Interoperability Standards and Specifications Report

December 11, 2016

Deliverable Code: D5.2
Version: 1.2
Dissemination level: Public

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

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

D5.2 – Interoperability Standards and Specifications Report

WP5 – Interoperability Framework	
WP participating organisations: UKP-TUDA, ARC, UNIMAN, INRA, OU, USFD, UvA, UoG, GESIS	
Contractual Delivery Date: 07/2016	**Actual Delivery Date:** 12/2016
Nature: Report	**Version:** 1.2
Public Deliverable	

Preparation slip

	Name	Organisation	Date
From	Richard Eckart de Castilho Mouhamadou Ba Penny Labropoulou Thomas Margoni Giulia Dore Wim Peters Matthew Shardlow Piotr Przybyla Jacob Carter	UKP-TUDA INRA ARC UoG UoG USFD UNIMAN UNIMAN UNIMAN	28/07/2016
Edited by	Richard Eckart de Castilho	UKP-TUDA	11/12/2016
Reviewed by	Robert Bossy Vangelis Floros John McNaught	INRA GRNET UNIMAN	25/10/2016
Approved by	Natalia Manola	ARC	11/12/2016
For delivery	Mike Hatzopoulos	ARC	13/12/2016

Document change record

Issue	Item	Reason for Change	Author/Editor	Organisation
V0.1	Draft version	Initial version sent for comments	Richard Eckart de Castilho	UKP-TUDA
V0.2	Draft version	Incorporated comments from reviewers	Richard Eckart de Castilho	UKP-TUDA
V1.0	First version	Finalizing document	Richard Eckart de Castilho	UKP-TUDA
V1.1	Revised version	Incorporating feedback from the approving instance	Richard Eckart de Castilho	UKP-TUDA
V1.2	Revised version	Incorporating feedback from the approving instance	Richard Eckart de Castilho	UKP-TUDA

1. Table of Contents

1. INTRODUCTION 8

1.1 METHODOLOGY 8
1.2 WORKING GROUPS 8

2. SUMMARY REPORTS 9

2.1 WG1 – RESOURCE METADATA 9
2.2 WG2 – KNOWLEDGE RESOURCES 12
2.3 WG3 – IPR AND LICENSING 13
2.4 WG4 – ANNOTATION AND WORKFLOWS 16
2.5 PUBLICATIONS 17
2.6 EXTERNAL EXPERTS 19

3. SCENARIOS 20

4. REQUIREMENTS 21

4.1 REQUIREMENT STRUCTURE 21
4.2 REQUIREMENTS OVERVIEW 22

5. COMPLIANCE 27

5.1 COMPLIANCE LEVELS 27
5.2 COMPLIANCE ASSESSMENTS 27

6. ACTIONS 29

6.1 WG 1 29
6.2 WG 2 31
6.3 WG 3 32
6.4 WG 4 33

7. LIST OF ATTACHMENTS 36

8. APPENDIX 37

8.1 OPENMINTED COMPONENT CLASSIFICATION (DRAFT) 37
8.2 WG1 - INVENTORY OF METADATA SCHEMAS AND RELATED EFFORTS 41
8.3 WG3 – COMPATIBILITY MATRIX: SUMMARY 46

2. Table of Tables

Table 1 - Requirements in status "draft" 23
Table 2 - Requirements in status "final" 24
Table 3 – Requirements in status "deprecated" 26
Table 4 - Assessed products and consulted sources 28
Table 5 – WG 1 summary of actions to improve compliance 30
Table 6 – WG 2 summary of actions to improve compliance 32
Table 7 – WG 3 summary of actions to improve compliance 33
Table 8 – WG 4 summary of actions to improve compliance 34
Table 9 - Compatibility Matrix (draft version 1.0): Contents 50
Table 10 - Compatibility Matrix (draft version 1.0): Software 51
Table 11 - Compatibility Matrix (draft version 1.0): Terms of Service 51
Table 12 - Compatibility Matrix (draft version 2.0): Concent 52
Table 13 - Compatibility Matrix (draft version 2.0): Software 53
Table 14 – Compatibility Matrix (draft version 2.0): Terms of Service 53

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.int/)

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

Acronyms

ARC	Athena Research Center; see ILSP
CC	Creative Commons (ttps://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 (https://jats.nlm.nih.gov)
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 (https://en.wikipedia.org/wiki/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 (https://en.wikipedia.org/wiki/Resource_Description_Framework)
SKOS	SKOS - Simple Knowledge Organisation System (https://en.wikipedia.org/wiki/Simple_Knowledge_Organisation_System)
TDM	Text and Data Mining
TheSOZ	Thesaurus for the Social Sciences (http://lod.gesis.org/thesoz/de.html)
UIMA	Unstructured Information Management Architecture; usually referring to the reference implementation Apache UIMA (https://uima.apache.org)
UNIMAN	University of Manchester, UK
UKP-TUDA	Ubiquitous Knowledge Processing (UKP) Lab, Technische Universität Darmstadt, Germany
UoG	University of Glasgow, UK
USFD	University of Sheffield, UK
WG	Working group
WP	Work package
XSD	XML Schema Definition (https://en.wikipedia.org/wiki/XML_Schema_(W3C))

Publishable Summary

The goal of the Interoperability Standards and Specifications report is to assess and 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 first in a series of three. It will be updated in M20 (D5.3) and M26 (D5.4). This report focusses on presenting a high-level overview of the progress achieved within the reporting period and on actions planned for the next period. The actual work documents released during the reporting period are provided as attachments to this deliverable.

1. Introduction

In this section, we briefly revisit the methodology by which WP 5.2 operates and the constitution of the interoperability working groups.

1.1 Methodology

In the milestone document MS5 "Working groups external experts list and work methodology", we outlined the methodology for building the interoperability specification. Since MS5 is presently only available on the project-internal wiki, we repeat the key concepts here:

1. **Scenario definition** – The WGs have created a set of 17 interoperability scenarios that highlight different aspects of interoperability from the perspective of the individual WGs (Section 3).
2. **Analysis** – The discussion of these scenarios together with external experts was the focus of the first OpenMinTeD Interoperability Workshop (cf. MS10 "Working groups interim meeting report I"). Furthermore, the scenarios have been analysed in the WGs to generate the requirements presented in Section 4.2.
3. **Prototype** – Several efforts are being undertaken to assess the feasibility and effort of implementing the proposed requirements and to provide further insight for their refinement. These efforts are listed in the summary reports (Section 2) of the respective WGs in the subsection "Progress in current period".
4. **Evaluation** – In order to evaluate the proposed requirements, we identified relevant products (e.g. TDM frameworks of the involved project partners) and assessed their compliance with our requirements. Based on the evaluation, we generated a set of actions designed to improve compliance with the interoperability requirements. These actions are meant to serve as a roadmap for the next reporting period. More details on the compliance assessment are provided in Section 4.3.
5. **Specification** – The requirements specification is a living document, continually updated as requirements are created, refined, and deprecated. Section 4.1 provides additional information about the requirement lifecycle.

The process was designed to ensure the participation of all stakeholders. It also pays attention on tightly involving those stakeholders that later may need to adjust their products in order to become compliant with our requirements.

1.2 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

2. *Summary Reports*

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

- **Mission statement** – short updated summary of the working group's mission statement
- **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 reporting period
- **Tasks planned for next period** – summary roadmap of the tasks planned for the next reporting period
- **State of operations** – short self-assessment of the current state of operations

The interoperability working groups are:

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

2.1 WG1 – Resource metadata

2.1.1 Mission statement

The focus of WG1 lies on the metadata required for describing resources targeted by the OpenMinTeD project in order to ensure their discoverability and achieve interoperability between them.

TDM involves a wide range of resource types: the resources to be mined (scholarly publications in the project), the text mining/language processing software per se and ancillary knowledge resources used for its operation (e.g. annotation schemas, linguistic tagsets, ontological resources used for annotating the resources to be mined, annotated textual corpora).

To describe these resources a core set of metadata elements can be used to capture their common properties (e.g. administrative information, such as contact details and identification data), while various sets of elements encode the particular properties they display (e.g. size and format for content resources vs. input specifications for components). Since processing activities involve the interaction of these resources, a subset of the resources' properties need to be described with the same vocabulary (e.g. the language of the contents of a publication and the language a tool or service can process, or the domain of a thesaurus and the domain of a publication). The definition and harmonisation of these metadata elements is the main objective of WG1. This endeavour is further hampered by the fact that these resources are the object of work for experts coming from different disciplines, with different theoretical backgrounds and conceptualisation of their work, often using different terms for the same or similar concepts. The clarification of these concepts and their semantic mapping as a means to establish a "common" vocabulary poses a challenge for WG1.

Interoperability for WG1 is, therefore, sought at two levels:

- **per resource type** - i.e. mapping metadata elements used by different schemas to describe the same property,

- **across resource types** - i.e. ensuring that the same metadata elements are used to describe their intersecting features.

2.1.2 Mode of operation

WG1 brings together experts from the different communities involved in the project, combining expertise in various fields: publishers, aggregators of scholarly publications, infrastructure specialists, developers of language processing and/or text mining services, experts in the creation and/or representation of language and knowledge resources, metadata specialists, legal experts etc.
The group holds regular teleconference calls where internal and external experts are invited; depending on the topic of the discussion, the attendance varies with a nucleus of the experts always present and a further set joining when the discussion relates to their particular expertise. Representatives from the other OpenMinTeD WPs (e.g. on use cases) also attend the meetings when in line with their interests. Extra-regular meetings dedicated to specific issues (e.g. metadata schemas of publications) have also been held. In addition, close collaboration is also sought with the other three working groups to ensure that their requirements as regards metadata encoding are properly met; attendance of their teleconference calls and working documents provides the appropriate input.
The group has created an inventory of metadata schemas and related efforts (taking into account the Deliverable D5.1 – Interoperability Landscaping Report) that present more interest for the WG1 objectives - cf. Section 8.2. The discussions of the group have focused on the contents of these schemas and on the interoperability requirements that were extracted from WP5.2 scenarios (cf. Section 4).
Finally, a document in the form of a working report, is collectively drafted.

2.1.3 Progress in current period

WG1 has produced the following:

- a set of **21 requirements for interoperability**[1] of the metadata descriptions between the various resource types. 8 additional requirements were generated but left for reconsideration for the next reporting period;
- a **selection of resources**, that will be directly involved in the project given that they belong to the partners or identified by them as standard for our purposes, has been assessed for compliance to the requirements:
    - OpenAIRE[2], CORE[3] and Frontiers schemas for describing publications;
    - TheSOZ[4], AGROVOC[5], JATS[6], OLIA[7] and LAPPS Grid[8] as knowledge resources;
    - a set of standard licences (e.g. CC, FOSS licences);

[1] https://openminted.github.io/openminted-site/releases/interop-spec/1.0.0/openminted-interoperability-spec.html#WG1
[2] https://www.openaire.eu
[3] https://core.ac.uk
[4] http://lod.gesis.org/thesoz/en.html
[5] http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus
[6] https://jats.nlm.nih.gov
[7] http://www.acoli.informatik.uni-frankfurt.de/resources/olia/
[8] http://www.lappsgrid.org

- Alvis NLP[1], Argo[2], DKPro Core[3] and ILSP software components[4];

- an **inventory of metadata schemas, vocabularies and ontologies** used for the description of these resource types (cf. Section 8.2)
- discussions taking as a base **the main metadata schemas** used by the consortium partners, i.e. META-SHARE for corpora, knowledge resources and software components, and OpenAIRE and CORE for publications; these, together with the WG1 interoperability requirements, the description of components performed by WG4 and the requirements of WG2, have been used as the basis for the Reference Metadata Schema of OpenMinTeD;
- a proposal for the **classification of software components** used in TDM processes (Section 8.1);
- the **first version of the OpenMinTeD Reference Metadata Schema in XSD**[5]; the schema covers in a harmonized way all the resource types of OpenMinTeD and caters for their satellite entities; the schema is currently under review by the members of all WGs and planned to be used for the registry first version.
- Publication (jointly with WG3): **Legal Interoperability Issues in the Framework of the OpenMinTeD Project: A Methodological Overview** (for details on the publication and a link, see Section 2.5)

2.1.4 Tasks planned for next period

WG1 will continue to work along the action lines that it has already initiated:

- finalise the interoperability requirements: a set of requirements has already been identified but the discussion showed that they are not mature yet to formulate with a unanimous consent; we expect also that analysis of other sources (e.g. WP4 requirements, use of the schema in the registry) will generate more requirements;
- update the inventory of schemas and vocabularies, as required;
- continue the compliance assessments and recommend ways of improving metadata descriptions based on their outcomes;
- evaluate the reference metadata schema by creating descriptions from scratch or by mappings from the currently used schemas for the resources of the consortium to populate the registry and improve it accordingly;
- document the schema in a user-friendly way and formulate recommendations and guidelines that will promote interoperability as identified in the requirements;
- work on mappings and walkthroughs with the most popular metadata schemas, according to the principles set by WG2.

