First version of the interoperability standards and specification report that guides interoperability considerations within and beyond the OpenMinTeD project.

Topic: EINFRA-1-2014
Managing, preserving and computing with big research data
Research & Innovation action
Grant Agreement 654021
D5.4 – Interoperability Standards and Specifications Report

WP5 – Interoperability Framework

WP participating organizations: UKP-TUDA, ARC, UNIMAN, INRA, OU, USFD, UvA, UoS, GESIS

Contractual Delivery Date: 07/2017  Actual Delivery Date: 10/2017

Nature: Report  Version: 1.0

Public Deliverable

Preparation slip

<table>
<thead>
<tr>
<th>Name</th>
<th>Organization</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>From</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Richard Eckart de Castilho</td>
<td>UKP-TUDA</td>
<td>27/09/2017</td>
</tr>
<tr>
<td>Mouhamadou Ba</td>
<td>INRA</td>
<td></td>
</tr>
<tr>
<td>Penny Labropoulou</td>
<td>ARC</td>
<td></td>
</tr>
<tr>
<td>Dimitris Galanis</td>
<td>ARC</td>
<td></td>
</tr>
<tr>
<td>Thomas Margoni</td>
<td>UoG</td>
<td></td>
</tr>
<tr>
<td>Giulia Dore</td>
<td>UoG</td>
<td></td>
</tr>
<tr>
<td>Mark Greenwood</td>
<td>USFD</td>
<td></td>
</tr>
<tr>
<td>Angus Roberts</td>
<td>USFD</td>
<td></td>
</tr>
<tr>
<td>Matthew Shardlow</td>
<td>UNIMAN</td>
<td></td>
</tr>
<tr>
<td>Jacob Carter</td>
<td>UNIMAN</td>
<td></td>
</tr>
<tr>
<td>Edited by</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Richard Eckart de Castilho</td>
<td>UKP-TUDA</td>
<td>27/09/2017</td>
</tr>
<tr>
<td>Reviewed by</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Panagiotis Zervas</td>
<td>AK</td>
<td>16/10/2017</td>
</tr>
<tr>
<td>Mappet Walker</td>
<td>FRONTIERS</td>
<td>16/10/2017</td>
</tr>
<tr>
<td>Approved by</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Androniki Pavlidou</td>
<td>ARC</td>
<td>9/11/2017</td>
</tr>
<tr>
<td>For delivery</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mike Hatzopoulos</td>
<td>ARC</td>
<td>10/11/2017</td>
</tr>
</tbody>
</table>
### Document change record

<table>
<thead>
<tr>
<th>Issue</th>
<th>Item</th>
<th>Reason for Change</th>
<th>Author</th>
<th>Organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>V0.1</td>
<td>Draft version</td>
<td>First draft</td>
<td>Richard Eckart de Castilho</td>
<td>UKP-TUDA</td>
</tr>
<tr>
<td>V0.2</td>
<td>Draft version</td>
<td>Version sent to comments to reviewers</td>
<td>Richard Eckart de Castilho</td>
<td>UKP-TUDA</td>
</tr>
<tr>
<td>V0.3</td>
<td>Draft version</td>
<td>Incorporating reviewer’s comments</td>
<td>Richard Eckart de Castilho</td>
<td>UKP-TUDA</td>
</tr>
<tr>
<td>V1.0</td>
<td>First version</td>
<td>Finalization</td>
<td>Richard Eckart de Castilho</td>
<td>UKP-TUDA</td>
</tr>
</tbody>
</table>
Table of Contents

TABLE OF CONTENTS ..................................................................................................................4

INDEX OF TABLES ......................................................................................................................6

1. INTRODUCTION .....................................................................................................................11
   1.1 WORKING GROUPS .............................................................................................................11

2. SUMMARY REPORTS .............................................................................................................12
   2.1 WG1 ......................................................................................................................................12
   2.1.1 MISSION STATEMENT UPDATE ....................................................................................12
   2.1.2 MODE OF OPERATION UPDATE ....................................................................................12
   2.1.3 PROGRESS IN THE CURRENT PERIOD ........................................................................12
   2.1.4 FUTURE WORK ...............................................................................................................15
   2.1.5 STATE OF OPERATIONS ..............................................................................................15
   2.2 WG2 ......................................................................................................................................15
   2.2.1 MISSION STATEMENT UPDATE ....................................................................................15
   2.2.2 MODE OF OPERATION UPDATE ....................................................................................15
   2.2.3 PROGRESS IN THE CURRENT PERIOD ........................................................................16
   2.2.4 FUTURE WORK ...............................................................................................................17
   2.2.5 STATE OF OPERATIONS ..............................................................................................17
   2.3 WG3 ......................................................................................................................................17
   2.3.1 MISSION STATEMENT .................................................................................................17
   2.3.2 MODE OF OPERATION .................................................................................................17
   2.3.3 PROGRESS IN THE CURRENT PERIOD ........................................................................18
   2.3.4 FUTURE WORK ...............................................................................................................19
   2.3.5 STATE OF OPERATIONS ..............................................................................................19
   2.4 WG4 ......................................................................................................................................19
   2.4.1 MISSION STATEMENT .................................................................................................20
   2.4.2 MODE OF OPERATION .................................................................................................20
   2.4.3 PROGRESS IN CURRENT PERIOD .................................................................................20
   2.4.4 FUTURE WORK ...............................................................................................................22
   2.4.5 STATE OF OPERATIONS ..............................................................................................22

3. REQUIREMENTS .......................................................................................................................23

4. COMPLIANCE ............................................................................................................................24

