openMINTED
Open Mining Infrastructure for Text & Data

Infrastructure data and service providers registration (M36)

June 15, 2018

Deliverable Code: D8.8
Version: 2.0
Dissemination level: Public

This deliverable presents the data and service providers of OpenMinTeD at month 36.

European Commission

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

D8.8 – Infrastructure data and service providers registration (M36)

WP8 – Operation and Maintenance	
WP participating organisations: ARC, OU, UNIMAN, UKP-TUDA, USFD, INRA, Frontiers, GESIS	
Contractual Delivery Date: 05/2018	**Actual Delivery Date:** 15/06/2018
Nature: Report	**Version:** 2.0
Public	

Preparation slip

	Name	Organisation	Date
From	Dimitrios Galanis Katerina Gkirtzou Penny Labropoulou	ARC ARC ARC	12/06/2018
Edited by	Penny Labropoulou	ARC	14/06/2018
Reviewed by	Sophie Aubin Byron Georgantopoulos	INRA GRNET	14/06/2018 14/06/2018
Approved by	Androniki Pavlidou	ARC	15/06/2018
For delivery	Mike Hatzopoulos	ARC	18/06/2018

Document change record

Issue	Item	Reason for Change	Author	Organisation
V1.0	Draft version	Initial version sent for comments	Dimitrios Galanis Katerina Gkirtzou Penny Labropoulou	ARC
v1.9	Draft version	Incorporating comments by reviewers	Penny Labropoulou	ARC
v2.0	Final version	Finalising the report	Penny Labropoulou	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).

Acronyms

ARC	Athena Research Center, Greece
JATS	Journal Article Tag Suite
INRA	French National Institute for Agricultural Research, France
NLP	Natural Language Processing
OU	Open University, UK
PDF	Portable Document Format
TDM	Text and Data Mining
UNIMAN	University of Manchester, UK
UKP-TUDA	Ubiquitous Knowledge Processing (UKP) Lab, Technische Universität Darmstadt, Germany
USFD	University of Sheffield, UK
XML	eXtenisible Markup Language

Publishable Summary

The goal of this report is to present all content (data) and service providers of OpenMinTeD in project month 36.

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 Text and Data Mining (TDM) through the OpenMinTeD platform must follow the relevant OpenMinTeD guidelines[1]. 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[2]. Each implemented connector can easily be integrated into the platform; when this happens, the registration of a provider is completed, and the content and metadata offered by him/her can be deployed. Each connector offers the following functionalities[3]:

- 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 a search functionality by using the proprietary search API of the data provider and returning the results (metadata) in a common format.
- Provides access to the full text (e.g. JATS XML[5], PDF) of the publications, allowing their collection into a new corpus following the criteria set by the user query.

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 36 of the project (as in month 24 and 30), two content providers are registered in the platform:

- OpenAIRE: 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 the OpenMinTeD platform and thoroughly tested.

[1] https://guidelines.openminted.eu/for-providers-of-content.html

[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/

[5] https://jats.nlm.nih.gov/

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

- CORE: 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 been fully integrated to the platform and tested.

In addition to these two aggregators, four major actors in the field, selected from the "Open Call for content providers" which was announced in October 2017 and closed in November of the same year, have provided their content to the OpenMinTeD platform through different mechanisms. More specifically:

- SciELO, an Open Access bibliographic database and digital library has developed a new version of their OAI-PMH service in order to provide all the resources (25,000 records in Health Sciences licensed under CC-BY) which are currently available through the OAI-PMH servers of each SciELO Network node published under the URLs shown in the following table.