2.1.5 State of operations

WG1 has established a very good level of operation and is on schedule.

[1] http://www.quaero.org/module_technologique/alvis-nlp-alvis-natural-language-processing/
[2] http://argo.nactem.ac.uk
[3] https://dkpro.github.io/dkpro-core/
[4] http://nlp.ilsp.gr/ws/
[5] https://openminted.github.io/openminted-site/releases/omtd-share/1.0.0/html/index.html

2.2 WG2 – Knowledge resources

2.2.1 Mission statement

Working Group 2 targets the interoperability of knowledge resources. Knowledge is specific information that is relevant for the linguistic and conceptual interpretation of text and the content exchange between TDM modules. This information is either exploited or produced by TDM modules and tools.
The definition encompasses a variety of resource subtypes:

- Language resources such as annotated text corpora;
- Ancillary resources for conceptual/linguistic interpretation within the TDM workflow such as lexicons, term banks, ontologies, thesauri and dictionaries.
- Processing resources that produce knowledge such as text mining tools/services like part of speech tags, dependency relations, vocabulary lookup and statistical classifiers.

WG2 aims to:

- Tackle semantic interoperability issues when integrating knowledge resources (linguistic, terminological and ontological resources) with TDM workflows. These issues arise a) when the same domain concept may be defined in different ways in knowledge resources, b) when the concept is modelled differently within a TDM component and a related knowledge resource. One example would be two ontologies on chemical compounds specifying these compounds at different level of detail and using different controlled vocabularies. Another example would be that the part-of-speech tag necessary to disambiguate a word during a dictionary lookup is using a different set of tags in the dictionary than it is produced by an automatic part-of-speech tagger.
- Define a specification for the representation of knowledge resource types such as lexicons, terminological sources, thesauri, ontologies, annotated corpora and tool outputs.
- Establish interoperability across different resources/tools for the purpose of their exploitation by text and data mining (TDM) applications.

The focus of this group is on ensuring the discoverability, interoperability and consistency of linguistic, terminological and ontological content at the granular representation level of individual knowledge elements. This knowledge is either contained within resources or produced by language processing and text mining tools. Its interoperability will foster common understanding, data sharing and reuse.
For this purpose, the group will establish a network of (de facto) standard reference vocabularies for the representation and linking of information elements required for interoperable text consumption and processing.

2.2.2 Mode of operation

The group consists of a number of internal experts, a subset of whom attends regular monthly teleconferences. The work that is discussed and distributed over the participating partners is done by means of the collaborative drafting of a document whose structure reflects that of the present report and its future iterations. We also use collaboratively maintained spreadsheets for the collection of knowledge resource schemas, linking information, and requirement formulation.
External experts are mostly consulted on a personal basis, and targeted for their particular expertise.

2.2.3 Progress in current period

WG2 has worked on the interoperability of Knowledge Resources (KRs): resources containing, producing or representing knowledge. A first set of schemas has been selected, which can be divided up into several subtypes:

- Schemas associated with OpenMinTeD contributed by OpenMinTeD partners (i.e. UIMA-based platforms, ALVIS, and GATE)
- Strategical, i.e. widely used and interconnected (de facto) standard vocabularies for linguistic/terminological/ontological metadata
- Representative set of use case driven schemas
- Publication: **Tackling Resource Interoperability: Principles, Strategies and Models** (for details on the publication and a link, see Section 2.5)

Interoperability between these schemas is in the process of being defined in terms of a number of linking relations based on OWL[1]/RDF[2] and SKOS[3] relations.
Once completed, the linked vocabularies form a reference network of linguistic, terminological and ontological conceptual elements that forms the core vocabulary for TDM information exchange.
Requirements for KR operationalisation within OpenMinTeD have been formulated. Eight of the KRs listed above have been checked with respect to their compliance with the requirements identified so far (cf. Section 5.2).

2.2.4 Tasks planned for next period

- Selection of a standard for annotation interoperability at the level of end-to-end system output
- The further selection of particular candidate standards for the provision of a core set of data category elements for data integration and exchange.
- The continuing creation of links between the elements of these vocabularies
- Requirement extension/adjustment
- Experimentation with the operationalisation of the schema network
- Incremental and collaborative creation of a draft specification report.

2.2.5 State of operations

The group operates efficiently and produces expected outputs according to the plan. The selection of use case driven resource schemas was performed on the basis of the WG2 member expertise. Alignment with other work packages, specifically WP4, has been established.

2.3 WG3 – IPR and licensing

2.3.1 Mission statement

The goal of WG3 "IPR and licensing" is to study and identify copyright and related rights (e.g. sui generis database right) restrictions and exceptions to the use and reuse of sources (both textual sources and

[1] https://www.w3.org/OWL/
[2] https://www.w3.org/RDF/
[3] http://www.w3.org/2004/02/skos/

text-mining services) in TDM activities. On this basis the WG will also identify contractual tools and schemes (e.g. licences) that can best serve the needs of TDM services.
In particular, it will examine which exceptions are currently available (e.g. the newly implemented TDM exception in the UK), which are upcoming and whether the current/proposed solutions embrace all the needs of the scientific and academic sector (e.g. is the non-commercial limitation necessary?).
The working group also focuses on the issue of legal compatibility and interoperability of licenses, aiming at determine whether multiple licenses that apply to different components can be deemed compatible and legal interoperable, particularly when there is the need to assess if the result (combination of components under difference licensing terms) can be redistributed or not.
Additionally, open licensing models for both the scientific related textual sources and the text-mining services will be explored and evaluated, by means of specific tools such as graphical representations of licenses compatibility (to be identified as compatibility matrix) and workflows that will guide the end users to choose the best applicable license and determine what licensing restrictions or rights statements limitations, if any, apply to specific uses.

2.3.2 Mode of operation

WG3, while focusing on legal interoperability issues, brings together experts from a variety of fields. These include: legal studies, publishers, technical experts (computer scientists, metadata experts, etc.), policy makers, academics, representative of different communities, groups and initiatives internationally active in the field.
WG3 regularly organise conference calls with external experts (once a month), internal experts (once a month) and dedicated conference calls with WG1 plus selected experts to discuss the specific issue of licence/right statements and metadata representation (again once a month), for a total of three conference calls a month.
Agenda items, minutes and summaries of all conference calls are maintained on the dedicated website.
WG3 maintains an updated list of working documents which include an inventory of licences and terms of use submitted by all the consortium members, which form the basis for another document dedicated to the compatibility of the licences and terms of use. A detailed bibliography of scholarly publications and policy documents is also maintained. Additionally, a glossary collecting the most recurring legal concepts with a brief explanation is likewise available.

2.3.3 Progress in current period

The group has two main goals: favouring licence compatibility and clarifying the legal landscape in the field of TDM. All this, with a view to the needs of TDM researchers, which implies the need to develop documents and tools that can be readily used by laymen.
The following items summarise the accomplishment achieved so far are:

- **Licence-related interoperability requirements**[1]
- **Licence Compatibility matrix** (see Section 8.3) - a schematic representation that represents (a) the type of data (contents, software and terms of services) and (b) the type of licences and/or rights statements, to determine whether or not there is compatibility between resources under respective licences. This matrix aims at facilitating the choice for users for the best licence to use and

[1] https://openminted.github.io/openminted-site/releases/interop-spec/1.0.0/openminted-interoperability-spec.html#WG3

share/distribute resources and particular TDM workflow results which may have been generated from multiple sources under heterogeneous licences. When multiple licences could be applied, also displays in brief what could be the legal implications of choosing one or the other licence.

- **Legal metadata and rights statement** (this document is still work in progress and not included with the present deliverable) - a document drafted in collaboration with WG1 that illustrates the roadmap to address the inference of legal metadata elements and rights statements for the purpose of TDM activities. This roadmap articulates in the following actions: (a) identifying applicable rights statements and build an OpenMinTeD inventory; (b) categorising these findings and comparing them with similar inventories (e.g. Europeana, OpenAIRE and CORE, but also CLARIN and META-SHARE); (c) identifying a common vocabulary while also (d) contemplating their machine-readability.
- **Licences <-> Rights Statements** (this document is still work in progress and not included with the present deliverable) - a synthetic representation of licences and rights statements' conditions to help users understand what a given license or a set of licenses allow them to do – including what they are required to do to properly perform their activities under those licensing terms – and what limitations or restrictions may apply to the use they wish to make of the resource..
- Publication (jointly with WG1): **Legal Interoperability Issues in the Framework of the OpenMinTeD Project: A Methodological Overview** (for details on the publication and a link, see Section 2.5)
- Publication: **Why We Need a Text and Data Mining Exception (But it is Not Enough) (conference extended abstract)** (for details on the publication and a link, see Section 2.5)

2.3.4 Tasks planned for next period

The main question to be addressed together with the other WG regards the "granularity" of the representation of legal information. In other words, whether the legal rules and the connected metadata should be represented at the licence level, or deconstructed further at the level of right statements. There is an ongoing discussion with internal and external experts (including the representatives of international projects active in this field) about the desirability and feasibility of a "rights statement" implementation.

On the basis of the outcome of this analysis, WG3 will implement the connected licence or rights statement compatibility table both at the "horizontal" level as already in draft version in the listed documents, as well as at the "multi-layer" level explained in the cited papers.

The extended abstract about TDM exception will be developed into a full paper.

The glossary will be accordingly expanded.

Many of the developed resources (case scenarios, bibliography, glossary, etc.), will form the basis for additional training material as requested by other WP (e.g. FAQs).

2.3.5 State of operations

The work of WG3 is on schedule. The discussion about "granularity" is revealing to be much more complex than originally thought, but this has not caused major delays. In the eventuality in which the discussion will not find an acceptable solution within reasonable time, a risk reduction plan has already been considered. The originally intended licence compatibility tools will be developed, in parallel to the discussion regarding rights statement. Given the modularity of the compatibility table and the multi-layer approach, an eventual implementation of a rights statement compatibility table within the existing licence compatibility can be easily achieved in legal, technical and scientific terms.

