Interoperability Standards and Specifications

June 20, 2017

Deliverable Code: D5.3
Version: 1.0
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

D5.3 – Interoperability Standards and Specifications Report

WP5 – Interoperability Framework	
WP participating organizations: UKP-TUDA, ARC, UNIMAN, INRA, OU, USFD, UvA, UoS, GESIS	
Contractual Delivery Date: 01/2017	**Actual Delivery Date:** 06/2017
Nature: Report	**Version:** 1.0
Public Deliverable	

Preparation slip

	Name	Organization	Date
From	Richard Eckart de Castilho Mouhamadou Ba Penny Labropoulou Thomas Margoni Giulia Dore Wim Peters Angus Roberts Matthew Shardlow Piotr Przybyla Jacob Carter	UKP-TUDA INRA ARC UoG UoG USFD USFD UNIMAN UNIMAN UNIMAN	02/07/2016
Edited by	Richard Eckart de Castilho	UKP-TUDA	06/06/2017
Reviewed by	Petr Knoth Marta Villegas	OU CNIO	12/05/2017 19/05/2017
Approved by	Androniki Pavlidou	ARC	20/06/2017
For delivery	Mike Hatzopoulos	ARC	20/06/2017

Document change record

Issue	Item	Reason for Change	Author	Organization
V0.1	Draft version	First draft	Richard Eckart de Castilho	UKP-TUDA
V0.2	Draft version	Expanded draft	Richard Eckart de Castilho	UKP-TUDA
V0.3	Draft version	Expanded draft	Richard Eckart de Castilho	UKP-TUDA
V0.4	Draft version	Expanded draft	Richard Eckart de Castilho	UKP-TUDA
V0.5	Draft version	Expanded draft	Richard Eckart de Castilho	UKP-TUDA
V0.6	Draft version	Version sent to comments to reviewers	Richard Eckart de Castilho	UKP-TUDA
V1.0	First version	Incorporating reviewers' comments	Richard Eckart de Castilho	UKP-TUDA

Table of Contents

Table of Tables

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
CAS	(UIMA) Common Analysis System (https://uima.apache.org/d/uimaj-2.9.0/references.html#ugr.ref.cas)
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)
ToS	Terms of service
ToU	Terms of use
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 second in a series of three. The first of the three, an earlier version of this deliverable, was delivered as D5.2 It will be updated again in 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 that constitute the bulk of the deliverable package are provided as **attachments** to the present document.

1. Introduction

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

Through an assessment of compliance with these requirements, it is possible to assess the level of interoperability of a content provider's systems, an analytics provider's systems, and other types of resources with OpenMinTeD. As part of the first version of the Interoperability Standards and Specifications Report (D5.2), 72 of these requirements had been identified and described. A set of relevant products and services have had their compliance with these requirements assessed. However, the majority (69) of the requirements reported in D5.2 were so-called "abstract" requirements, i.e. they were described without committing to specific standards, technologies, or implementations. For the preparation of the present deliverable (D5.3), these have been augmented by so-called "concrete" requirements which describe how in particular the abstract requirements can be fulfilled by OpenMinTeD. For example, the abstract requirement REQ-51[1] asserts that a license should be attached to a resource. The related concrete requirement REQ-92[2] specifies the exact location and format of a license file for resources provided as ZIP or JAR files.

Section 2 provides a short summary of the work that has been taking place in the interoperability working groups during the reporting period (M15 – M20). Section 3 provides an overview of the updates to the interoperability requirements specification.

1.1 Working groups

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

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

[1] https://openminted.github.io/releases/interop-spec/1.1.0/openminted-interoperability-spec/#REQ-51
[2] https://openminted.github.io/releases/interop-spec/1.1.0/openminted-interoperability-spec/#REQ-92

2. *Summary Reports*

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

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

2.1 WG1

2.1.1 Mission statement update

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

2.1.2 Mode of operation update

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

2.1.3 Progress in the current period

The current period has been dedicated the following actions:

- **Concrete requirements:** suggestions for making the interoperability requirements concrete; this was carried out through a document shared with other WGs since some requirements are shared among them; feedback for requirements of other WGs has also been provided. The requirements that have been finalised have been integrated into the OpenMinTeD Interoperability Specification[1] published on GitHub. As a part of this activity, also existing requirements have been revised and/or deprecated.
- **OMTD-SHARE schema:** work on the metadata schema has continued: comments by resource providers, as they convert their current descriptors or create new metadata records in the OMTD-SHARE schema have been gathered and influenced changes in the current version of the schema[2]; at the same time, the formalisation of the schema with the assignment of URIs to each schema element and a study into its RDFization (as an OWL ontology) with adoption of the SKOS model for the controlled vocabularies is in progress.
- **Glossary:** the lack of a common terminology has been identified as a problem when discussing strategy in the project. Thus, led by WG1 and in collaboration with the other WGs, work has been initiated on the creation of a common glossary which has been be published online for future reference. [3]

