Infrastructure data and service providers registration (M24)

November, 2017

Deliverable Code: D8.6

Version: V1.0 – 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

D8.6 - Infrastructure data and service providers registration (M24)

WP8 – Operation and Maintenance	
WP participating organizations: GRNET, ARC	
Contractual Delivery Date: 31/05/2017	**Actual Delivery Date:** 19/12/2017
Nature: Report	**Version:** 1.0
Public Deliverable	

Preparation slip

	Name	Organization	Date
From	Dimitrios Galanis Katerina Gkirtzou	ARC	1/12/2017
Edited by	Dimitrios Galanis	ARC	5/12/2017
Reviewed by	Penny Labropoulou	ARC	18/12/2017
Approved by	Androniki Pavlidou	ARC	19/12/2017
For delivery	Mike Hatzopoulos	ARC	19/12/2017

Document change record

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

V1.0	First version	Incorporating reviewers' 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 24. We also briefly describe how they are registered.

1. Data and service providers

D8.6 – "Infrastructure data and service providers registration (M24) "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/1VNPU3Llu1Trds31mNIifrY3txNDqso6tjZINgpoov2o/edit#

The information of the Google document is the following due to M24 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 guidelines (https://guidelines.openminted.eu/guidelines_for_providers_of_publications/). 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 GitHub[1]. 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 functionalities[2]:

- Performs mapping both from the OpenMinTeD metadata schema (OMTD-SHARE metadata schema[3]) to the external provider's schema and the reverse, 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 of the publications, allowing the construction of new corpora following the criteria set by the user query.

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

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

[3] 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 24 of the project (as in month 18), two content providers are registered in the platform:

- **OpenAIRE**[4]: 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[5]. It has fully been integrated to OpenMinTeD platform and thoroughly tested.
- **CORE**[6]: CORE is an aggregator of content stored in Open Access repositories and journals. The respective connector is implemented and available at GitHub[7]. 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[8]" 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 at https://guidelines.openminted.eu/guidelines_for_providers_of_sw_resources/sharing-software-through-openminted.html; the main requisites are:

1. All the required metadata according to OMTD-SHARE metadata schema are provided and stored in the Registry[9].
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.

In month 24, four organizations are preparing their components and applications for registration, e.g.,

[4] https://www.openaire.eu/
[5] https://github.com/openminted/content-connector-openaire
[6] https://core.ac.uk/
[7] https://github.com/openminted/content-connector-core
[8] http://openminted.eu/wp-content/uploads/2016/12/D5.2-Interoperability-Standards-and-Specifications-Report-v1.2.pdf
[9] see D6.1 - "Platform Architectural Specification" for more information on the Registry.

- 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.
- Possibly preparing harvesters for extracting the metadata from the source code, software artifacts etc.
- Possibly creating docker images of their software (if needed).

The aforementioned four providers are partners in the OpenMinTeD project:

- **TECHNISCHE UNIVERSITAET DARMSTADT (UKP-TUDA)[10]:** It offers DKPro[11], a repository of UIMA components for NLP.
- **The University of Sheffield (USFD)[12]:** It offers a repository of GATE[13] components for NLP.
- **UNIMAN - National Centre for Text Mining (NaCTeM)[14]**: It offers a repository of UIMA components for NLP.
- **Institut National de la Recherche Agronomigue (INRA):** It offers a repository of NLP components created with the Alvis NLP framework[15].

In addition, in the context of WP9 a set of applications[16] are being prepared (from OMTD partners) to be registered in the platform by

- either using the workflow editor and creating the respective workflows
- or by creating end-user applications.

However (in month 24), the components and applications described above were not completely ready (at least all of them) for registration. So, in order to test and demonstrate the OMTD Registry and the OMTD workflow execution backend, we have prepared manually for a small set of components the respective OMTD-SHARE metadata files as well as the XML files for the Galaxy workflow engine. The set consists of components provided by ARC, UKP-TUDA and UNIMAN; all of them have been integrated in workflows (they are also mentioned in MS60) as follows:

- Two text mining components developed by ARC that have been integrated and deployed in the OpenMinTeD platform:

10 http://www.tu-darmstadt.de/
11 https://dkpro.github.io/
12 http://www.sheffield.ac.uk/
13 https://gate.ac.uk/
14 http://www.nactem.ac.uk/
15 http://en.www.quaero.org.systranlinks.net/module_technologique/alvis-nlp-alvis-natural-language-processing/
16 See D9.1- "Community Driven Applications Design Report"

- madIS funding mining extractor: This component mines publications' fulltexts 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 (https://www.datacite.org/).

- Both aforementioned components require as input files in XMI format, so a PDF-to-XMI component (that reads PDF files and extracts the text) has also been registered in the platform in order to be used in building the workflows.
- Additionally, 3 more components were registered and used in the respective Galaxy worfklows (that consist of only one processing component):
  - Named entity recognition for Social Sciences: It is based on the CoreNLP named entity recognizer.
  - Chemical named entity recognition: The component is deployed as a Web service at University of Manchester and extracts chemical entities.
  - Topic modelling: It uses the DKPro Core component MalletLdaTopicModelInferencer that infers the topic distribution over documents using MALLET library.