2.4 WG4 – Annotation and workflows

2.4.1 Mission statement

This working group studies interoperability aspects of text annotation and workflows. It includes supported input/output formats, annotation encoding models, workflow architectures, service access modes, type system alignment and others. An interface between workflow management systems and components is a key interoperability issue, as it includes the problems of how their functionality is packaged, what metadata are included and how they are interpreted by a system, but is also related to what type of information is processed and how it is represented, serialised as input/output files or transmitted.

2.4.2 Mode of operation

The group activities are based on the expertise of its members, representing institutions developing some of the leading text mining frameworks: University of Manchester (ARGO, U-Compare), University of Sheffield (GATE, AnnoMarket), University of Darmstadt (DKPro Core), French National Institute for Agricultural Research (Alvis) and Athena Research and Innovation Center (ILSP). The group meets at regular conference calls every two weeks and if necessary consults external experts, representing other major TM centres. A cycle of technical presentations on workflow systems has also been initiated, starting with a description and discussion on distributed execution in Argo.

2.4.3 Progress in current period

So far the group has produced the following resources:

- A set of **33 requirements for components**[1] to assure workflow interoperability. They have been created based on experiences of group members and analysis of interoperability scenarios.
- An **alignment of 6 type systems**[2] used in existing platforms (Alvis, Argo, DKPro Core, GATE, ILSP, LAPPS Grid). The alignment maps equivalent types and features, which shows concepts and approaches that are consistent or overlapping. On the other hand, it also help to identify differences and see whether they come from different focus (e.g. concentrating on biomedical concepts missing from other systems) or different conventions of data representation.
- An initial **directory of 556 components**[3] currently available in libraries of the considered workflow systems, including their short description, automatically assigned categories, parameters and machine-readable descriptors in META-SHARE format. The directory is created through an automatic process that aggregates metadata from multiple sources. The aggregation processes is work in progress and continually being improved.
- Initial work on a prototype solution allowing to build workflows including components coming from different platforms (initially DKPro Core (UIMA) and GATE, also looking into Alvis, Argo, ILSP, LAPPS Grid) in the form if the domain-specific programming language "OpenMinTeD Script".[4] This prototype serves as a sandbox to investigate interoperability issues in terms of

[1] https://openminted.github.io/openminted-site/releases/interop-spec/1.0.0/openminted-interoperability-spec.html#WG4
[2] https://openminted.github.io/openminted-site/releases/interop-spec/1.0.0/typealignment.html
[3] https://openminted.github.io/openminted-site/releases/interop-spec/1.0.0/components.html
[4] https://github.com/openminted/openminted-script

component lifecycle, deployment, and data transformation. In particular, it allows us to generate discussions and insights on these topics independently of the OpenMinTeD Workflow service which will be delivered later in the project. In fact, we expect that lessons learned from OpenMinTeD Script will have an impact on the design of the OpenMinTeD Workflow service – potentially parts of OpenMinTeD Script can even evolve to be integrated into the workflow service, e.g. the data transformation functionality.

- Publication: **Interoperability of corpus processing work-flow engines: the case of Alvis NLP/ML in OpenMinTeD** (for details on the publication and a link, see Section 2.5)

2.4.4 Tasks planned for next period

The following tasks are necessary to improve the requirements set:

- Finalising all requirements: some of the proposed requirements have sparked off a discussion that hasn't been concluded by unanimous agreement yet. These requirements have received 'draft' status and now need to be further discussed and finalised.
- Creating concrete requirements: so far all of the created requirements are 'abstract', i.e. they describe some desired functionality (e.g. components should be described by machine-readable metadata), but without technical details (e.g. a format of the metadata). For each abstract requirement, at least one concrete counterpart should be created in the next period. The process of creating the concrete requirements will also inform the interoperability guideline deliverables (D5.5 and D5.6).

2.4.5 State of operations

The group operates efficiently and produces expected outputs according to the plan.

2.5 Publications

This section lists peer-reviewed publications relevant to this deliverable from project partners within the reporting period. All the publications are available online[1] as open access under CC-BY-NC licence[2].

- P. Labropoulou and S. Piperidis and T. Margoni, 2016. **Legal Interoperability Issues in the Framework of the OpenMinTeD Project: a Methodological Overview**. In Proceedings of the Workshop on Cross-Platform Text Mining and Natural Language Processing Interoperability (INTEROP 2016) at LREC 2016, p.60-63, Portorož, Slovenia, DOI 10.5281/zenodo.182497
- T. Margoni and G. Dore, 2016. **Why We Need a Text and Data Mining Exception (but it is not enough) (Extended Abstract)**. In Proceedings of the Workshop on Cross-Platform Text Mining and Natural Language Processing Interoperability (INTEROP 2016) at LREC 2016, p.57-59, Portorož, Slovenia
- W. Peters, 2016. **Tackling Resource Interoperability: Principles, Strategies and Models**. In Proceedings of the Workshop on Cross-Platform Text Mining and Natural Language Processing Interoperability (INTEROP 2016) at LREC 2016, p.34-37, Portorož, Slovenia

[1] http://interop2016.github.io//Program and http://www.lrec-conf.org/proceedings/lrec2016/index.html
[2] http://lrec2016.lrec-conf.org/en/submission/authors-kit/

- M. Ba and R. Bossy, 2016. **Interoperability of corpus processing work-flow engines: the case of AlvisNLP/ML in OpenMinTeD**. In Proceedings of the Workshop on Cross-Platform Text Mining and Natural Language Processing Interoperability (INTEROP 2016) at LREC 2016, p.15-18, Portorož, Slovenia
- P. Knoth and N. Pontika, 2016. **Aggregating Research Papers from Publishers' Systems to Support Text and Data Mining: Deliberate Lack of Interoperability or Not?**. In Proceedings of the Workshop on Cross-Platform Text Mining and Natural Language Processing Interoperability (INTEROP 2016) at LREC 2016, p.1-4, Portorož, Slovenia, DOI 10.5281/zenodo.194788
- R. Eckart de Castilho, 2016. **Interoperability = f(community, division of labour)**. In Proceedings of the Workshop on Cross-Platform Text Mining and Natural Language Processing Interoperability (INTEROP 2016) at LREC 2016, p.24-28, Portorož, Slovenia, DOI 10.5281/zenodo.161848

2.6 External Experts

The following persons act as external experts on one or more of the interoperability working groups.

Name	Affiliation	WG1	WG2	WG3	WG3
Andreas Kempf	Deutsche Zentralbibiothek für Wirtschaftswissenschaften, Germany		X		
Christian Chiarcos	Goethe-Universität Frankfurt am Main, Germany	X			
Christopher Cieri	LDC, USA			X	
Daan Broeder	MPI for Psycholinguistics, Netherlands	X			
Diane Peters	Creative Commons HQ			X	
Dominique Estival	Western Sydney University, Australia		X		X
Enrique Alonso	Consejo de Estado			X	
Eric Nyberg	Carnegie Mellon University, USA				X
Federico Morando	Nexa Center for Internet & Society, Italia			X	
Geoffrey Bilder	Crossref	X		X	
Giulia Ajmone Marsan	The Organisation for Economic Co-operation and Development (OECD)			X	
Gwen Franck	Creative Commons, EIFL			X	
Ineke Schuurman	CCL, University of Leuven		X		
Jin-Dong Kim	Database Center for Life Science, Research Organisation of Information and Systems				X
Jochen Schirrwagen	Universität Bielefeld, Germany	X			
John McCrae	National University of Ireland, Galway, Ireland		X		
Keith Suderman	Vassar College,USA (LAPPS Grid)	X	X		X
Kristofer Erickson	CREATe			X	
Lars Bjørnshauge	SPARC Europe			X	
Liam Earney	JISC, UK			X	
Lukasz Bolikowski	University of Warsaw, Poland	X			X
Maarten van Gompel	Radboud University Nijmegen, NL		X		
Maarten Zeinstra	Kennisland, NL			X	
Marc Verhagen	Brandeis University, USA (LAPPS Grid)				X
Mark Perry	University of new England, Australia			X	
Maurizio Borghi	Bournemouth University, UK			X	
Menzo Windhouwer	MPI for Psycholinguistics, Netherlands		X		
Nancy Ide	Vassar College, USA (LAPPS Grid)		X		X
Paul Keller	Kennisland, NL			X	
Paul Uhlir	National Academy of Sciences			X	
Pawel Kamocki	Institut für Deutsche Sprache, Germany			X	
Peter Suber	Berkman Klein Centre, Harvard University			X	
Piek Vossen	VU University Amsterdam, Netherlands				X
Prodromos Tsiavos	The Media Institute			X	
Rafal Rak	UberResearch, UK				X
Steve Cassidy	Macquarie University Sydney, Australia		X		X
Thilo Götz	IBM, Germany				X

3. *Scenarios*

In preparation of generating interoperability requirements (Section 4), the WGs prepared a set of 17 scenarios. These scenarios highlighted particular aspects of interoperability from the perspective of the respective WG. They were identified and described by the participants from the respective working groups through introspection and subsequently described and refined in a collaborative process involving external experts, cross-WG communication, as well as communication with WP4. In particular, the first OpenMinTeD Interoperability Workshop held on Nov 12, 2015 in The Hague, NL revolved around the interoperability scenarios and focussed on deriving a first seed set of interoperability requirements from them which was later elaborated.

WG1

- Scenario 1 — Discover resources of various types at various locations
- Scenario 2 — SME running research analytics for funders within the European Research Area
- Scenario 3 –- A content provider using text mining tools to enrich their content
- Scenario 4 — Provide comprehensive statistical metadata for resources
- Scenario 5 — Domain specific researcher using a text mining tool or service to promote their research or use applied research results within their setting.

WG2

- Scenario 1 — Combining heterogeneous resources for information extraction
- Scenario 2 — Including Custom Knowledge
- Scenario 3 — The relation between documents and knowledge bases through keywords

WG3

- Scenario 1 — Legal status of aggregations: focus on content
- Scenario 2 — Focus on TDM tools and TDM services
- Scenario 3 — The type and nature of TDM results (or How far do copyright and SGDR reach)?

WG4

- Scenario 1 — Transferability of components between ecosystems
- Scenario 2 — Comparison of competing components or parameters
- Scenario 3 — Non-expert provider of TDM resource
- Scenario 4 — Reproducibility of TDM-related research
- Scenario 5 — Integration of a TDM workflow in a service/embedding in an application
- Scenario 6 — Development of TDM resources

A full description of the scenarios is omitted in this document. They are provided as an attachment as well as on our publicly GitHub repository.[1]

[1] https://openminted.github.io/openminted-site/releases/interop-spec/1.0.0/openminted-interoperability-scenarios.html

4. Requirements

This section outlines the structure of requirements and presents the requirements generated so far. The section provides an overview of the requirement's structure (Section 4.1) and a high-level overview of the actual requirements (Section 4.2).

4.1 Requirement Structure

ID - Every requirement has an ID. We start counting from 1 and every new requirement increments the ID by 1. The ID is encoded in the requirement filename, e.g. 1.adoc.

Concreteness - The OpenMinTeD infrastructure aims to be open, sustainable, and able to cope with change in the community and in technology. As such, it needs to be able to support multiple popular technologies and standards. As popularity is changing over time and as new standards and technologies are evolving, OpenMinTeD will have to evolve as well. As supporting too many technologies and standards in parallel is also unsustainable. Thus, the supported by OpenMinTeD at any time will be limited to a few. However, third-parties that would like to develop and maintain external modules for OpenMinTeD to support additional technologies and standards are welcome and these third-parties can refer to the interoperability requirements to estimate the feasibility of creating such an external module. The distinction between abstract and concrete interoperability requirements that we make here allows us to answer two questions:

- How difficult is it for a new technology or standard to be incorporated into OpenMinTeD?
- How difficult is it to integrate new components based on already supported technologies and standards into OpenMinTeD?