5. LIST OF ATTACHMENTS ..........................................................................................................25
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.</td>
<td>APPENDIX</td>
</tr>
<tr>
<td>6.1</td>
<td>CONTENT-CONNECTOR-OPENAIRE</td>
</tr>
<tr>
<td>6.1.1</td>
<td>SEARCH</td>
</tr>
<tr>
<td>6.1.2</td>
<td>FETCHMETADATA</td>
</tr>
<tr>
<td>6.1.3</td>
<td>DOWNLOADFULLTEXT</td>
</tr>
<tr>
<td>6.2</td>
<td>CONTENT-CONNECTOR-CORE</td>
</tr>
<tr>
<td>6.2.1</td>
<td>SEARCH</td>
</tr>
<tr>
<td>6.2.2</td>
<td>FETCHMETADATA</td>
</tr>
<tr>
<td>6.2.3</td>
<td>DOWNLOADFULLTEXT</td>
</tr>
</tbody>
</table>
Index of Tables

Table 1 – Knowledge resources by domain and use—case application

16
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; see ILSP
CAS       (UIMA) Common Analysis System
          (https://uima.apache.org/d/uimaj-2.9.0/references.html#ugr.ref.cas)
CC        Creative Commons
          (https://creativecommons.org)
CCR       CLARIN Concept Registry
          (https://www.clarin.eu/ccr)
CLARIN    Common Language Resources and Technology Infrastructure
          (https://www.clarin.eu)
CM        Compatibility Matrix
D(number) (Project) deliverable
ELRA      European Language Resources Association
          (http://www.elra.info)
FOSS      Free and open-source software (https://en.wikipedia.org/wiki/Free_and_open-
          source_software)
INRA      French National Institute for Agricultural Research
ILSP      Institute for Language and Speech Processing (ILSP/"Athena" R.C.) aka ARC,
          Greece
JATS      Journal Article Tag Suite
KR        Knowledge resource
NACTEM    National Centre for Text Mining, University of Manchester, UK
NIST      National Institute of Standards and Technology, USA
          (https://www.nist.gov)
NLP       Natural Language Processing
M(number) Month counting from project start
MS(number) (Project) milestone
ODRL      Open Digital Rights Language
          (https://www.w3.org/community/odrl/)
OLIA      Ontologies of Linguistic Annotation
          (http://www.acoli.informatik.uni-frankfurt.de/resources/olia/)
OWL       Web Ontology Language
LAPPS Grid Language Application Grid
          (http://www.lappsgrid.org)
LDC       Linguistic Data Consortium
          (https://www.ldc.upenn.edu)
LR        Language Resource
LT        Language Technology
RDF       Resource Description Framework
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDM</td>
<td>Text and Data Mining</td>
</tr>
<tr>
<td>TheSOZ</td>
<td>Thesaurus for the Social Sciences (<a href="http://lod.gesis.org/thesoz/de.html">http://lod.gesis.org/thesoz/de.html</a>)</td>
</tr>
<tr>
<td>ToS</td>
<td>Terms of service</td>
</tr>
<tr>
<td>ToU</td>
<td>Terms of use</td>
</tr>
<tr>
<td>UIMA</td>
<td>Unstructured Information Management Architecture; usually referring to the reference implementation Apache UIMA (<a href="https://uima.apache.org">https://uima.apache.org</a>)</td>
</tr>
<tr>
<td>UNIMAN</td>
<td>University of Manchester, UK</td>
</tr>
<tr>
<td>UKP-TUDA</td>
<td>Ubiquitous Knowledge Processing (UKP) Lab, Technische Universität Darmstadt, Germany</td>
</tr>
<tr>
<td>UoG</td>
<td>University of Glasgow, UK</td>
</tr>
<tr>
<td>USFD</td>
<td>University of Sheffield, UK</td>
</tr>
<tr>
<td>WG</td>
<td>Working group</td>
</tr>
<tr>
<td>WP</td>
<td>Work package</td>
</tr>
</tbody>
</table>
Publishable Summary

The goal of the Interoperability Standards and Specifications report is to document measures taken to improve interoperability between relevant products from the TDM and NLP domains, in particular those involved and associated with the OpenMinTeD project. The process underlying the document is designed to closely involve internal and external stakeholders in the definition of requirements necessary to achieve better interoperability, with the aim also of committing these stakeholders to actually perform the necessary adjustments to their respective systems. This document is the third and last in a series of three. The first of the three, an earlier version of this deliverable, was delivered as D5.2\(^1\). It was updated again in D5.3\(^2\). The present deliverable focusses on presenting a high-level overview of the progress achieved within the reporting period (since D5.3) and on future work to be carried out in the context of the platform implementation work packages. The actual work documents that constitute the bulk of the deliverable package are provided as attachments to the present document. The sections regarding the compliance of relevant products with our interoperability requirements and specifications are no longer part of this deliverable as they are covered by the milestone reports MS34 and MS35 (Progress report of alignment of service and content provider systems). While these milestone reports are internal, they are still based on the current versions of the same publicly available compliance assessments records on which the relevant sections in D5.2 and D5.3 had been based.


1. **Introduction**

The main output of T5.2 “Infrastructure interoperability specifications” are interoperability standards and specifications. According to the methodology defined in MS5 “Working groups external experts list and work methodology” and elaborated in the first two version of the “Interoperability Standards and Specifications Report” (D5.2³, D5.3⁴), these are rendered primarily as “interoperability requirements”.

Through an assessment of compliance with these requirements, it is possible to assess the level of interoperability of a content provider’s systems, an analytics provider’s systems, and other types of resources with OpenMinTeD. As part of the first version of the Interoperability Standards and Specifications Report (D5.2), 72 of these requirements had been identified and described. A set of relevant products and services have had their compliance with these requirements assessed. However, the majority (69) of the requirements reported in D5.2 were so-called “abstract” requirements; i.e. they were described without fixing specific standards, technologies, or implementations. For the preparation of the second version (D5.3), these have been augmented by so-called “concrete” requirements which describe how in particular the abstract requirements can be fulfilled in a way supported by OpenMinTeD. For example, the abstract requirement REQ-51⁵ asserts that a license should be attached to a resource. The related concrete requirement REQ-92⁶ specifies the exact location and format of a license file for resources provided as ZIP or JAR files. During the preparation of the present deliverable (D5.4), work has focussed on creating new specifications for those areas of interoperability that been found not to be sufficiently covered by existing standards and specifications. The set of requirements has remained largely stable in comparison to D5.3.