Argentina	https://oaipmh.scielo.org/ar/
Bolivia	https://oaipmh.scielo.org/bo/
Brazil	https://oaipmh.scielo.org/br/
Chile	https://oaipmh.scielo.org/ch/
Colombia	https://oaipmh.scielo.org/co/
Costa Rica	https://oaipmh.scielo.org/cr/
Cuba	https://oaipmh.scielo.org/cu/
Mexico	https://oaipmh.scielo.org/mx/
Paraguay	https://oaipmh.scielo.org/py/
Peru	https://oaipmh.scielo.org/pe/
Portugal	https://oaipmh.scielo.org/pt/
Spain	https://oaipmh.scielo.org/es/
Uruguay	https://oaipmh.scielo.org/uy/
Venezuela	https://oaipmh.scielo.org/ve/

- IBECS, a bibliographic database with 150.000 registries, and abstracts from journals edited in Spain since 2000 in the health domain, has implemented an application that enables the IBECS resource maintainers to transform their LILACS 3 bibliographic metadata and abstracts into the OMTD-SHARE format and feed their contents into the OpenMinTeD platform
- INIA, the National Institute for Agronomic Research, has aligned its Scientific Journals' metadata with OpenAIRE 3 Guidelines in order to make them available for further exploitation using text mining techniques and has included them in the OpenMinTeD platform through the OpenAIRE aggregator.
- ISTEX, which is part of the "Investments for the Future" program initiated by the French Ministry and hosts 21 million of digital documents, has implemented a connector between the ISTEX repository and the OpenMinTeD registry. The inclusion of the ISTEX content in OpenMinTeD will facilitate users belonging to the French research community to run TDM applications offered through the OpenMinTeD platform on ISTEX resources.

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

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/combined in the creation of new applications.

The offered components and applications must comply with the **interoperability specifications** of the platform; more information is available in the respective deliverables, i.e., D5.4 - “Interoperability Standards and Specifications Report”[8] and D5.6 - “Platform Interoperability Guidelines”[9], and the sections dedicated to software providers in the online OpenMinTeD Interoperability Guidelines[10].

Software can be registered in OpenMinTeD following the procedure described in the Guidelines[11]; the main requisites are:

1. All the required metadata according to OMTD-SHARE metadata schema are provided and stored in the Registry[12].
2. The actual software resource (e.g. Maven[13] artifact or Docker[14] image) can be downloaded from where it is provided (e.g. Maven Repository or Docker Hub) and deployed at OpenMinTeD.

In month 36, five organizations from the OpenMinTeD consortium have prepared their components and applications for registration, by 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 providers are:

- **TECHNISCHE UNIVERSITÄT DARMSTADT - UKP Lab (UKP-TUDA)[15]:** It offers DKPro Core**[16]**, a repository of UIMA components for NLP; the latest release DKPro Core 1.9.2 contains 128 components as well as reading and/or writing support for 64 data formats/data sources; for most of these components, OMTD-SHARE metadata files are available. For some components,

[8] http://openminted.eu/wp-content/uploads/2016/12/D5.4_Interoperability-Standards-and-Specifications-Report.pdf

[9] http://openminted.eu/wp-content/uploads/2017/11/OpenMinTeD_D5.6-PlatformGuidelines_Final.pdf

[10] https://guidelines.openminted.eu/guidelines_for_providers_of_sw_resources/

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

[12] See D6.3 - “Platform Architectural Specification” for more information on the Registry.

[13] https://mvnrepository.com/

[14] https://www.docker.com/, https://hub.docker.com/

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

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

the OMTD-SHARE metadata is not generated because the users would not be able to sensibly configure and run the components on the OpenMinTeD platform.

- **The University of Sheffield (USFD)[17]:** It offers a repository of GATE**[18]** components for NLP. Approximately 50 components are currently available with the respective OMTD-SHARE metadata files.
- **UNIMAN - National Centre for Text Mining (NaCTeM)[19]**: It offers a subset (7 components) of their UIMA components for biomedical text mining[20]; the components integrated in the platform run remotely as web services.
- **Institut National de la Recherche Agronomique (INRA):** It offers 8 components[21] for text-mining and information extraction created with the AlvisNLP framework**[22]**. The components are either generic or dedicated to the biodiversity domain.
- **Barcelona Supercomputing Center (BSC)** has provided the following component:
  - The *NLProt component*[23] is a command line tool that combines dictionary- and rule-based filtering with several support vector machines (SVMs) to tag protein names in PubMed abstracts and delivers its output in its own XML format.