Abstract requirements are agnostic to concrete technologies and standards and help assessing compliance with them; helps answering the first question. Concrete requirements refer to specific implementation details and help answering the second question.
Requirement concreteness values

- **Abstract** - the requirement specifies a need, but does not go into details how this need must be fulfilled. The requirement may provide examples of techniques or implementations that fulfil the requirement, but does not mandate their use.
- **Concrete** - the requirement specifies a need and prescribes the use of specific techniques, standards, implementations, etc.

Strength - Requirement strength values

- **Mandatory** - compliance with a mandatory requirement is obligatory. Non-compliance with any mandatory requirement entails non-compliance with the specification as a whole.
- **Recommended** - compliance with a recommended requirement is not obligatory but strongly desired.
- **Optional** - compliance with an optional requirement is not obligatory and not strongly desired, but considered beneficial.

Status - The requirement status indicates how far it has proceeded in its lifecycle. If and which changes may be made to a requirement depends on this status.

Requirement status values

- **Draft** - the requirement is a suggestion and can be changed substantially in any respect.
- **Final** - the requirement is ready for release. Changes to a final requirement are only allowed if they do not affect the compliance status of any product, component, format, etc. that has already been evaluated against the requirement specification. If a change would trigger a change in any compliance status, instead of changing an existing requirement, a new requirement must be created under a new ID and compliance must be evaluated against this new requirement specification in the next iteration. The previous requirement must be moved to deprecated status.
- **Deprecated** - the requirement is no longer to be used for compliance assessment. The requirement specification must not be changed. Exceptions are amendments adding pointers to potential new versions of the requirement and providing a rationale for the deprecation.

Category - The category of a requirement is used to anchor it in the document structure of the interoperability specification. A requirement may be associated with multiple categories.

4.2 Requirements Overview

This section provides a high-level overview of the interoperability requirements that have been generated during the reporting period. A total of 72 requirements have been generated by the WGs, many of which are applicable across the WGs (WG1: 21, WG2: 17, WG3: 23, WG4: 33). These can be broken down by status:

- 22 requirements in status “draft” (Table 1)
- 40 requirements in status “final” (Table 2)
- 10 requirements in status “deprecated” (Table 3)

Here, we provide only a tabular overview over the requirements generated so far. Each of these requirements has a more detailed description which can be found in the interoperability specification document[1] that is also attached to the present deliverable.
The generation of requirements happens per WG. It is possible that very similar requirements are being generated in multiple WGs. When this happened, we kept on of the requirements and deprecated the others, merging compliance assessment into the remaining requirement if necessary – this is an ongoing process and continues as more requirements are added and as existing requirements become better understood. Several requirements that were generated by the WGs were later considered to be functional requirements for one of the OpenMinTeD services (e.g. the registry service or the workflow services) rather than interoperability requirements. These have also been marked as deprecated and scheduled for inclusion in the functional specification document D4.3.
Most of the requirements are recommendations (41), a core set of requirements is mandatory (6), and a few are optional (6).
We provide in this document only the requirement overview with their short titles. The full requirement specification is provided as an attachment to this document and is also publicly hosted on our GitHub

[1] https://openminted.github.io/openminted-site/releases/interop-spec/1.0.0/openminted-interoperability-spec.html

repository. Browsing the requirements hosted on GitHub is the preferred method as it is a highly cross-referenced hypertext.

Table 1 - Requirements in status "draft"

ID	Requirement	Concreteness	Strength	WG's
5	Components should detail all their environmental requirements for execution	abstract	mandatory	WG4
6	Components should have a unique identifier and a version number	abstract	mandatory	WG4
10	Components should specify the types of the annotations that they input and output	abstract	mandatory	WG4, WG2
11	Components should declare whether they can be scaled within a workflow	abstract	mandatory	WG4
13	Citation information for component	abstract	recommended	WG1, WG4
14	Components must maintain Licence information	abstract	mandatory	WG4
18	Workflows should be described using an uniform language	abstract	recommended	WG4
51	Licence should be attached	abstract	recommended	WG3
53	Licensor must be entitled to grant licence	abstract	recommended	WG3
54	Licensees should remain with a copy of the licence	abstract	recommended	WG3
55	Standard licences should be used	abstract	recommended	WG3
56	Licence should be machine readable	abstract	recommended	WG3
57	Licence should be understandable by non-lawyers	abstract	recommended	WG3
58	TDM must be explicitly allowed	abstract	recommended	WG3
59	Right for (temporary) reproduction must be granted	abstract	recommended	WG3
60	Boundary for derivative work must be clearly defined	abstract	recommended	WG3
61	No restrictions on TDM results which are not derived works	abstract	recommended	WG3
62	World-wide and irrevocable licence grant	abstract	recommended	WG3
63	Licence must qualify for Open Access rights	abstract	recommended	WG3
64	Licence must qualify for Open Access uses	abstract	recommended	WG3
65	Licence must qualify for Open Access must not restrict use in any way	abstract	recommended	WG3
66	Licence must qualify for Open Access may include attribution requirements	abstract	recommended	WG3

Table 2 - Requirements in status "final"

ID	Requirement	Concreteness	Strength	WG's
1	Components should be described by machine-readable metadata	abstract	mandatory	WG4
2	Component metadata should be embedded into the component source code	abstract	recommended	WG4
3	Component metadata is separable from the component	abstract	mandatory	WG4
4	URL to actual content must be discoverable	abstract	mandatory	WG1, WG2, WG3
7	Components should have a fully qualified name that follows the Java class naming conventions	concrete	mandatory	WG4
8	Components should associate themselves with categories defined by the OpenMinTeD project	abstract	mandatory	WG4
9	Components should declare their annotation schema dependencies	abstract	mandatory	WG4
12	Components should provide documentation describing their functionality	abstract	recommended	WG4
15	Human readable information should be provided by each resource	abstract	recommended	WG1, WG4
16	Models/resources should be useable across different component collections/platforms	abstract	recommended	WG4
17	Components should be stateless	concrete	recommended	WG4
21	Configuration and parametrisable options of the components should be identified and documented	abstract	recommended	WG4
24	Using/treating workflows as components	abstract	mandatory	WG4
26	Ability to determine source of an annotation/assigned category	abstract	recommended	WG4
27	Components should handle failures gracefully	abstract	recommended	WG4
28	Processing components should be downloadable	abstract	recommended	WG4
30	Metrics for the confidence level of the TDM operation should be included in the metadata	abstract	optional	WG1, WG4
31	Metrics for the performance of the TDM operation should be included in the metadata	abstract	optional	WG1, WG4
32	Version must be included in the metadata description for all resources	abstract	mandatory	WG1, WG2, WG3, WG4
33	Licensing information must be included in the metadata	abstract	mandatory	WG1, WG3

ID	Requirement	Concreteness	Strength	WG's
34	Licensing information should be expressed in a machine-readable form	abstract	recommended	WG1, WG3
35	All resources must include a unique persistent identifier	abstract	mandatory	WG1, WG2, WG3, WG4
36	Classification metadata should be included, where applicable, in the metadata record of the resource	abstract	recommended	WG1, WG2
37	Information on the structural annotation (layout) of resources should be included in the metadata of the resource	abstract	recommended	WG1
38	Access mode of resources must be included in the metadata	abstract	mandatory	WG1, WG2, WG4
39	Content resources must include metadata on their format (e.g. XML, DOCX etc.)	abstract	mandatory	WG1
40	Component metadata must include standardised categories/tags that make them easy to discover	abstract	mandatory	WG1, WG4
41	Content resources must include metadata on their language(s)	abstract	mandatory	WG1, WG2
43	S/W (tools, web services, workflows) must indicate whether they are language-independent or the language(s) of the resources they take as input and output	abstract	mandatory	WG1, WG4
44	Statistical metadata that allow monitoring of resource versions may accompany resources	abstract	optional	WG1, WG2
45	S/W (tools, web services, workflows) must indicate format of their output	abstract	mandatory	WG1, WG4
47	Information on funding of resources may be included in the metadata	abstract	optional	WG1, WG2, WG3, WG4
48	All resource metadata records must include a reference to the metadata schema used for their description	abstract	mandatory	WG1, WG2, WG3, WG4
50	Documentation references should be versioned	abstract	recommended	WG1, WG2, WG3, WG4
67	Knowledge Resource Element Id	abstract	recommended	WG2
68	Data Category Linking Vocabulary	abstract	recommended	WG2
69	Interoperability between elements from different knowledge resource schemas should be expressed through RDF statements.	abstract	recommended	WG2
70	All KR content elements need to be added as text annotations within a TDM workflow.	abstract	mandatory	WG2

ID	Requirement	Concreteness	Strength	WG's
71	The KR should be ingestible through a URI	abstract	recommended	WG2
72	The KR format should be in a standard format such as XML, JSON or RDF.	abstract	recommended	WG2

Table 3 – Requirements in status "deprecated"

ID	Requirement	Concreteness	Strength	WG's
19	Components that use external knowledge resources should delegate access to a resource adapter instead of handling it themselves	abstract	optional	WG2, WG4
20	Workflow engines should not require to see data	concrete	recommended	WG2, WG4
22	The Workflow Engine Should Permit Saving Experimental Conditions in a Workflow	abstract	recommended	WG1, WG4
23	The Workflow Engine should permit Licence Aggregation in Workflows	abstract	recommended	WG3, WG4
25	Incorporation of multiple resources in parallel	abstract	recommended	WG4
29	The actual content of all content resources must be discoverable	abstract	mandatory	WG1, WG2, WG3
42	The metadata can include the information on which projects/workflows involve the resource	abstract	optional	WG1, WG2, WG3, WG4
46	Output resources of web services/workflows must be accompanied by provenance metadata	abstract	mandatory	WG1, WG4
49	Metadata of tools should contain information about the models available for them	abstract	recommended	WG1, WG4
52	Licence information must be in metadata	abstract	recommended	WG1, WG3

5. 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 this section, we analyse the compliance with the requirements so far. This provides us with a basis for determining how to effectively improve compliance and thus interoperability between the relevant products as well as with the OpenMinTeD infrastructure.

5.1 Compliance levels

As part of the compliance assessment process, the following compliance levels are assigned:

- **Full** - fully compliant
- **Partial** - partially compliant. E.g. some parts of a product are compliant but not all. This is typically the case if a product is in a state of transition from a non-compliant to a compliant state.
- **No** - not compliant.
- **N/A** - not applicable. This is expected to occur mainly for concrete requirements if a certain requirement is not applicable for a certain implementation, e.g. a requirement on remote API access on a tool which does not offer a remote API. Abstract requirements should be formulated in such a way that they are always applicable.

When a requirement is changed, compliance assessments may have to be updated as well. Thus, compliance assessments should only be made on requirements that have been marked as “final”, i.e. whose description must no longer be changed. However, in preparation of the present deliverable, we have also performed compliance assessments for those requirements which are still in “draft” status. Those assessments will have to be updated when the requirements are promoted to the “final” status.

5.2 Compliance assessments

In this section, we list the products taken into account for the compliance assessment. For every interoperability requirement, there are relevant classes of products:

- Resources that have been developed by the consortium partners and where the creation of metadata is the responsibility of the respective partners (Frontiers, Alvis, Argo/U-Compare, DKPro Core, ILSP)
- Resources that are already used in TDM processes and/or are being examined for use in OpenMinTeD and are, therefore, not directly responsible for the metadata descriptions (TheSOZ, AGROVOC, JATS, OLIA, LAPPS Grid, licences)
- Resources that are being collected from the original data providers who also supply the metadata descriptions (CORE, OpenAIRE).