Section 2 provides a short summary of the work that has been taking place in the interoperability working groups since the preparation of D5.3. Section 3 provides an overview of the updates to the interoperability requirements specification.

1.1 **Working groups**

Four working groups (WG) consisting of project members and external experts are contributing to the OpenMinTeD Interoperability Standards and Specification series of deliverables. These WGs are:

- WG1 – Resource metadata
- WG2 – Knowledge resources
- WG3 – IPR and licensing
- WG4 – Annotation workflows

---


⁵ https://openminted.github.io/releases/interop-spec/1.1.0/openminted-interoperability-spec/#REQ-51

⁶ https://openminted.github.io/releases/interop-spec/1.1.0/openminted-interoperability-spec/#REQ-92
2. Summary Reports

In this section, we provide short summary reports for each of the interoperability working groups covering the following aspects:

- **Changes in mission statement** – if applicable, includes a short, updated summary of the working group’s mission statement
- **Changes in mode of operation** – each of the working groups opted for slightly different modes of operations due to the heterogeneous scientific backgrounds and working habits
- **Progress in current period** – short summary of the progress and achievements from the current report period
- **Future work** – summary roadmap of the tasks planned for the remainder of the project
- **State of operations** – short self-assessment of the current state of operations

2.1 WG1

2.1.1 Mission statement update
The mission statement remains unchanged. For details, please refer to D5.2^7.

2.1.2 Mode of operation update
The mode of operation remains the same. For details, please refer to D5.2.

2.1.3 Progress in the current period
In the current period, WG1 has started consolidating and evaluating the work done over the previous periods. The main activities revolved around the implementation of the interoperability specifications regarding metadata through the OMTD-SHARE creation of descriptors for most of the resource types targeted by the project and initiatives aiming to improve the schema functionalities. The main outcomes of this period are the following:

- **OMTD-SHARE schema**: a new version (v2.0.1) of the metadata schema has been released; the XSD's, documentation and examples per resource type are available from our GitHub site^8. The new version incorporates comments from metadata editors and the updated specifications of WG4 addressing the operational needs of dockerized components and creation of workflows.
- **Glossary**: The OpenMinTeD Glossary, a joint work with WG3, has been published as a SKOS vocabulary^9. The glossary aims to establish a common working vocabulary among the various communities engaged in OpenMinTeD (researchers, developers of TDM applications, content providers, etc.) and clarify the concepts used in the OpenMinTeD infrastructure for the end users. Mappings to concepts of similar vocabularies and ontologies are also provided.

---


^8https://openminted.github.io/releases/omtd-share/2.0.1/

• **Ontology:** Work on the OpenMinTeD ontology has started. Using WebProtege, an online tool for collaborative editing of ontologies, a subgroup of WG1 with the FOSTER team has focused on concepts related to:
  - component types (RDF-izing the component types already enumerated in the OMTD-SHARE schema)
  - TDM operations
  - annotation types
  - data formats
  - TDM methods
  - domains

The aim of this effort is to increase the flexibility of the schema by introducing these concepts as recommended enumerations of the relevant metadata elements and allowing for input of new concepts by users. Once finalized, the ontology will be made publicly available, while its integration in the schema and the Registry will be coordinated with the Registry developers.

• **Creation of OMTD-SHARE metadata records (aka. descriptors)**:
  - **Content to be mined (scholarly content):** OpenAIRE & CORE have implemented a connector that allows harvesting from the respective aggregators metadata and data (full texts) of scholarly publications. The connector API serves as the interoperability layer between them and is designed in such a way that each side keeps their own functionalities, mechanisms and schemas, without interferences. The metadata from both sources (OpenAIRE & CORE) are combined in a joint catalogue in the OpenMinTeD platform and are used to build "user corpora", i.e. filtered subsets based on user queries; these corpora can then be submitted for processing by the TDM applications registered in OpenMinTeD. More information on the technical implementation of the connectors for OpenAIRE and CORE in Appendices I & II.
  - **Knowledge resources:** A set of terminologies, ontologies and lexical resources have been selected by WG2 to suit the OpenMinTeD use cases, and are being described using the OpenMinTeD minimal schema. This is being carried out as a manual process by the domain experts who are most familiar with the resources. The work is carried out in a spreadsheet and a script will be used to automatically generate the OMTD-SHARE metadata records. This work is still in progress.
  - **Components:** OMTD-SHARE descriptors are in the process of being generated by automatic processes by the responsible consortium partners. The following paragraphs give a short description on the operations adopted by each of them and the background

---

10 [https://webprotege.stanford.edu](https://webprotege.stanford.edu)

11 The ontology (working version) is available for viewing at: [https://webprotege.stanford.edu/#projects/a61c05a8-33ed-4347-a44f-4e34b4c394a0/edit/Classes](https://webprotege.stanford.edu/#projects/a61c05a8-33ed-4347-a44f-4e34b4c394a0/edit/Classes)

12 Sample metadata records from the consortium partners are uploaded at [https://openminted.github.io/releases/omtd-share/2.0.0/](https://openminted.github.io/releases/omtd-share/2.0.0/).

13 For a list of the resources, see section 2.2 on WG2.
Interoperability Standards and Specification

- **DKPro Core**: DKPro Core includes rich component metadata that are natively maintained in various forms, e.g. in the form of JavaDoc-based documentation or uimaFIT-based Java annotations allowing to mark parameters, supported languages, and other component-related metadata directly inside the component’s source code. Using the uimaFIT Maven Plugin, the metadata elements are aggregated into UIMA XML descriptors. Additional information, e.g. related to license, communication channels, and other organizational aspects, is maintained inside Maven project descriptors. In a second step, the OpenMinTeD Maven Plugin is used to merge the information from the UIMA XML descriptors with the metadata contained in the Maven project descriptors into OMTD-SHARE descriptors.