Many of these components have been deployed in the creation of new applications using the Galaxy workflow editor[24] (Fig. 1) and the applications tested in OpenMinTeD with real data from the content offered in the platform.

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

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

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

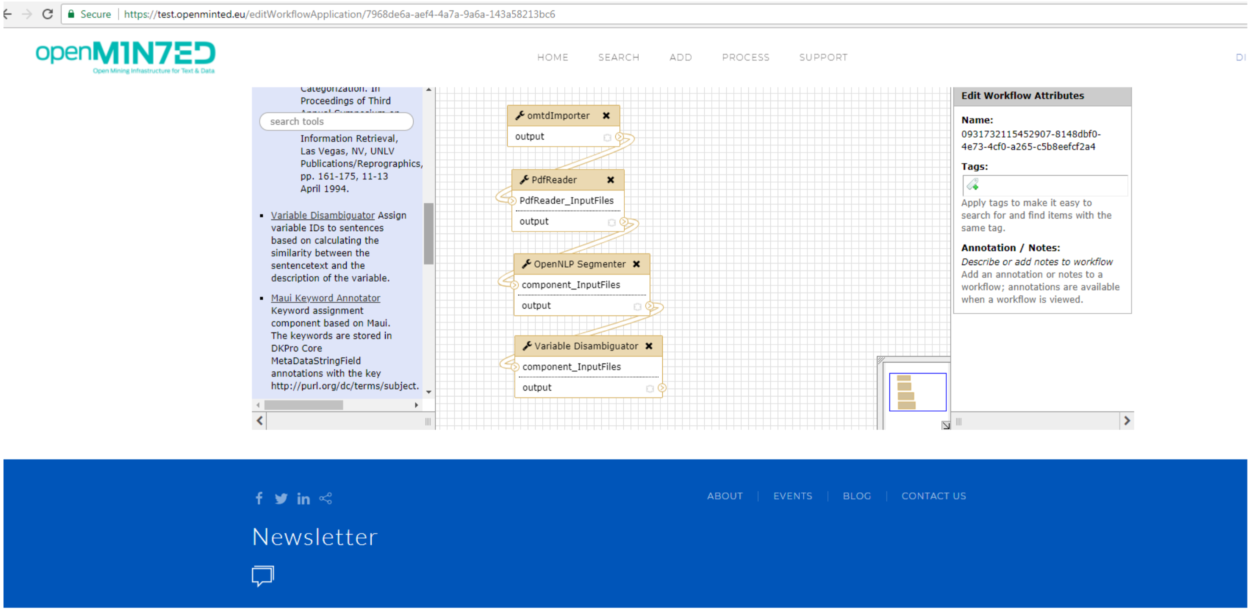
[20] https://openminted.github.io/releases/interop-spec/1.0.0/components/#__original_nactem_uima_18

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

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

[23] Under testing

[24] https://galaxyproject.org/, https://usegalaxy.org/



https://test.openminted.eu/editWorkflowApplication/7968de6a-aef4-4a7a-9a6a-143a58213bc6

Proceedings of Third Annual Symposium on Information Retrieval, Las Vegas, NV, UNLV Publications/Reprographics, pp. 161-175, 11-13 April 1994.

- Variable Disambiguator Assign variable IDs to sentences based on calculating the similarity between the sentencetext and the description of the variable.
- Maui Keyword Annotator Keyword assignment component based on Maui. The keywords are stored in DKPro Core MetaDataStringField annotations with the key http://purl.org/dc/terms/subject.

Edit Workflow Attributes

Name: 0931732115452907-8148dbf0-4e73-4cf0-a265-c5b8eefcf2a4

Tags:

Apply tags to make it easy to search for and find items with the same tag.

Annotation / Notes:

Describe or add notes to workflow

Add an annotation or notes to a workflow; annotations are available when a workflow is viewed.