An overview of the assessed products can be found in Table 4. The detailed assessment can be found in the Detailed Interoperability Specification v1[1].

[1] https://openminted.github.io/openminted-site/releases/interop-spec/1.0.0/openminted-interoperability-spec.html#_compliance

Table 4 - Assessed products and consulted sources

Product	Assessed requirements (excl. deprecated)	Source
ARGO	35	http://argo.nactem.ac.uk/
AGROVOC	16	http://aims.fao.org/aos/agrovoc/void.ttl
Alvis	36	http://www.quaero.org/module_technologique/alvis-nlp-alvis-natural-language-processing/
CLARIN CR	5	https://www.clarin.eu/ccr
CORE	11	https://core.ac.uk
DKPro Core	36	https://dkpro.github.io/dkpro-core/documentation/
Frontiers	11	http://home.frontiersin.org/about/author-guidelines
GATE	35	https://gate.ac.uk/sale/tao/split.html
ILSP	13	https://inventory.clarin.gr/
JATS	15	http://jats.nlm.nih.gov/about.html
LAPPS Grid	15	http://vocab.lappsgrid.org/
Licences	6	variety of standard licences, such as CC and FOSS
OLiA	15	http://acoli.cs.uni-frankfurt.de/resources/olia/
Ontolex	5	https://www.w3.org/community/ontolex/
OpenAIRE	11	https://guidelines.openaire.eu/en/latest/
TheSOZ	16	http://lod.gesis.org/thesoz/de.html
Apache UIMA	1	https://uima.apache.org
schema.org	5	http://schema.org

6. *Actions*

Based on the compliance assessment, each WG has identified actions that need to be performed in order to improve the compliance of relevant products with the OpenMinTeD interoperability requirements. These actions shall guide the work of the WGs in the next reporting period(s), will provide input to T5.4 "Alignment of service and content provider systems" and T5.5 "Data interoperability toolkit for repositories, publishers' systems and OpenMinTeD subsystems" and shall also be taken into account for the implementation of OpenMinTeD services (WP 6).
Most of the requirements generated so far (69 out of 72) are "abstract", i.e. endorsed ways to be compliant with these requirements through the use of specific standards, have not yet been specified. Nevertheless, various relevant products are already compliant with these abstract requirements, although potentially in very different ways.
A major focus across all WGs for the next reporting period will be to add suitable "concrete" requirements explicating the specific standards and mechanisms endorsed and supported by OpenMinTeD. Where no suitable standards and mechanisms exist, the WGs will - in collaboration with WP 6 (Implementation) - propose to make use of respective mechanisms pioneered and implemented by OpenMinTeD and include their respective specifications in future versions of this deliverable.
A second measure of ensuring the applicability, practicability, and completeness of the interoperability requirements going forward is the continued development of interoperability prototypes. These prototypes are also meant to be carried over into the actual implementation of OpenMinTeD.

6.1 WG 1

With one exception, when the metadata descriptions fall under the responsibility of the consortium members, the results of the assessments were rather satisfactory, when the requirement applied to the specific resource type.
Strategic actions – need to be undertaken to resolve these issues include at a higher level:

- Promoting and supporting the creation and enrichment of formal metadata descriptions
- Standardising, where possible, the metadata elements and values and recommending best practices for filling them in.

Immediate actions – the immediate actions that can and should be taken in the OpenMinTeD framework, to ensure interoperability at least for the project's purposes:

- Create formal metadata descriptions for all the resources; for those that are not developed by the consortium partners, this will be allocated to relevant qualified members
- Conversion of the existing metadata descriptions to the reference metadata schema and enrichment thereof with the lacking metadata elements; this will also help in spotting other interoperability issues as well.

Adding/improving formal metadata descriptions for JATS, LAPPS and Alvis; for Alvis, these can be provided in the next phase, given that the developer is part of the consortium; for JATS and LAPPS, these will need to be provided by other partners.

- Lack of appropriate metadata elements in the used metadata schemas to encode the required information (e.g. REQ-4[1] and REQ-33 for access point and licensing respectively).
- Lack of standardised vocabulary to encode the information, even though the elements are considered important and may already be present in other forms (e.g. as free text documentation); for instance, for REQ-30, REQ-31 (quality metrics) and REQ-32 (version), it is important to arrive at a consensus on the encoding practices before adding it to the metadata description in a harmonised way.
- Absence of the information in the metadata descriptions, despite the existence of the appropriate elements; such cases are, for instance, REQ-39 (format for content resources) and REQ-41 (language for content resources); this category includes both technical and administrative information, and the reasons behind this non-compliance can be that the information is usually optional and regarded by the providers as redundant.

Prototypes – the component overview[2] represents a prototype for the aggregation and transformation of existing component metadata descriptions from different sources (GATE, UIMA, Alvis, Maven, etc.) into a common scheme. Its development has provided insights that have been integrated into the development of the first version of the OpenMinTeD Metadata Schema. Parts of its functionality, in particular functionality related to the harvesting of metadata, is also now being transferred into the OpenMinTeD registry. WG1 will accompany the evolution of this prototype as it is being integrated into the registry and as its harvesting functionalities are expanded and update the interoperability specification as new requirements come up.

Table 5 – WG 1 summary of actions to improve compliance

Action	Products	Related requirements
Create formal metadata descriptions	Alvis, JATS, LAPPS	All WG1 requirements
Enrich metadata descriptions with specific elements (added in the reference metadata schema), at least regarding required elements	Alvis, Argo/U-Compare, DKPro Core, ILSP	4, 33, 38, 40, 44, 47, 48
Discuss in order to standardise vocabulary and add to metadata descriptions	all	30, 31, 32, 35, 36, 44
Promote the enrichment of metadata descriptions, especially for required metadata elements, in the case of resources owned by others	OpenAIRE, CORE, original providers of depending resources	4, 33, 37, 39, 41, 47, 48
Evolve metadata harvesting and aggregation prototype	all	All WG1 requirements

[1] https://openminted.github.io/openminted-site/releases/interop-spec/1.0.0/openminted-interoperability-spec.html#REQ-4
[2] https://openminted.github.io/openminted-site/releases/interop-spec/1.0.0/components.html

6.2 WG 2

In general, the requirements were deemed essential and relevant. All KRs were deemed compliant with REQ-67, REQ-71 and REQ-72. For this reason, we consider making these requirements mandatory in the next iteration of the specification.
Requirement REQ-68 and REQ-69 check the level at which KRs already resource-internally make use of:

- links to elements from external vocabularies
- elements from OWL/RDF/SKOS linking vocabularies

The set of (de facto) standard reference vocabularies to be used for establishing data category interoperability is the following:

- Schemas associated with OpenMinTeD
  - GATE
  - LAPPS Exchange Vocabulary Type Hierarchy
  - DKPro Core
- Strategical, i.e. widely used and interconnected (de facto) standard vocabularies for linguistic/terminological/ontological metadata
  - Ontolex
  - OLIA Reference Model
  - CLARIN Concept Registry
  - Schema.org
  - Penn Treebank
  - Universal Dependency
- Representative set of use case driven schemas
  - TheSOZ (social sciences)
  - JATS (structure of scholarly articles)
  - Agrovoc (agriculture)
  - BioLexicon (life sciences)

This set will be extended where necessary in order to ensure full coverage for the interoperability of TDM KR elements required in OpenMinTeD.
REQ-70 was considered as potentially a general WG4 requirement, and was therefore not included in the compliance check.

Using the schema alignments created in this period, a prototype will be set up during the next reporting period to investigate the application of the alignments in practical workflows. To this end, we will involve in particular those OpenMinTeD partners and external experts that work extensively with knowledge resources. In this way, we expect to generate further abstract and concrete interoperability requirements for OpenMinTeD. This prototype can be implemented in conjunction with another prototype originating from WG4: OpenMinTeD Script.

Table 6 – WG 2 summary of actions to improve compliance

Action	Product	Related requirements
Implementation of prototype integrating schema alignments with actual workflows	all	All WG2 requirements
Possible extension of WG2 requirements with requirements produced by WP4.	all	
Check overlap with/migration to requirements from WG1/3/4	all	70
Further extension of reference vocabularies set through inclusion of additional standards.	all	69
Definition of initial linking structure in the form of a spreadsheet: class name/feature/feature value/SKOS linking relation/class name/feature/feature value	all	69
Further definition of linking strategies. • Links between simple RDF classes with SKOS relations • Explore the viability of the definition in RDF of complex data categories (e.g. classes with specific feature values) as named graphs. • Link these named graphs to simple classes or other named graphs with SKOS relations.	all	69
Further linking of reference vocabulary elements.	all	69
Selection of relevant/necessary GATE plugins and their annotation types; create GATE schema.	all	67
Creation of links between GATE data categories and reference vocabularies.	all	69
Creation of RDF serialized network of (de facto) standard KR elements and links. This network will be used for interoperability purposes as the mediating KR for text annotation type harmonisation.	all	All WG2 requirements

6.3 WG 3

Due to the nature of the WG, the possible immediate actions are limited. The primary relevant products for this WG are licences and terms of use that are created by third parties. Hence, actions with external influence are generally of a strategic nature. Other actions are coordinated with other WGs, mainly WG1, regarding the inclusion of licence metadata with resources being submitted to and accessed through OpenMinTeD.

Strategic actions – All resources ingested by OpenMinTeD or produced as the result of a TDM process must carry a licence. The licence has to be expressed in both legal and metadata terms. An additional layer of information regarding the main rights and obligations should also be added (e.g. commons deed).

Licences should comply with the product requirements (licences) identified by WG3. In particular, all licences (inbound; outbound) should be chosen among standard licences with clear compatibility standards. Ad-Hoc licences are deprecated. A major issue here is connected with the fact that typically terms of use for services are not standardised. This is an aspect that will be addressed.
An assessment of the legal status of a resource (other than licence or absence of licences) is dependent on applicable legislation. Ad-hoc analysis in this case seems unavoidable.
Immediate actions – We continue discussion with internal and external experts the "rights statement" implementation and feed the output of this discussion into the implementation of the connected licence or rights statement compatibility table.
Prototypes – For the next reporting period, we plan the implementation of a licence selector prototype turning the compatibility matrix into a user-oriented application. Additionally, we shall investigate the way in which certain permissions or obligations are (not) granted or imposed in particular licence texts through. To this end, we will conduct an experiment in which experts with legal training will analyse licence texts and annotate phrases or sentences in the licence texts with their legal implications.

Table 7 – WG 3 summary of actions to improve compliance

Action	Product	Related requirements
Apply licence to your resources	all	33, 51, 53, 54
Choose only resources with a licence	all	33
Apply licence properly (legal; metadata)	all	33, 56, 62
For resources with no clear licence statement look at applicable law (TDM exception?)	all	58, 59, 60, 61

6.4 WG 4

Immediate actions – Based on the compliance assessment, we identify three areas that require immediate action in the next reporting period:

1. **Core requirements** – necessary for workflow execution, e.g. regarding component input/output definition, metadata, dependencies specification. These are fairly well supported across the products and only a few improvements are necessary to achieve full compliance:
   - Making component metadata available both from its source (REQ-2) and separately (REQ-3).
   - Creating a unique identifier for each component (REQ-6).
   - Enabling using workflows as components (REQ-24).
   - Enforcing specifying input/output annotation types for components (REQ-10).
2. **Additional requirements** – where all (or nearly all) of the products are non-compliant, thus significant changes are necessary to satisfy them:
   - Non-technical information in component metadata, including citable publications (REQ-13), component category (REQ-8), associated licences (REQ-14) and licence aggregation for a whole workflow (REQ-23).