- **ALVIS**: The ALVIS components are described via internal XML wrappers. There is an XML wrapper that provides some human readable information for each component (e.g., description, input and output parameters). Compared to the OMTD-SHARE Schema, the schema of the wrappers lacks information such as component type, licensing, or validation information. A program allows currently to convert the existing wrappers to OMTD-SHARE descriptors limited to the main mandatory elements pending the production of additional information. It is planned to use the OMTD-SHARE Annotations (OMTD Maven plugin) after an adaptation of the annotation system of Alvis.

- **ARGO**: Argo Components support the UIMA Framework and are written in Java, using Maven to build these. Recently, new maven archetypes have been deployed\(^\text{14,15}\) allowing Argo components to be easily built. Argo components support uimaFIT Java annotations to provide metadata and so the same approach to generate OMTD-SHARE descriptors can be taken for Argo components as outlined for DKPro Core components above, i.e., using the uimaFIT Maven Plugin to generate UIMA descriptors and then using the OpenMinTeD Maven plugin to generate an OMTD-SHARE descriptor. We have recently deployed a patch for the uimaFIT Maven plugin which allows us to automatically include type systems when generating descriptors. The generation of OMTD-SHARE descriptors for Argo Components is ongoing, carried out as part of the application development in WP9.

- **GATE**: GATE components are natively described via existing Java annotations that allow us to embed rich metadata within the source code to document parameters, parameter types, documentation etc. When building a GATE plugin this metadata is automatically extracted to create a creole.xml file which can be interrogated for information on the components and resources within the plugin.

\(^{14}\) [http://search.maven.org/#search%7Cga%7C1%7Ca%3A%22argo-analysis-engine-archetype%22](http://search.maven.org/#search%7Cga%7C1%7Ca%3A%22argo-analysis-engine-archetype%22)

\(^{15}\) [http://search.maven.org/#search%7Cga%7C1%7Ca%3A%22argo-reader-archetype%22](http://search.maven.org/#search%7Cga%7C1%7Ca%3A%22argo-reader-archetype%22)
without recourse to the GATE API. Other useful metadata is also contained within the Maven pom.xml file (information such as version numbers, licenses, source code locations etc.). These two sources of metadata are then combined using the OpenMinTeD Maven plugin to produce one OMTD-SHARE descriptor file per component which are packaged inside the JAR file for easy access by other OpenMinTeD components, or possibly other text mining frameworks.

- **Dissemination and community engagement:** The OMTD-SHARE schema with an emphasis on the interoperability between data and software has been presented at the 2nd Datathon on Linguistic Linked Open Data. The guidelines have also been announced over the social media asking for feedback and will also be sent out to focused audiences.

### 2.1.4 Future work

- **Documentation and dissemination of OMTD-SHARE:** With each new release of the OMTD-SHARE metadata schema, documentation at the site and the Interoperability Guidelines (D5.5 and D5.6) is updated. Example XML files have also been made available and will be further enriched. The guidelines and presentation of the schema will be deployed for disseminating the schema to user communities and getting feedback from them.

- **Creation of OMTD-SHARE metadata records:** The focus up to now has been on the generation of metadata records for specific resource types. The consortium still needs to produce metadata for models and other resources used by these components; discussions on how to best describe the links between these and components are ongoing mainly with the DKPro Core team. All component developers that use models and other ancillary resources will also be required to produce metadata records for them.

- **Registration of resources in the OpenMinTeD platform:** Although not an interoperability task per se, the registration of all resources in the platform and the testing of their operation is influencing work in WG1 as it creates new or modified requirements for metadata.

### 2.1.5 State of operations

WG1 work has progressed significantly during this period, mainly through the implementation of the OMTD-SHARE descriptors and the discussions raised with WG4 for the deployment of metadata elements in the workflow descriptors.

### 2.2 WG2

#### 2.2.1 Mission statement update

The mission statement remains unchanged. For details, please refer to D5.2.

#### 2.2.2 Mode of operation update

The mode of operation remains the same. For details, please refer to D5.2.

---


2.2.3 Progress in the current period

The current period has been dedicated the following actions:

- **Strategy**: In the previous period, WG2 had shifted focus from the creation of mappings between knowledge resources and type systems towards examining suitable representations for knowledge level interoperability, in particular Web Annotation\(^\text{18}\) and UIMA XMI\(^\text{19,20}\) (with an appropriate type system). Internal project discussions decided these representations would be best examined within WG4, and so WG2 has turned in this period to the examination of specific knowledge resources and their match to the OpenMinTeD interoperability specifications.

- **WG2 teleconferences** have been convened every 2 to 4 weeks with members of the OpenMinTeD to consider the knowledge resource requirements of the OpenMinTeD use cases. The amended list of knowledge resources drawn up is given in the below table:

<table>
<thead>
<tr>
<th>Resource</th>
<th>Domain</th>
<th>Use case</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGROVOC</td>
<td>Agriculture</td>
<td>AS-A</td>
</tr>
<tr>
<td>AK-MGV GrapeVarieties</td>
<td>Agriculture</td>
<td>AS-A</td>
</tr>
<tr>
<td>OntoBiotope</td>
<td>Agriculture</td>
<td>AS-C</td>
</tr>
<tr>
<td>WheatPhenotype</td>
<td>Agriculture</td>
<td>AS-D</td>
</tr>
<tr>
<td>Plant ontology</td>
<td>Agriculture</td>
<td>none</td>
</tr>
<tr>
<td>Schema.org</td>
<td>General</td>
<td>none</td>
</tr>
<tr>
<td>SKOS</td>
<td>General</td>
<td>none</td>
</tr>
<tr>
<td>TERMCAT</td>
<td>General</td>
<td>none</td>
</tr>
<tr>
<td>GeoNames</td>
<td>Geographical</td>
<td>AS-B</td>
</tr>
<tr>
<td>FAO Geopolitical ontology</td>
<td>Geographical, Political</td>
<td>AS-B</td>
</tr>
<tr>
<td>MeSH</td>
<td>Life sciences</td>
<td>LS-A</td>
</tr>
<tr>
<td>NCBI taxonomy</td>
<td>Life sciences</td>
<td>LS-A</td>
</tr>
<tr>
<td>PubChem</td>
<td>Life sciences</td>
<td>LS-A</td>
</tr>
<tr>
<td>PubMed / Europe PMC</td>
<td>Life sciences</td>
<td>LS-A</td>
</tr>
<tr>
<td>BioLexicon</td>
<td>Life sciences</td>
<td>LS-B</td>
</tr>
<tr>
<td>NeuroLex</td>
<td>Life sciences</td>
<td>LS-B</td>
</tr>
<tr>
<td>OLIA</td>
<td>Linguistics</td>
<td>none</td>
</tr>
<tr>
<td>TheSoz</td>
<td>Social sciences</td>
<td>SS-A</td>
</tr>
<tr>
<td>GermaNet</td>
<td>General</td>
<td>SS-A</td>
</tr>
<tr>
<td>WordNet</td>
<td>General</td>
<td>SS-A</td>
</tr>
<tr>
<td>(no resources defined yet for this use case)</td>
<td>Life sciences</td>
<td>LS-C(^\text{21})</td>
</tr>
<tr>
<td>FLAGdb++ resources</td>
<td>Agriculture</td>
<td>AS-E</td>
</tr>
<tr>
<td>Arabidopsis Information Resource (TAIR)</td>
<td>Agriculture</td>
<td>AS-E</td>
</tr>
<tr>
<td>The microRNA database</td>
<td>Agriculture</td>
<td>AS-E</td>
</tr>
</tbody>
</table>

---

18  [https://www.w3.org/annotation/](https://www.w3.org/annotation/)
20  [https://uima.apache.org/d/uimaj-2.9.0/references.html#ugr.ref.xmi](https://uima.apache.org/d/uimaj-2.9.0/references.html#ugr.ref.xmi)
21  The use case and associated resources were still being developed in the project as this document was being finalised. Resource metadata collection and checking of compliance with OpenMinTeD standards is therefore still ongoing.
2.2.4 Future work

- Work is ongoing to specify inclusion of knowledge resources in the OpenMinTeD registry.
- Work is nearing completion on selection of knowledge resources, and checking of their compliance.
- Work on an example workflow continues to evolve as the OpenMinTeD platform develops, and is expected to be completed in the following weeks.

2.2.5 State of operations

In the previous period, WG2 committed to engage more fully with use case partners. This has been the major achievement of the WG in this period, with all use case partners involved in developing the work with respect to selection of resources, and checking their compliance with project requirements.

2.3 WG3

2.3.1 Mission statement

The mission statement remains unchanged. For details, please refer to D5.2.

2.3.2 Mode of operation

The mode of operation remains unchanged. For details, please refer to D5.2.

---

2.3.3  Progress in the current period

The current period has been dedicated the following actions:

- **Licenses and rights statements:** The WG3 “Licenses and rights statements” working plan, implemented as reported in D5.3, has undergone further developments. In particular, regarding the legal annotation experiment, a set of licenses (for content and software) and terms of services has been annotated in full by at least two independent annotators. The related annotation guidelines, drafted in collaboration with WG1 to provide instructions for both internal and external annotators, are being expanded upon the preliminary results after curation.

- **Compatibility matrix:** The WG3 Compatibility matrix has been further developed and reviewed. An online questionnaire was also designed to help with the reviewing process. Aligning with the recommendations of reviewers, the structure of the matrix was edited and its content augmented and standardized to follow a consistent and uniform pattern. The working document that illustrates the Compatibility Tool (now version 1.2) was presented and discussed on the 28th of June 2017 during the CREATe IP Summer Summit at the University of Glasgow.

- **Models training and licensing issues:** Following intense discussions with other WGs about the concrete inference of the current legal framework on text and data mining activities, a new task was initiated in collaboration with WG1, which focuses on the problems related to the legal standing of Machine Learning (ML) models with uncertain copyright status. Aimed at investigating, first, whether training models on corpora is to be considered a copyright relevant activity and, second, what is the copyright status and the licensing conditions of trained output, the WG is currently looking at some particularly problematic cases especially those concerning models from OpenMinTeD partners but trained on external corpora as well as models from other sources.

- **Open Access and Open Science:** in collaboration with OpenAire and Creative Commons UK, the first version of two important documents in support of Open Access and Open Sciences was released, namely the Fact Sheet on Creative Commons & Open Science (Creative Commons UK 2017, of August 9, 2017) and the Frequently Asked Questions on Creative Commons & Open Access (Creative Commons UK 2017, of August 9, 2017). The former contains questions and responses to common concerns surrounding open science and the implications of licensing data under Creative Commons licenses. The latter includes questions and responses to common concerns surrounding Open Access, Creative Commons, and the publication of research. Both documents are intended to aid researchers, teachers, librarians, administrators and many others using and encountering Creative Commons licenses and the Open Access movement in their work.

- **Concrete requirements:** No further changes have been made to the concrete requirements for legal interoperability linked to WP5 Interoperability Specifications following D5.3.

---

23  https://openminted.github.io/releases/license-matrix/
24  https://doi.org/10.5281/zenodo.840652
25  https://doi.org/10.5281/zenodo.841086
• **Glossary**: In line with the updated OpenMinTeD glossary of interoperability terms designed by WG1, the legal glossary has been further edited to reflect and match the latest version of the former.

• **Dissemination and community engagement**: The set of legal Frequent Asked Questions (FAQs) designed under task T3.3 (led by WP3) for both training and policies purposes, as reported in D5.3 (in its version 0.1), has been revised and upgraded to version 1.0. Some of these FAQs were also included in the general FAQs published in the OMTD platform (beta version). The webinar “The Law on TDM in Europe: an introduction” was held on the 21st of March 2017 in collaboration with UKSG and the recording of the session are available upon registration at the UKSG website.