Figure 1: A workflow example in Galaxy editor

In addition, in the context of WP9, a set of applications[25] have been prepared (by OpenMinTeD partners) and registered in the platform via one of the following options:

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

More specifically:

- Three text mining components developed by **ARC** have been adapted to the OpenMinTeD specifications and used in the respective applications (for **use case SC-A**):
  - the *MadIS funding mining extractor* mines publications' full texts and extracts links to acknowledged projects; various funders are supported including European Commission (FP7/H2020), NSF, FCT, Wellcome Trust etc.
  - the *DataCite mining component* mines publications' full texts, searches the references section and extracts links to DataCite[26] .
  - the *Classification component* mines publications' full texts and classifies them using given taxonomies.
- Three components have been prepared from **UKP-TUDA** and **GESIS**[27] (for **Use Case SS-A**):
  - a *PDF to text converter* for scientific publications based on CERMINE,

[25] For a detailed list of components and applications that are available in the context of WP9 see D9.2 - “Community Driven Applications Design Report”

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

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

    - a *Named Entity Recognition model* based on Social sciences papers by using an appropriate machine learning model trained on in-domain data for the DKPro Core Stanford Named Entity Recognizer ,
    - *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 - applications for the life science use-cases:
    - The *ChEBI web service*[28] (for **Use Case LS-A**) can detect entities in six categories: 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.
    - The *Neuroscience web service*[29] (for **Use Case LS-B**) 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 workflow. The model for each tagger was trained on an in-house corpus using Conditional Random Fields.
    - A web service (for **Use Case LS-B**) to normalise four categories of neuroscience entities including Neuron, Brain Region, Ionic Current, and Model Organisms to the NIFSTD ontology.
- **AgroKnow** has adapted the following extractors to the specifications, registered them in the OpenMinTeD platform[30] and created the respective applications:
    - The *ArgoVoc Extractor* (for **Use Case AS-A**) discovers AgroVoc terms within text segments, as defined by the AGROVOC Thesaurus
    - The *Grapevine Extractor* (for **Use Case AS-A**) discovers grape variety names within text segments, as defined by the OIV specification.
    - The *Geopolitical Extractor* (for **Use case AS-B**) discovers Geolocation terms of the FAO Geopolitical Ontology within text segments.
    - The *Geonames Extractor* (for **Use case AS-B**) discovers GeoName entities within text segments.
- **INRA** has created three applications for the agriculture science use case[31]:
    - The *Habitat-Phenotype Relation Extractor for Microbes* (**Use case AS-C**) is an application that can recognize microorganism taxa, their habitats and their phenotypes. It categorizes them with ontologies (NCBI taxonomy and OntoBiotope ontology). It

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

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

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

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

identifies lives-in relationships between taxa and habitats, and exhibits relationships between taxa and phenotypes.

  - The *Wheat Phenotypic Information Extractor* (**Use case AS-D**) is an application that can recognize phenotypes, genes, markers, and wheat-related taxa. It categorizes the phenotypes with the Wheat Trait Ontology.
  - The *Arabidopsis Gene Regulation Extractor* (**Use case AS-E**) is an application that can recognize Gene, Protein and RNA of Arabidopsis thaliana. It normalizes them with Gene Locus and identifies genic interactions.
- **OU** has built the following components[32]:
  - the *CORE recommender[33]* web service (**Use Case SC-B**) that offers content-based recommendation,
  - the *Structure extractor from pdf* component (**Use Case SC-B**) that makes use of Grobid[34] machine learning for extracting, parsing and restructuring raw documents,
  - the *Citation analytics web service* (**Use Case SC-C**) that offers citation counts for given documents utilising the MAG[35] corpus,
  - the *Structure extractor from pdf* component (**Use Case SC-C**), similar to the Grobid extractor above, using ScienceParse[36] machine learning models instead.
- **Frontiers** has built applications based on the following components:
  - The *Liver diseases and progression app* (**Use case LS-C**) is an application that works on PDF articles and annotates them with liver diseases (including synonyms) and progression keywords.
  - The *Date Range Annotation App* is an application that annotates a corpus with ancient kind of dates.
  - The *PDF Tables Extraction App* is an application that extracts tables form PDF as CSV files. The pages containing tables are saved as PNG.
  - The *Leica Model Annotation App* (**Use Case SC-D**) is an application that annotates a corpus with Leica Microsystems products.