- Handling external resources used in a workflow: determining a source of a specific annotation element (REQ-26) or ensuring re-usability of resources across different platforms (REQ-16).
- Letting users decide on how a workflow is deployed, which may be necessary because of legal reasons by making sure a workflow engine doesn't see the processed data (REQ-20) or downloading components for local use (REQ-28).
- Writing documentation for components that lack it (REQ-12).

3. **Vague requirements** – where additional work is necessary to clarify the formulation, so that it will become evident what actions are necessary:
   - Components declaring their annotation schema (REQ-9) and environmental requirements (REQ-5).
   - Statelessness of components (REQ-17).
   - Uniform workflow description language (REQ-18).

Prototypes – During the present reporting period, WG4 has created OpenMinTeD Script as a prototype to investigate interoperability issues in cross-platform workflows (i.e. workflows involving components from UIMA and from GATE). This was a necessary step in order to deepen the discussion around interoperability in workflows towards the generation of concrete interoperability requirements. It also can serve as a temporary research substitute for the OpenMinTeD workflow service, which is to be delivered later in the project. In fact, we expect that parts of OpenMinTeD Script can be transferred into the design and implementation of the OpenMinTeD workflow service. As such, we shall continue evolving this prototype during the next reporting period, in particular incorporating more platforms (e.g. web-services from ILSP, from UNIMAN, or from LAPPS Grid). This also entails intensified investigation into the data transformation processes necessary to bridge the technical and semantic gaps between the different platforms.

Table 8 – WG 4 summary of actions to improve compliance

Action	Product	Related requirements
Continue evolving the OpenMinTeD Script prototype with a focus on the integration of additional platforms and on data transformation	all	All WG4 requirements
Improve existing component metadata (UIMA XML) model to make it available both from its source and separately.	Alvis, Argo	2, 3
Agree on a format of id and version for components and apply it.	all	6
Add functions to export all the configuration parameters of a workflow as a single file (considered difficult because of architectural limitations).	Alvis, Argo, ILSP	22
Add changes to execution environment and user interface that would enable running workflows as components	Argo	24

Action	Product	Related requirements
Revise implementations and metadata for existing components to make sure they specify input/output types	Alvis, Argo	10
Expand the component metadata schemata to include all desired additional fields. That seems to be fairly easily achievable in all products, but it requires a lot of effort to fill in that information (esp. licences) for all existing components.	all	8, 13, 14
Unify handling of resources and models.	all	16, 26
Prepare the execution model in a way that guarantees that a user may choose where the processing happens, which is important because of legal restrictions. Achieving this is considered possible but difficult, as it requires major changes in the systems.	Alvis, Argo, GATE, ILSP	28
Write documentation for undocumented components.	Alvis, Argo, ILSP	12
Offer classes for component authors to extend, so that they will handle failures properly	Argo	27
Define exactly what kind of type system and environmental information is necessary, how it's going to be used, and how to encode it.	Alvis, Argo, DKPro Core	5, 9
Define (or choose) a workflow representation language to be used before we can assess ability to comply with this	all	18

7. List of attachments

- Detailed Interoperability Specification v1
  - https://openminted.github.io/openminted-site/releases/interop-spec/1.0.0/openminted-interoperability-spec.html
- Detailed Interoperability Scenarios v1
  - https://openminted.github.io/openminted-site/releases/interop-spec/1.0.0/openminted-interoperability-scenarios.html
- Detailed type system alignment v1
  - Can presently not be included as PDF because of technical reasons
  - https://openminted.github.io/openminted-site/releases/interop-spec/1.0.0/typealignment.html
- Detailed overview of components from partners involved in WG4 v1
  - Can presently not be included as PDF because of technical reasons
  - https://openminted.github.io/openminted-site/releases/interop-spec/1.0.0/components.html
- OpenMinTeD Metadata Scheme
  - https://openminted.github.io/openminted-site/releases/omtd-share/1.0.0/html/index.html

8. *Appendix*

8.1 OpenMinTeD Component Classification (Draft)

This section provides an overview of the draft categorisation system for components. This is an excerpt from a work-in-progress document. The actual document also includes information on how to map these categories to other categorisation systems, e.g. the META-SHARE vocabulary.

Category Level 1	Category Level 2	Category Level 3	Category Level 4	Description (not formal definition!)
Access Component				
	Reader			A component that reads content of various types (pdf, txt, xml etc.)
	Writer			A component that writes processing results in various formats
Support Component				A component that provides support to developers
	Visualiser			A component or interface that renders the contents of a resource in a graphic way for visualisation purposes
	Debugger			A component that helps in the debugging process
	Validator			A component used to confirm that a system/resource meets the specifications and fulfills its intended purpose
	Viewer			A component that provides access to the contents of a resource but intended only for access by humans
		Corpus Viewer		A component that provides access to the contents of a corpus but intended only for access by humans
		Lexicon Viewer		A component that provides access to the contents of a lexical/conceptual resoruces but intended only for access by humans
	Editor			A component that allows humans to edit the contents of a resource
	ML Trainer			A component that is used in training models for machine learning
	ML Predictor			A component that is used in predicting based on machine learning models
	Feature Extractor			A component that is used for extracting features
	Data Splitter			A component that performs data splitting for cross validation purposes

	Data Merger		A component that supports data merging from various sources
	Converter		A component that performs conversion between formats of a resource
	Evaluator		A component that is used in the evaluation of the performance of a component
	Flow Controller		A component that supports controlling flows
	Script-Based Analyzer		A component that performs analysis tasks based on a script
	Matcher		A component that allows matching (mapping) of elements
		Gazetteer-based Matcher	A component that allows matching of elements based on a gazeteer
	Crowd Sourcing Component		A component that supports crowd sourcing operations
	Data Collector		A component that collects (retrieves) data from various sources
		Crawler	A component that crawls the web and collects data from various web sites
Processor			A component that is used in processing operations
	Annotator		A component that annotates any language data (text, video, audio etc.), i.e. adds any descriptive or analytic notations (structural, linguistic, etc) to raw language data
		Segmenter	A component that segments a text into structural untis (chapters, paragraphs, sentences, words, tokens etc.)
		Stemmer	A component that extracts stems from words in a text, usually by removing the commoner morphological and inflectional endings from words
		Lemmatizer	A component that annotates the tokens of a text with lemma information
		Morphological Tagger	A component that annotates tokens of a text with morphological information (part-of-speech and morphological features)
		Chunker	A component that groups tokens of a text into chunks
		Parser	A component that parses sentences and ?
		Coreference Annotator	A component that annotates tokens of a text with coreference information, i.e. annotating

				expressions that refer to the same entity in the text
		Named Entity Recognizer		A component that seeks to locate and classify elements in a text into pre-defined categories such as the names of persons, organisations, locations, expressions of times, etc.
		Semantics Annotator		A component that annotates the tokens of a text with semantic features
			SRL Annotator	A component that annotates the tokens of a text with Semantic Role labels
			Readability Annotator	A component that annotates the tokens of a text with readability scores
		Aligner		A component that detects and annotates equivalence relations between items (corpora, texts, paragraphs, sentences, phrases, words) in two languages
	Generator			A component that generates (semi-)automatically natural language texts (based on non-linguistic data, keywords, logical forms, knowledge bases...)
		Summarizer		A component that produces a natural language synopsis of a longer text
		Simplifier		A component that outputs a simpler rendition of a given item (sentence, text etc.)
	preOrPostProcessingComponent			A component that is used at pre- or post-processing stages in order to normalize input/output
		Spelling Checker		A component that corrects spelling mistakes in a text
		Grammar Checker		A component that corrects grammatical mistakes in a text
		Normalizer		A component that removes unwanted material from text, usually as a pre-processing step
		Filters		
	Analyzer			A component that is used for analysing an input text in order to perform extraction of features/information (e.g. word list), or characterisation of the whole text
		Topic Extractor		A component that guesses the topic of a text
		Document Classifier		A component that tries to classify a document into one or more categories
		Language Identifier		A component that identifies the language of a given text based on its contents

		Sentiment Analyzer	A component that tries to identify sentences that express the author's negative or positive feelings on something; (Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials (wikipedia))
		Keywords Extractor	A component that tries to extract keywords from a given text
		Term Extractor	A component that tries to extract terms from a corpus
		Contradiction Detector	A component that tries to automatically recognise elements that reveal contradiction in a text
		Event Extractor	A component that tries to extract information related to incidents referred to in a text
		Persuasive Expression Miner	A component that tries to identify persuasive expressions in a given text
		Information Extractor	A component that automatically extracts structured information from unstructured and/or semi-structured machine-readable documents
		Lexicon Extractor From Corpora	A component that extracts structured lexical resources from corpora
		Lexicon Extractor From Lexica	A component that extracts specific lexical information contained in other lexica
		Word Sense Disambiguator	A component that tries to identify which sense of a word (i.e. meaning) is used in a sentence, when the word has multiple meanings (Source: wikipedia)
		Qualitative Analyzer	
platform			A technology that eases the development of new tools and services in the NLP field
infrastructure			
architecture			A technology that supports the flexible development of NLP applications, together with all the requested resources
NLP development environment			A technology that supports the development of data resources, like lexicons, grammars, corpora, etc. Can be included in an Architecture or in a Platform
other			

8.2 WG1 - Inventory of metadata schemas and related efforts

Title	Full title	Type	Description	Publications	Lexica, ontologies, etc.	Corpora	S/w tools	Web services	Workflows	Comment
UIMA Component descriptors		TDM components	Describe language processing components and their parameters	no	no	no	yes	?	yes (aggregate components)	already used by partners
bibo[1]		publications	ontology for bibliographic citations	yes	yes	no	no	no	no	popular but similar to other bibliographic resources
DC / DCMI[2]	Dublin Core Metadata Initiative	general	metadata schema for digital resources	yes	yes	yes	yes	no	no	the most widespread at least for exchange purposes
ALVEO[3]	A Virtual Lab for Human Communication Science	language resources	infrastructure for finding, accessing and processing datasets for NLP	no	no	yes	no	no	yes	metadata mostly for linguistics; interesting mainly for galaxy
OLAC[4]	Open Language Archives	language resources	repository & metadata schema for language resources	no	yes	yes	no	no	no	metadata very general, based on DC; mainly for OAI-PMH
TEI[5]	Text Encoding Initiative	language resources	metadata schema for encoding external & internal information of text resources	yes	yes	yes	no	no	no	metadata mainly for humanities; more interesting for external structure
CERIF[6]	Common European Research Information Framework	projects, people, publications	schema for research entities, covering projects, funding, researchers, research organisations etc.	yes	yes	yes	no	no	no	used already by OpenAIRE; mainly for research entities; might be interesting for satellite entities
CrossRef[7]		publications	provider of DOI's for citation, linking & access; mainly for publications	yes	no	no	no	no	no	widespread for publications

[1] http://bibliontology.com/
[2] http://dublincore.org/documents/dcmi-terms/
[3] http://alveo.edu.au/
[4] http://www.language-archives.org/OLAC/metadata.html
[5] http://www.tei-c.org/Guidelines/P5/
[6] http://eurocris.org/cerif/main-features-cerif
[7] http://www.crossref.org/ & http://doi.crossref.org/schemas/unixref1.0.xsd