• **Policy**: the WG has recently provided a preliminary draft of the Terms and Conditions (v 1.0) for the use of the OpenMinTeD Platform.

2.3.4 Future work

• As anticipated in D5.3, one of the priorities of WG3 is to make an automated tool out of the static version of the Compatibility Matrix, with one of its main feature be to determine whether the content, the software or the service licensed or otherwise conditioned may be combined and their combination feasibly result in a derivative work.

• The flow chart (released under version 0.1) that illustrates some of the key legal dynamics applicable to text and data mining will be expanded and updated to version 1.0.

• In addition to the tasks so far undertaken, WG3 plans to conclude the Annotation experiment with the curation of results, which will be also used to complete the third layer of the compatibility matrix that is dedicated to terms of services.

2.3.5 State of operations

The activity of the WG proceeds in alignment with the wider working plan, with regular contribution to the reports initiated by WP5 and other WPs.

The temporary decrement of external expert participation in WG3 con-calls, which was first reported in D5.3, has been successfully tackled with more tailored con-calls and individual involvement of external experts.

A minor delay in delivering the WG3 interoperability tools, in particular the compatibility matrix, mainly due to the complex and scrupulous revision of its contents and structure, has been overcome and the WG3 is ready to disclose their latest version.

2.4 WG4

[26](https://vocabularies.openminted.eu/skosmos/omtd_glossary/en/)

[27](https://services.openminted.eu/support/faqLegal)

[28](https://www.uksg.org/webinars/lawontdm)

[29](https://services.openminted.eu/assets/files/OpenMinTeD_Terms_and_Conditions_v1.0.pdf)
2.4.1 Mission statement
The mission statement remains unchanged. For details, please refer to D5.2\textsuperscript{30}.

2.4.2 Mode of operation
The mode of operation remains unchanged. For details, please refer to D5.2.

2.4.3 Progress in current period
The current period has been dedicated the following actions:

- **Interoperability requirements**: work on the interoperability requirements has continued. Existing requirements have been partially refined or deprecated as some were found to be rather functional requirements e.g. for the OpenMinTeD Registry Service or OpenMinTeD Workflow Service. New concrete requirements were defined such that most of the abstract requirements are now paired with at least one concrete requirement explaining a specific approach or technology supported by OpenMinTeD that can be used to meet the abstract requirement.

- **Dockerized TDM components**: All TDM components in OpenMinTeD are required to be in Docker images; dockerization helps in avoiding several interoperability problems that arise when
  - a “difficult-to-deploy” TDM component has to be used in a workflow; i.e., a component that has various software dependencies.
  - we want to combine TDM components in a workflow that are built with different frameworks (UIMA, GATE, AlvisNLP), programming languages (Java, Python), operating systems, etc.
  - Docker images are easily portable and are used to run software in virtual containers, in OMTD this facilitates the definition and execution of complex workflows in a cluster (cloud infrastructure) or a (single) local machine. For more information on Docker see “Interoperability Landscaping Report” (D5.1). In this direction, we defined a **Docker interface** for OMTD TDM components, we designed an **automatic dockerization** process for non-dockerized components and we defined a common **Galaxy xml tool generation process** from OMTD-SHARE files.

- **Docker Interface**: Work to provide a common specification for an interface for Docker to expose\textsuperscript{31}. This will ensure that people wishing to make their components and workflows compliant with OpenMinTeD will be able to construct their Docker images in a way that is interpretable by the OMTD workflow engine. It also contains specifications for the OMTD-SHARE descriptor, specifying what elements are mandatory when uploading a new component.

- **Automatic Dockerization**: For UIMA and GATE components that are provided via Maven central repository and are not dockerized we (OMTD platform/registry) automatically dockerize them by adding in a Docker image the required dependencies and software (following the Docker interface spec.). For example, we add


\textsuperscript{31} https://openminted.github.io/releases/docker-spec/0.1.0/specification
- Java.
- the executors for executing a UIMA or GATE component.
- fetchers for downloading (if needed) the maven dependencies of a component.
- The Docker images that we produce along with the Galaxy XML wrappers (see below for more information on them) are the two things needed for executing a component in Galaxy engine.

**Galaxy XML wrappers generation:**

- Definition\(^{32}\): *The XML File for a Galaxy tool, generally referred to as the “tool config file” or “wrapper”, serves a number of purposes. First, it lays out the user interface for the tool (e.g. form fields, text, help, etc.). Second, it provides the glue that links your tool to Galaxy by telling Galaxy how to invoke it, what options to pass, and what files it will produce as output.*

- A first version of a generator that produces Galaxy tool XML files is now available at GitHub\(^{33}\). The generator takes as input the OMTD-SHARE metadata for a OMTD component (e.g. a UIMA or GATE component) and creates a tool XML file with all the information that are required in Galaxy UI and for executing the tool with Galaxy execution engine. For example, based on the metadata we can set/specify for Galaxy:
  - the id, name and version of the tool
  - a description of it.
  - the command for calling/executing the tool
  - the input parameters and attributes of them; e.g. optional or not, default values, descriptions of them.

**Web service framework:** A framework providing a seamless interface for integrating UIMA components as web services has been specified\(^{34}\) and implemented\(^{35}\). It can be used when deploying a component requiring a large model (i.e., to avoid repeated initialization), or when licensing issues prevent a component from being distributed outside of an institution.

