts and extracts links to acknowledged projects; various funders are supported including European Commission (FP7/H2020), NSF, FCT, Wellcome Trust etc.
  - DataCite mining component: This component mines publications' fulltexts, searches the references section and extracts links to DataCite[19] .

[18] For a detailed list of components and applications that are being prepared in the context of WP9 see D9.2- "Community Driven Applications Design Report"

[19] https://www.datacite.org/

- Three components have been prepared from UKP-TUDA and GESIS[20]
  - A Named Entity Recognition component for Social sciences papers by using an appropriate machine learning model trained on in-domain data.
  - To be integrated soon are two additional components for assigning keywords to documents and to detect mentions of survey variables in social sciences publications.
- UNIMAN has developed three web services for the life science use-cases
  - ChEBI web service can detect entities in six categories[21]: Metabolite, Chemical, Protein, Species, Biological Activity, and Spectral Data. The workflow underlying the web service consists of nine Argo components: Lingpipe Sentence Splitter, OSCAR Tokenisation, GENIA Tagger, and six NERSuite Taggers for six categories trained on the ChEBI corpus.
  - Neuroscience web service[22] can extract nine types of entities: Neurons, Brain Regions, Scientific Values, Scientific Units, Ionic Currents, Ionic Channels, Ionic Conductances, Synapse, and Model Organisms, and five types of modelling parameters: Neuron Density, Maximal Ionic Conductance, Resting Membrane Potential, Volume of Brain Region, and Ohmic Input Resistance. The workflow underlying the web service consists of three basic NLP Argo components: Lingpipe Sentence Splitter, OSCAR Tokenisation, and GENIA Tagger. For each category of entities, a NERSuite Tagger was included into the worfkflow. The model for each tagger was trained on an in-house corpus using Conditional Random Fields.
  - A web service to normalise four categories of neuroscience entities including Neuron, Brain Region, Ionic Current, and Model Organisms to the NIFSTD ontology.
- ArgoKnow has made the following extractors to be compatible with the OMTD platform[23]:
  - ArgoVoc Extractor: Search and Rank Agricultural Terms in PDF documents
  - Geopolitical Extractor: Search Geolocation terms in the FAO Geopolitical Ontology
  - Grapevine Extractor: Search and Rank Grape Varieties in PDF documents
  - Geonames Extractor: Search and Rank Geolocation Terms in PDF documents
- INRA has done three workflows for the agriculture science use case[24].
  - Habitat-Phenotype Relation Extractor for Microbes: a workflow that can recognize microorganism taxa, their habitats and their phenotypes. It categorizes them with ontologies (NCBI taxonomy and OntoBiotope ontology). It identifies lives-in

[20] https://github.com/openminted/uc-tdm-socialsciences

[21] https://github.com/openminted/uc-tdm-LS-A

[22] https://github.com/openminted/uc-tdm-LS-B

[23] https://github.com/openminted/uc-tdm-agriculture

[24] https://github.com/openminted/alvis-docker/tree/master/openminted-components

relationships between taxa and habitats and exhibits relationships between taxa and phenotypes.
    - Wheat Phenotypic Information Extractor: a workflow that can recognize phenotypes, genes, markers, and wheat-related taxa. It categorizes the phenotypes with the Wheat Trait Ontology.
    - Arabidopsis Gene Regulation Extractor: a workflow that can recognize Gene, Protein and RNA of Arabidopsis thaliana. It normalizes them with Gene Locus and identifies, interacts with relationships between Gene and Protein.
- Open University has built the following components:
    - CORE recommender[25] web service that offers content-based recommendation.
    - Structure extractor from pdf, component makes use of Grobid[26] machine learning for extracting, parsing and restructuring raw documents
    - Citation analytics web service that offers citation counts for given documents utilising the MAG[27] corpus
    - Structure extractor from pdf, component similar to Grobid extractor above, using ScienceParse[28] machine learning models instead
- Frontiers has built
    - a TDM application on articles related to Health State Modelling in Chronic Liver Diseases. **Health State Modelling** represents different health states to provide the right intervention for the right patient at the right time and dose. This application uses the DKPro OpenNLP Segmenter, a Liver Disease Tagger, and a Co-occurrences Graph Builder.
    - a **Rock Art Mining application** that is about assisting a researcher finding ancient artistic artefacts in the archaeological scientific literature. It is necessary to strictly mine scientific articles since the trust in the results constitutes a critical factor. Artistic artefacts can be rock paintings, petroglyphs, pictographs or engravings. The following properties must be extracted: Artefact type, Site name, Date, Dating method. The components that are used for this application are a PDF Tables Extractor, a Tables Features Extractor, and a Date Range Annotator.
    - **Text Mining application of Articles Related to Leica Microscopes**: The goal of this use case is to identify citations of Leica related products. Leica is manufacturer of optical microscope. Their products play a big role in the research community. Usually

[25] https://core.ac.uk/services#recommender

[26] http://grobid.readthedocs.io/en/latest/

[27] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/

[28] https://github.com/allenai/science-parse

citations analytics focus on citations of articles. Tools used as part of research are also a factor of their success. Identifying tools used in successful research could lead to better research. Such information is also very valuable to manufacturing companies not only to know what their most cited products are but as well when concurrent products are used instead. A better understanding of what is used could give a better insight of the demand and would therefore lead to better products. The tool used is a Leica Model Annotator.

All the aforementioned components/applications are planned to be available in the OMTD Registry and execution backend at the end of the project.