Title	Full title	Type	Description	Publications	Lexica, ontologies, etc.	Corpora	S/w tools	Web services	Workflows	Comment
JATS[1]	Journal Article Tag Suite	publications	XML tags for journal articles; schema for contents of publications	yes	no	no	no	no	no	mainly for structure & article types
OpenAIRE[2]	OpenAIRE	publications	aggregator; schema for publications & research data; guidelines for literature providers, data archives & CRIS managers	yes	no	no	no	no	no	for open access publications; used by partner
Beta SHARE Metadata Schema[3]		general	metadata schema for research objects	yes	no	yes	no	no	no	general for exchange
Maven Project Object Model[4]		software and resource artifacts	Describes software libraries and other software artifacts (can also be resource packages) in the Java world.	no	no	no	yes	no	no	used by partners
MARC21[5]	MARC21 format for bibliographic data	publications	MARC21 provides a complete but complex description of bibliographic metadata using code numbers to describe data; for different types of printed materials and digital media	yes	no	no	no	no	no	MARC21 provides a complete but complex description of bibliographic metadata using code numbers to describe data. MARC21 presents some inconveniences, such as its high complexity and its inability to be easily read by humans.
FaBiO[6]	FRBR (functional requirements for bibliographic records)-aligned Bibliographic Ontology	publications	an ontology for recording and publishing on the Semantic Web descriptions of entities that are published or potentially publishable, and that contain or are referred to by bibliographic references,	yes	no	no	no	no	no	FaBiO allows for the semantic description of a variety of bibliographic objects, such as research articles, journal articles, and journal volumes, to clearly separate each part of the publishing process, the people

[1] http://jats4r.org
[2] https://guidelines.openaire.eu/en/latest/
[3] https://osf.io/wur56/wiki/Schema/
[4] https://maven.apache.org/pom.html
[5] http://www.loc.gov/marc/bibliographic/
[6] http://www.sparontologies.net/ontologies/fabio/source.html

Title	Full title	Type	Description	Publications	Lexica, ontologies, etc.	Corpora	S/w tools	Web services	Workflows	Comment
			or entities used to define such bibliographic references.							involved in the publication process, and the various versions of documents (electronic or physical); may check for next version
EDAMontology[1]	EMBRACE Data and Methods	bioinformatics	ontology of well established, familiar concepts that are prevalent within bioinformatics, including types of data and data identifiers, data formats, operations and topics	yes	no	no	no	yes	no	seems to cover datasets, web services & publications; domain-specific; check for use cases later
CMDI[2]	Component Metadata Initiative	language resources	metadata schema modeller for language resources & registry for metadata components & profiles	no	yes	yes	yes	yes	no	wide range of profiles; check specific profiles later on
MetaShare[3]	MetaShare	language resources	repository & metadata schema for language resources	no	yes	yes	yes	yes	no	widespread for language resources
CORE[4]	COnnecting REpositories	publications	aggregates scholarly publications (metadata and full-text content) that is available as Open Access; it looks like the metadata comes from OLAC/DC (given the OAI/PMH protocol)	yes	no	no	no	yes	no	for publications; used by OU
swso[5]	Semantic Web Services Ontology	web services	ontology for web services	no	no	no	no	yes	yes	from Semantic Web; to check more thoroughly together with workflows for next version
DCAT[6]	Data Catalogue	datasets	schema for catalogues and datasets	no	no	yes	no	no	no	general for publishing catalogues

[1] http://edamontology.org/page
[2] http://media.dwds.de/clarin/userguide/text/metadata_CMDI.xhtml
[3] http://www.meta-share.org/portal/knowledgebase/home
[4] http://core.ac.uk
[5] http://www.w3.org/Submission/SWSF-SWSO/
[6] http://www.w3.org/TR/vocab-dcat/

Title	Full title	Type	Description	Publications	Lexica, ontologies, etc.	Corpora	S/w tools	Web services	Workflows	Comment
CC-REL[1]	Creative Commons - REL	licensing	ontology for legal metadata	N/A	N/A	N/A	N/A	N/A	N/A	complementary to legal metadata, if WG3 decides to go for machine readable licences
ODRL[2]	Open Digital Rights Ontology	licensing	ontology for representing legal rights	N/A	N/A	N/A	N/A	N/A	N/A	complementary to legal metadata, if WG3 decides to go for machine readable licences
LRE Map[3]	LRE Map	language resources	user-provided descriptions of language resources	no	yes	yes	yes	no	no	general, free values; user filled in
CCR[4]	CLARIN Concept Registry	metadata (external & linguistic)	registry for metadata elements and values	no	yes	yes	yes	yes	yes	follow-up of ISOcat; good for checking, but not all elements and values are validated; difficult to select those that are needed fro external metadata
Prov-O[5]		provenance	ontology for provenance information	N/A	N/A	N/A	N/A	N/A	N/A	for provenance information regardless of resource type; check for next version
DataCite[6]		publications	citation of datasets; DOI's; collaboration with CrossRef	yes	yes	yes	yes	no	no	focusing on citation rather than description
DOI[7]	Digital Object Identifier	publications	provider of PID's	N/A	N/A	N/A	N/A	N/A	N/A	
RIOXX[8]		publications	metadata application profile & guidelines for research publications, incl. research grants etc.; mapping to OpenAIRE	yes	no	no	no	no	no	

1 https://wiki.creativecommons.org/wiki/CC_REL
2 http://www.w3.org/ns/odrl/2/
3 http://www.resourcebook.eu/searchll.php
4 http://www.clarin.eu/ccr/
5 http://www.w3.org/TR/prov-o/
6 https://www.datacite.org/
7 http://www.doi.org/
8 http://rioxx.net/guidelines/

Title	Full title	Type	Description	Publications	Lexica, ontologies, etc.	Corpora	S/w tools	Web services	Workflows	Comment
swrc[1]		publications	ontology for modelling entities of research communities such as persons, organisations, publications (bibliographic metadata) and their relationships	yes	no	no	no	no	no	for satellite entities mainly
DOAJ[2]		publications	DOAJ (Directory of Open Access Journal) article format	yes	no	no	no	no	no	DC-based;
EDM[3]	EUROPEANA Data Model	cultural heritage objects		N/A	N/A	N/A	N/A	N/A	N/A	not so interesting for our scope; keep in mind only for the licensing model
NLM[4]	NLM (National Library of Medicine) Journal Archiving and Interchange Tag suite	publications		yes	no	no	no	no	no	obsolete; continues as JATS;
NISO[5]	National Information Standards Organisation	standards	standards for content publishers, libraries & s/w publishers	yes	no	no	no	no	no	general link to standards
Handle PID	PID	PID's	provider of PID's	N/A	N/A	N/A	N/A	N/A	N/A	
LAPPS Grid[6]	Language Application Grid	web services	open, interoperable web service platform for natural language processing (NLP) research and development	no	no	yes	yes	yes	yes	for components and type systems; checked by WG4
CREOLE descriptors[7]		TDM components		no	no	no	yes	?	yes	already used by partners

[1] http://ontoware.org/swrc/
[2] https://doaj.org/features
[3] http://pro.europeana.eu/page/edm-documentation
[4] http://dtd.nlm.nih.gov/
[5] http://www.niso.org/standards/
[6] http://www.lappsgrid.org/
[7] https://gate.ac.uk/sale/tao/splitch4.html

8.3 WG3 – Compatibility Matrix: Summary

8.3.1 Preliminary considerations

The terms of copyright licences are often unclear and not standardized, with the consequence that an effective and interoperable use of resources through TDM is extremely limited. Among the objectives of OpenMinTeD, licence standardisation and interoperability is of crucial importance. The aim of the WG3 is to provide guidance for users who wish to undertake TDM activities by developing a tool that would help overcoming the ambiguity of the current setting.
The main idea of the WG3 is to draft a **Compatibility Matrix** that considers

a) the type of content,
b) the type of licence, resulting
c) in whether there is compatibility or not among different licences.

This matrix should help users to share and distribute their resources under the appropriate licence and, at the same time, to comprehend the legal implications of choosing one or the other licence.

8.3.2 License Compatibility Tools

In order to develop the OpenMinTeD Compatibility Matrix, WG3 has collected existing related work. The first step was to collect a list of examples of how to provide or represent graphically information regarding licence conditions and, for some, compatibility. For each example, main features and limits have been indicated.
Exemplary tools can be grouped into three main categories:

1. licence calculators or selectors;
2. licence descriptors;
3. comparative tables and graphics.

The first category includes tools that use slightly different sets of criteria (licensing conditions) for the classification of licences and end up with a number of licences that satisfy these criteria. The second offers a representation of licences with visualisation of basic licensing conditions. The third provides a figurative illustration of licences and licence rules.

8.3.2.1 License Calculators/Selectors

Among this first group of tools, the following examples have been considered: LICENTIA¸ ELRA Licence Wizard; LINDAT Open Licence Selector; CLARIN Licence category calculator; RDF Representation of licences; OSSWATCH Licence differentiator.

LICENTIA[1] is a suite of services that support users in finding a suitable licence for their data. It is a licence calculator/selector based on ODRL representations of commonly used licences and can be used in three modes: users select conditions of use (obligations, permissions and prohibitions) and compatible

[1] http://licentia.inria.fr/

licences are shown; users select a licence and see whether it's compatible with certain conditions of use; users select a licence, view it with a visualisation tool and export a RDF representation.
The tool appears easy to use if users know their preferences in terms of permissions, obligations and prohibitions. However, the partition into permissions, obligations and prohibitions may be tricky if not supported by a clear legal definition. Besides, it does not allow multiple choices and it does not provide a broader illustration of licences.

ELRA Licence Wizard[1] is a web configurator that enables to choose among a number of legal features and consequently obtain a suitable licence distribute contents adjusted to their selection. It covers 24 licences (ELRA, Creative Commons and META-SHARE) which are classified according to nine criteria (e.g. use type, whether it requires electronic signatures etc.).
The tool guides users to make their choices with the help of explanatory text. It also allows the user to state multiple preferences. It provides a broader picture of available licences based on users' selection. The language used to guide users' choices is not however always clear (explanations under the question mark that aims to explain the criteria are not always clear) and it does not always have a corresponding legal meaning (see, for instance, the distinction between implicit and explicit). Moreover, it does not provide a graphical illustration that could help users to better visualize the licences' rules.

Similarly, **LINDAT Open Licence Selector**[2] asks users a number of questions (based on licensing conditions, again) and concludes with a set of licences that match his/her answers and which the user can use for his/her resource.
The tool is quite appealing in terms of interface, also allowing a free search and provides a summary of potentially applicable licences. Nevertheless, it does not allow multiple choices neither it really guides users (especially non-professional users) to make their choice.

Another similar tool is **CLARIN Licence category calculator**[3], which suggests a number of labels for conditions of use (aka “Laundry Tags”). It helps users to classify the licence that they would like to use for a certain resource according to the CLARIN licence categories. If the user has not chosen a licence, it also provides a link to a "ready-made" legal text conformant with the licensing conditions the user has selected. If the conditions do not require user identification, it provides a link to the LINDAT Open Licence Selector.
The tool guides the users providing a number of conditions of use identified by labels. However, it seems to be confined to CLARIN, as it implies certain specifications that are narrowed to CLARIN categories. In addition, the meaning of each condition is too synthetic and more examples could be added.

Finally, **OSSWATCH Licence differentiator**[4] aims at helping users to understand their preferences in relation to free and open source software licences.
It guides users specifying in detail the content of their choices. It makes also explicit that users fully read and understand their chosen licence (by stating that “it is no substitute for reading the licences

[1] http://wizard.elra.info/index.php
[2] http://ufal.github.io/public-license-selector/
[3] http://www.helsinki.fi/finclarin/calculator/ClarinLicenseCategory.html
[4] http://oss-watch.ac.uk/apps/licdiff/

themselves [and] the classifications of licence type that enable this tool to work are by necessity somewhat reductive, and therefore output of this tool cannot and must not be thought of as legal advice"), which is often not the case for most of users who do not have a legal training.