**AERO protocol:** The first version of the Annotation Editor Remote Operation (AERO) Protocol\(^{36}\) has been finalized. The protocol can be used to provision annotation protocols into project-oriented text annotation editors such as WebAnno or AlvisAE. The upcoming WebAnno 3.3.0 will include the first implementation of the protocol. Later in the project, the OpenMinTeD Registry is meant to use AERO to communicate with the Annotation Editor Service. AERO provides functionalities to manage annotation projects, documents to be annotated, annotations created by users, as well as curated annotations. The RESTful protocol leaves certain aspects of the data being exchanged with the

---

\(^{32}\) Source: https://docs.galaxyproject.org/en/master/dev/schema.html

\(^{33}\) https://github.com/galanisd/omtd-workflows-executor/tree/master/omtd-component-galaxywrappers

\(^{34}\) https://openminted.github.io/releases/processing-web-services/1.0.0/specification

\(^{35}\) https://github.com/argo-nactem/omtd-remote-execution

\(^{36}\) https://openminted.github.io/releases/aero-spec/1.0.0/omtd-aero/
AERO-enabled service underspecified, e.g. which concrete document or annotation formats are supported. While this could be perceived as a gap in the interoperability provided by AERO, it is actually by design. The underspecification in AERO 1.0.0 facilitates the implementation of the protocol in AERO servers and to offer clients direct access to proprietary APIs. We conjecture the underspecification may be partially removed after AERO has found some adoption in the community and readiness for further consensus has formed in the community.

2.4.4 Future work

- Further concretisation and refinement of the interoperability requirements.
- Suggest a common type system.

2.4.5 State of operations

- WG is performing strongly with regular conference calls well attended by OpenMinTeD partners.
- Communication with external experts happens largely through individual members of the WG. External experts do not participate directly in the conference calls, but they do sometimes contribute to the Redmine forum.
- Given the strongly growing number of requirements (specifically from WG1 and WG4), an approach needs to be defined how to organize and/or refine these to make them better accessible from a user’s perspective.
- This group has now postponed future calls, pending the end of WP5. Further interoperability related questions will be addressed through ongoing work and calls in WP6.
3. Requirements

For the preparation of the present deliverable, we have focussed on creating specifications for those areas that were found not to be sufficiently covered during the analysis and compliance evaluation of the requirements defined as part of the previous versions of this deliverable (D5.2\textsuperscript{37} and D5.3\textsuperscript{38}). An updated version of the interoperability requirements will be published online along with updates to the compliance assessments as part of the next “Progress Report of the Alignment of Service and Content Provider Systems” milestone (MS35).


\footnotesize{\textsuperscript{38} http://openminted.eu/wp-content/uploads/2017/06/D5.3_Interoperability-Standards-and-Specifications_v1.0.pdf}
4. Compliance

In the previous section, we discussed the requirements for interoperability that WGs in OpenMinTeD have identified so far. But unless relevant products are compliant with these, the requirements are ineffective.

In D5.3, the previous and first version of these interoperability standards and specifications, compliance levels were described and a list of products given with the number of requirements against which they were assessed. Since that report, a fuller compliance assessment was given in the MS34 report, “Progress Report of the Alignment of Service and Content Provider Systems”. This included detailed breakdowns of product compliance produced from the requirements GitHub site, and a set of recommendations on future compliance checking. In keeping with this approach, compliance will again be reported in the next relevant milestone report, MS35.

In summary, the conclusions of the MS34 report, as adopted after discussion within WP5. were:

- There is a statement in the specification that “If a product does not appear at all in a compliance table, then that is because there never was any question that it is not applicable and no justification is required.” However, this means it is unclear when a product compliance is not recorded against a requirement, whether this is because it is not applicable, or because recording has failed. The “not applicable” compliance level should be enforced in all cases.
- The “platinum” level of compliance does not seem to be especially relevant, given the low number of optional requirements.
- Consider the creation of further tooling to support the compliance process, adapting what we already have, and weighing up the required effort and expected gain. Tooling should:
  - Enforce a “not applicable” compliance level, so that whether or not a requirement is relevant to a product is recorded (including a justification).
  - Add a field signifying which type(s) of product a requirement applies to (e.g. content, services, resources) - as not all requirements are relevant to all types of product
  - Add a formally defined link between abstract requirements and their concrete equivalents.
5. List of attachments

New for this deliverable

- Dockerization of Components v0.1.0
  - https://openminted.github.io/releases/docker-spec/0.1.0/specification
- OMTD Processing Web Service API Specification v1.0.0
  - https://openminted.github.io/releases/processing-web-services/1.0.0/specification
- OpenMinTeD AERO Protocol v1.0.0
  - https://openminted.github.io/releases/aero-spec/1.0.0/omtd-aero/

Updated for this deliverable

- Legal FAQ
  - https://services.openminted.eu/support/faqLegal
- License compatibility matrix
  - https://openminted.github.io/releases/license-matrix/
- Terms and Conditions
  - https://services.openminted.eu/assets/files/OpenMinTeD_Terms_and_Conditions_v1.0.pdf
- Fact Sheet on Creative Commons & Open Science
  - https://doi.org/10.5281/zenodo.840652
- Frequently Asked Questions on Creative Commons & Open Access
  - https://doi.org/10.5281/zenodo.841086

No change since D5.3

- Detailed Interoperability Specification v1.1.0
  - https://openminted.github.io/releases/interop-spec/1.1.0/openminted-interoperability-spec/
- Detailed Interoperability Scenarios v1.0.0
  - https://openminted.github.io/releases/interop-spec/1.0.0/openminted-interoperability-scenarios/
- Detailed type system alignment v1.0.0
  - Can presently not be included as PDF because of technical reasons
  - https://openminted.github.io/releases/interop-spec/1.0.0/typealignment/
- Detailed overview of components from partners involved in WG4 v1.0.0
  - Can presently not be included as PDF because of technical reasons
  - https://openminted.github.io/releases/interop-spec/1.0.0/components/
- OpenMinTeD Metadata Scheme v2.0.1
  - https://openminted.github.io/releases/omtd-share/2.0.1/
6. **Appendix**

6.1 **content-connector-openaire**

The connection with **OpenAIRE** is accomplished via the following mechanism:

- Replication at the OpenMinTeD premises of the Solr index used at OpenAIRE for storing the metadata records they are harvesting; during this process, filtering out of resources that are not of interest to OpenMinTeD (e.g., closed access resources or publications without a link to full text and/or abstract) is also performed.
- Installation of an additional local index for linking the contents of each document with its metadata record, which is further enriched with information about the full text: path, hash key and data format (mime type).