As in the case of content, the successful applicants of the "Open Call for TDM software providers including TDM knowledge resources" (announced in December 2017), have registered several TDM software components and applications as well as knowledge resources at the OpenMinTeD platform, as described below:

- Agroportal: The OMTD-AgroPortal wrapper allows the OpenMinTeD platform to consume semantic knowledge resources from AgroPortal (a repository of knowledge resources for agriculture, food science, plant science and biodiversity) and by extension from the NCBO BioPortal or any other instances of the NCBO technology.

[32] Under testing
[33] https://core.ac.uk/services#recommender
[34] http://grobid.readthedocs.io/en/latest/
[35] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/
[36] https://github.com/allenai/science-parse

- BabelNetExtractor: A UIMA component to detect BabelNet terms in text packaged with BabelNet V3.7 and the corresponding libraries.
- BTH-OGER: Integration of resources developed by the OntoGene/BioMeXT group in the OpenMinTeD infrastructure, in particular: the OGER (OntoGene's Biomedical Entity Recognizer), an accurate, fast, efficient and robust Named Entity Recognition solution for biomedical entities which is built on top of the Bio Term Hub, an aggregator of terminologies from reference databases, and thus deploys up-to-date terminologies from it.
- FIMDA: The "SNP Extraction Tool for Human Variations" (SETH) is a named entity recognition component identifying a large set of different mutation types, ranging from simple substitutions/deletions/insertions to more complex mutations, such as translocations or inversions.
- FreeLing: A component including several tools (one for each language provided by FreeLing: Asturian, Catalan, English, French, German, Galician, Italian, Norwegian, Portuguese, Spanish, Russian, Slovene and Welsh) and a general tool that takes the language from the UIMA document metadata.
- IXA Pipes: IXA pipes, a set of ready to use NLP tools for Basque, Catalan, Galician, Spanish, Dutch, English, French, German and Italian: a tokenizer and sentence segmenter, a statistical lemmatizer and PartOfSpeech tagger and a state of the art statistical NER tagger, and it includes the following tools with a variety of language support: a chunker (Basque and English), Aspect Based Sentiment Analysis based on ABSA datasets (English, Spanish, French and Dutch) and Statistical Document Classifier.
- JONES: Two components running in the PubRunner framework that extract (i) abbreviations and (ii) subj-predicate-obj triples.
- Journalism: A series of accessible text mining tools to quickly give journalists answers to the five W's by analysing academic literature. These form the core of a suite of NLP tools for journalists, aimed at providing relevant information on which to build their reporting process.
- Relation Classifier: A classifier that extracts relations between entities using contextual information and biomedical ontologies. It combines deep learning with biomedical ontologies to improve relation classification solutions, more specifically it uses word embeddings and common ancestors of the two entities (as features) to train a neural network model.
- Summarizer: Text summarization services for automatically identifying the most important information of a research article. The project focuses on the implementation of a text analysis system for scientific papers and the implementation of a service for the computation of sentence relevance values.
- TermSuite: TermSuite is a toolbox for terminology extraction and multilingual term alignment that deals with multi-word and compound term detection, morpho-syntactic analysis, term variant detection, term specificity computation and many other features. It extracts monolingual terminologies and generates bilingual dictionaries from these terminologies by the means of distributional and compositional methods. The languages covered are English, French, German, Spanish and Russian.

- UPFMT: MLPLA text processing platform, a freely available NLP framework that supports more than 50 languages. MLPLA performs the following tasks either individually or as a pipeline: tokenization, sentence splitting, compound word expansion, lemmatization, PoS tagging and dependency parsing. The models are trained using the Universal Dependencies corpus.
- VineSum: An open source executable software component that, given a collection of documents: (i) performs NER extraction, identifying four entity types: vine varieties, persons, locations and dates and (ii) clusters the documents by taking into account the extracted entities and/or other extracted keywords.