8.3.2.2 License Descriptors

Among the license descriptors, the dataset provided by **RDF Representation of licences**[1] contains 126 licenses that are expressed as RDF, while licenses can be also accessed directly.
The tool is a representation of licences with visualisation of basic licensing conditions: a set of commonly used licences with their RDF representation (ODRL & CC-REL). While the dataset contains many licenses, it does not provide guidance to users.

Likewise, **RDF License dataset**[2] also includes licences represented in RDF.
Similar to the previous tool, it is makes use of the MS-rights vocabulary[3]. It contains the same list of licenses as the previously mentioned dataset, but it adds the value of including a keyword- based summary. At the same time, conditions of use could be expanded.

8.3.2.3 Comparative Tables and Graphics

The first example considered is the **META-SHARE Table**[4], which appears in the D6.1.1 META-SHARE Report related to the use of language resources (LRs) and language technologies (LTs) within the framework of META-SHARE.
The tool has the aim to cover as many elements as possible and condense it into a concise graphical representation. Limited to ELRA, LDC (& NIST), CC licenses, it yet provides some unclear information (e.g. with ref. to "Remark" or "Implicit/Explicit"), therefore it is not always straightforward to follow and risks to confuse users to some extent.

A similar tool is the **ORACLE Table**[5], which is intended to compare the major attributes of the most popular Free and Open Source Software licenses.
The chart compares and graphically represents a number of licenses, with the aim to visually compare the main features of most popular free and open source software licenses. However, it appears too synthetic and its own compiler understands the related limits acknowledging the difficulty of fully understand the differences among licenses.
Another comparative graphic tool is the **GNU Table**[6], which knowingly aims at covering also the New Compatible Licenses. In addition to the GNU licenses list, the graphic illustrates some licensing rules in relation to new compatible licenses.

[1] https://datahub.io/dataset/rdflicense
[2] http://rdflicense.appspot.com/
[3] http://purl.org/NET/ms-rights
[4] http://www.meta-net.eu/public_documents/t4me/META-NET-D6.1.1-Final.pdf
[5] https://blogs.oracle.com/davidleetodd/entry/free_and_open_source_license
[6] http://www.gnu.org/licenses/quick-guide-gplv3.html

The chart aims at clarifying the compatibility of a number of free software licenses with GPL and now also GPLv3. Although it offers a quite clear and schematic picture of the relations among licenses, the scope of the chart is inevitably too narrow.

To conclude with, WG3 has considered a **list of other graphical representations**, such as **TLDRLegal**[1] and **GitHub Tool**[2], but also - although of different nature and scope tools like the **European Public Domain Calculator**[3] and **Public Domain Sherpa**[4], the US copyright term calculator.
Regarding these specific tools, during the previous conference calls with WG3 internal and external experts, it emerged that, although they cannot be considered precisely applicable to the scenarios considered by OpenMinTeD nor precisely to the extent that they could be directly applied to the WG3 Compatibility Matrix, they still offer a basis for comparison and therefore they can be considered simply as examples, especially in terms of the methodology behind them and their graphical interface, to look at when developing a more OpenMinTeD tailored tool.

8.3.3 The OpenMinTeD Compatibility Matrix (CM)

As well argued by Labropoulou, Piperidis and Margoni in the framework of the *LREC 2016 Workshop on Cross-Platform Text Mining and Natural Language Processing Interoperability*:

> *"In the field of TDM it is important to properly address the licence compatibility issue by employing a "multi-layer licence approach". The starting point is of course to focus on just one "layer", e.g. content licences or software licences or terms of use, and try to resolve compatibility issues "within" the same type of licences. This means to verify the compatibility of the same kind of licences in order to determine whether two or more content licences can be combined, or two or more software licences can be combined. A multi-layer approach applies the same compatibility principle across the 3 categories identified (content licences, tools or software licences, and service agreements). In this way, it will be possible to develop an interoperability model or matrix that is not limited to content, tools or services individually considered, but that, by taking a holistic approach, is able to offer a more complete analysis of the licence compatibility issues faced by TDM researchers. In other words, this formulation, instead*

[1] https://tldrlegal.com/ and https://tldrlegal.com/compare
[2] http://choosealicense.com/
[3] http://archive.outofcopyright.eu/calculator.html
[4] http://www.publicdomainsherpa.com/calculator.html

of taking a theoretical legal approach, puts at its centre the needs and the skills of TDM researchers, who usually are not legally trained".[1]

A **first draft of the CMs** for (1) contents, (2) software, and (3) terms of use, could look at first like the following. It is important to note at this point that the third column ("Are they compatible?") refers to the possibility to combine the subject matter of column 1 and 2 in a way that under copyright law they form a so called "derivative work[2]". When the combination of two works does not lead to the creation of a derivative work there should be no restriction to the possibility to combine them. Nevertheless, there are cases where the difference is not clear cut and the specific terminology employed by the licences can become decisive. There are instances, however, where two licences interpret their respective terms in different ways. When this is the case, this will be noted in the 4th column.

Note: Tables Table 9, Table 10, and Table 11 include only some of the licenses to be considered. These tables are now being substituted with two axis graphical representations (Tables Table 12, Table 13, and Table 14). The present tables are still in a draft version. Updated versions are to be included with D5.3.

Table 9 - Compatibility Matrix (draft version 1.0): Contents

Licence for resource A	Licence for resource B	Are they compatible?	Under which conditions?
CC BY 4.0	CC BY 4.0	Yes	No restrictions
CC BY 4.0	CC-BY-SA 4.0	Yes	Results under SA
CC BY 4.0	CC-BY-NC 4.0	Yes	
CC BY 4.0	CC-BY-ND 4.0	Yes	
CC BY 4.0	CC-BY-NC-SA 4.0	Yes	But SA
CC BY 4.0	CC-BY-NC-ND 4.0	No	
CC-BY-ND 4.0	CC-BY-ND 4.0	No	
CC-BY-SA 4.0	CC-BY-SA 4.0	Yes	e.g. BY-SA 1.0 only with BY-SA 1.0[3]
CC-BY-SA 4.0	CC-BY-NC 4.0	Yes	Both restrictions apply
CC-BY-SA 4.0	CC-BY-NC-SA 4.0	No	
MS Commons BY	MS Commons BY	Yes	

[1] P. Labropoulou, S. Piperidis, T. Margoni. *Legal Interoperability Issues in the Framework of the OpenMinTeD Project: A Methodological Overview (Abstract)*, in R. Eckart de Castilho, S. Ananiadou, T. Margoni, W. Peters, S. Piperidis (eds.). Proceedings of LREC 2016, Tenth International Conference on Language Resources and Evaluation, May 23-28, 2016, Portorož, Slovenia, *LREC 2016 Workshop on Cross-Platform Text Mining and Natural Language Processing Interoperability*, p. 62, available at: http://www.lrec-conf.org/proceedings/lrec2016/index.html.

[2] Derivative work is an expression that may vary from legal system to legal system. In fact, the term is not present in every Copyright Act. We use this definition: "*A work that is based upon or otherwise derived from another work or a number of works, in particular by means of adapting, editing, modifying, translating the pre-existing work/s regardless of the medium used.*"

[3] https://wiki.creativecommons.org/wiki/ShareAlike_compatibility

Table 10 - Compatibility Matrix (draft version 1.0): Software

Licence for software A	Licence for software B	Are they compatible?	Under which conditions?
GPLv3	GPLv3	Yes	
GPLv3	GPLv2	No	Unless GPLv2 "or any later version"
GPLv3	Apachev2	Yes	The Licences are compatible as long as derivative works are distributed under GPLv3. Apache foundation[1] points out that Apachev2 software may be included in GPLv3 projects, but NOT *vice versa*
GPLv3	EPL	No	(both copyleft)
GPLv3	LGPLv3	Yes	
GPLv2	GPLv2	Yes	
GPLv2	Apachev2	No	
GPLv2	EPL	No	
Apachev2	Apachev2	Yes	
Apachev2	EPL	Yes	
Apachev2	LGPLv2	No	
Apachev2	LGPLv3	No	

Table 11 - Compatibility Matrix (draft version 1.0): Terms of Service

ToS for Service A	ToS for Service B	Are they compatible?	At what conditions?
Google Translate[2]	Google search engine[3]	Yes	
Google Search engine	Twitter[4]	Yes	Their services interact at different levels, e.g. Twitter contents being indexed in the search engine
Google Search engine	Facebook[5]	Yes	However, some conflict may arise at some point (as it occurs with privacy)

[1] This condition is specified by the steward of the license.
[2] https://cloud.google.com/translate/v2/terms
[3] https://www.google.com/policies/terms/
[4] https://dev.twitter.com/overview/terms/agreement-and-policy
[5] https://developers.facebook.com/policy

ToS for Service A	ToS for Service B	Are they compatible?	At what conditions?
Twitter	Facebook	Yes	As long as the licenses to use their services are not conflicting
Twitter Facebook	Dropbox[1]	Yes	As long as the licenses to use their services are not conflicting
CLARIN[2]	Twitter Facebook	No	No third party access to CLARIN resources
Mendeley[3]	SSRN[4]	Yes	
Mendeley	Zotero[5]	Yes	
ContentMine[6]	Mendeley	Yes	

Table 12 - Compatibility Matrix (draft version 2.0): Concent

	CC-0	CC-BY 4.0	CC-BY-NC 4.0	CC-BY-SA 4.0	CC-BY-ND 4.0	CC-BY-NC-ND 4.0	CC-BY-NC-SA 4.0
CC-0	Yes, no restrictions	Yes, no restrictions	Yes, results under NC	Yes, results under SA			
CC-BY 4.0	Yes, no restrictions	Yes, no restrictions	Yes, results under NC	Yes, results under SA			
CC-BY-NC 4.0	Yes, results under NC	Yes, results under NC	Yes, results under NC	Yes, results under both restrictions			
CC-BY-SA 4.0	Yes, results under SA	Yes, results under SA	Yes, results under both restrictions	Yes, results under SA			
CC-BY-ND 4.0					No		
CC-BY-NC-ND 4.0		No					
CC-BY-NC-SA 4.0		Yes, results under SA		No			

[1] https://www.dropbox.com/terms
[2] https://www.clarin.eu/content/licenses-agreements-legal-terms
[3] https://www.mendeley.com/terms/
[4] http://www.ssrn.com/en/index.cfm/terms-of-use/
[5] https://www.zotero.org/support/terms/terms_of_service
[6] http://discuss.contentmine.org/tos

Table 13 - Compatibility Matrix (draft version 2.0): Software

	GPLv3	GPLv2	Apachev2	EPL	LGPLv3	LGPLv2
GPLv3	Yes	No, unless GPLv2 "or any later version"	Yes, but as long as derivative works are distributed under GPLv3	No		
GPLv2	No, unless GPLv2 "or any later version"	Yes	No	No		
Apachev2	Yes, but as long as derivative works are distributed under GPLv3		Yes	Yes	No	No
EPL	No	No				
LGPLv3	Yes		No			
LGPLv2			No			

Table 14 – Compatibility Matrix (draft version 2.0): Terms of Service

	Google search	Google Translate	Twitter	Facebook	CLARIN	Dropbox	Mendeley	SSRN	Zotero	Content Mine
Google search	Yes	Yes	Yes	Yes						
Google Translate						Yes				
Twitter			Yes	Yes	No	Yes				
Facebook					No					
CLARIN			No	No						
Dropbox		Yes	Yes							
Mendeley								Yes	Yes	Yes
SSRN							Yes			
Zotero							Yes			
Content Mine							Yes			