It should be noted that the decision to replicate the metadata records and full texts in OpenMinTeD was taken in the second phase of the project, in order to facilitate and speed up the querying and retrieval processes.

The content-connector-openaire, as mentioned above, implements the content-connector-api interface which supports the **OpenMinTeD corpus building process**, through three main methods: search, fetchMetadata, downloadFulltext.

6.1.1 **Search**

This method queries the Solr index and returns all the values used for the metadata fields that have been selected from the OMTD-SHARE minimal schema\(^\text{39}\) for the faceted browsing of the scholarly works, namely *Publication Year*, *Access rights/License*, *Publication Type*, *Document Language*, *Document Type* and the *Content Source* (i.e. OpenAIRE or CORE, for the time being); all the values statistical information on the contents of the registry, i.e. the amount of scholarly works in total and per content provider (OpenAIRE and CORE)

Users can select a subset of the works through a google-like free text field or by clicking on specific values from one or more of the facets ("faceted query").

For input, **search** gets a Query object describing a query (Figure 1) and it returns an object with information drawn from the the metadata results (Figure 2). An example of a request is the following: https:.../searchForPublications;publicationYear=2013;query=DIGITAL;licence=Open%20Access;publicationType=Research

For the faceted query, the user uses the OMTD-SHARE metadata schema; thus, there is a mechanism that converts these to the equivalent OpenAIRE metadata for the querying and back again when retrieved (cf. Table 1 of mappings). The conversion is performed on-the-fly without storage of the converted metadata records anywhere.

### 6.1.2 fetchMetadata

It queries the Solr index of OpenAIRE and it returns all the metadata that are retrieved by the query. Like in the above search method, the initial query is provided by the user with OMTD-SHARE values and a mechanism is responsible for converting them to OpenAIRE and committing the query to the Solr index.

Each metadata record that is retrieved from the Solr index is then converted to the corresponding OMTD-SHARE metadata record (in XML format). This method is mainly used in the corpus building process and the retrieved metadata records are stored for the user to use.
6.1.3 downloadFulltext
It queries the Solr index for a particular document via its metadata record identifier and it returns it. This method too is mainly used in the corpus building process and the retrieved documents are stored as part of the corpus that the user can further process.

6.2 content-connector-core
The connection of OpenMinTeD registry to CORE is achieved through an implementation of the interface defined in content-connector-api. This API essentially consists of the following 3 main functions: search, fetchMetadata, downloadFulltext.

6.2.1 search
The function receives as an input a Query object which consists of the user query keyword, the desired facets, parameters and offset/size (used for pagination in the OMTD platform). This is then translated into a filtered query of the backend ElasticSearch of CORE. An example of this conversion (OMTD query to CORE's ElasticSearch query) is given in Figure 3.

<table>
<thead>
<tr>
<th>OMTD query:</th>
</tr>
</thead>
<tbody>
<tr>
<td>{</td>
</tr>
<tr>
<td>&quot;keyword&quot;: &quot;Deep Learning&quot;,</td>
</tr>
<tr>
<td>&quot;params&quot;: {</td>
</tr>
<tr>
<td>&quot;documentLanguage&quot;: [&quot;english&quot;, &quot;german&quot;],</td>
</tr>
<tr>
<td>&quot;publicationYear&quot;: [&quot;2012&quot;, &quot;2013&quot;]</td>
</tr>
<tr>
<td>},</td>
</tr>
<tr>
<td>&quot;facets&quot;: [&quot;publicationYear&quot;, &quot;documentLanguage&quot;, &quot;publicationType&quot;],</td>
</tr>
<tr>
<td>&quot;from&quot;: 1,</td>
</tr>
<tr>
<td>&quot;to&quot;: 15</td>
</tr>
<tr>
<td>}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CORE’s elasticsearch query</th>
</tr>
</thead>
<tbody>
<tr>
<td>{</td>
</tr>
<tr>
<td>&quot;query&quot;: {</td>
</tr>
<tr>
<td>&quot;filtered&quot;: {</td>
</tr>
<tr>
<td>&quot;query&quot;: {</td>
</tr>
<tr>
<td>&quot;bool&quot;: {</td>
</tr>
<tr>
<td>&quot;must&quot;: [{</td>
</tr>
<tr>
<td>&quot;query_string&quot;: {</td>
</tr>
</tbody>
</table>
|        "query": "Deep learning"
|      }
|     }
|    }
|   }
|  }
|  "filter": { |
"bool": {
  "must": [
    "exists": {
      "field": "fullText"
    },
    
    "term": {
      "deleted": "ALLOWED"
    },
    
    "bool": {
      "should": [
        "term": {
          "language.name": "english"
        },
        
        "term": {
          "language.name": "german"
        },
        
        "bool": {
          "should": [
            "term": {
              "year": "2012"
            },
            
            "term": {
              "year": "2013"
            }
          ]
        }
      ]
    }]
  }
}
Figure 3 - OMTD query to CORE's ElasticSearch query

The result set is then converted into the OMTD-SHARE schema following the recommended guidelines for publications\textsuperscript{40}. Details on this conversion on the granularity level of each field is given (aggregated with OpenAIRE conversion information) in Table 1.

6.2.2 fetchMetadata
The function receives a similar query object like search method and retrieves in serialized form all the records (in OMTD share schema) matching the specified query. In order to overcome deep-pagination problems, conventional pagination (with offset and page size) is avoided. Instead a scan-and-scroll query is used (equivalent to Solr’s cursor) allowing for fetching big result sets regardless the size (the method was successfully tested for result sets up to 50k records).

\textsuperscript{40} https://guidelines.openminted.eu/guidelines_for_providers_of_publications/recommended_schema_for_publications.html
6.2.3  downloadFullText
The function receives an identifier literal and retrieves the PDF (fullText) associated with the record identified by it. Essentially it queries CORE for records matching this identifier, retrieves the CORE identifier of the record and then constructs a download URL for this record that is used to get the record.