[1] https://openminted.github.io/releases/interop-spec/1.1.0/openminted-interoperability-spec/
[2] https://github.com/openminted/omtd-share_metadata_schema/tree/master/OMTD-SHARE%20v200%20XSD
[3] http://vocabularies.openminted.eu/glossary/OMTDglossary

- **Dissemination and community engagement:** The webinar "Achieving interoperability between resources involved in TDM at the level of metadata"[1] has been presented on November 10 as part of the OpenMinTeD Fall 2016 Webinar Series on Interoperability (see Appendix 7.3). Details have been reported in MS22 “Working groups interim meeting report 3”.

2.1.4 Tasks planned for the next period

- **Documentation and dissemination of OMTD-SHARE:** the OMTD-SHARE metadata schema that is being developed by WG1 will be further documented in the Interoperability Guidelines (D5.5 and D5.6) and the support material generated within T5.3. The guidelines and presentation of the schema will be deployed for disseminating the schema to user communities and getting feedback from them. This dissemination is expected to drive the improvement of compliance of related products with the requirements from the OpenMinTeD interoperability specification (D5.2 – D5.4).
- **Incorporation of feedback into OMTD-SHARE:** Given the increasing level of development activity at the level of the platform and the testing of the related software, we anticipate a lot of feedback regarding improvements of the OMTD-SHARE schema.
- **Elaboration and refinement of requirements:** Work on the finalisation of the interoperability requirements will also continue, and, if deemed necessary, addition of new requirements will be investigated.

2.1.5 State of operations

WG1 is advancing work at a steady pace; however, the task of developing concrete requirements has proven more difficult than anticipated and has shown certain problems with some of them, which have produced interesting discussions and, in some cases, reviews or postponements.
Overall, 18 abstract requirements have been studied with the aim to produce concrete recommendations for them. Of these, 4 requirements have been deprecated and 3 have been postponed as they need further discussions. 16 new concrete requirements have been created, with a reference to the abstract requirement(s) they address; in addition, WG1 has contributed to 10 concrete requirements issued by the other WGs in so far as they had some reference to metadata issues.

2.2 WG2

2.2.1 Mission statement update

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

2.2.2 Mode of operation update

The mode of operation remains the same. For details, please refer to D5.2.
Following the finalization of D5.2, there has been a change in the leadership of WG2.

2.2.3 Progress in the current period

The current period has been dedicated the following actions:

[1] https://www.fosteropenscience.eu/content/achieving-interoperability-between-resources-involved-tdm-level-metadata

- **Strategy:** Following the change in leadership, the scope of WG2 has been put to discussion in order to ensure that the WG continues working towards a shared strategic vision. This has involved a re-examination of the use of linked data and semantic web paradigms by the WG, and the extent to which OpenMinTeD should provide mappings between resources. As a result, the attention of the group shifted from the creation of mappings between knowledge resources and type systems towards examining suitable representations for knowledge level interoperability. Web Annotation[1] and UIMA XMI[2,3] (with appropriate type system) were examined, an example use-case based on these formats created, and a decision taken to use Web Annotation as an exchange format for external interoperability, and UIMA XMI for internal interoperability. Further work in this area has subsequently been passed to WG4 (Workflow), by the project management board. WG2 has subsequently moved back to considering individual knowledge resources. By involving subject domain representatives, we are building a complete picture of all resources used by use cases, and checking the fit between these and OpenMinTeD interoperability requirements.
- Discussions on the linked data and semantic web paradigms for knowledge level interoperability. In particular, WG2 has examined the extent to which these paradigms are used within our use case domains (generally, quite widely), and whether adoption of a pure linked data / semantic web approach would exclude some communities, and some important resources. We have adopted an approach of encouraging the use and uptake of linked data resources, whilst pragmatically recognising that we will still have to support other representations.
- Examination of life science knowledge resources, and their use in a prototype annotation service.
- **Dissemination and community engagement:** The webinar "Interoperability at the knowledge level"[4] has been presented on November 6 as part of the OpenMinTeD Fall 2016 Webinar Series on Interoperability (see Appendix 7.3).

2.2.4 Tasks planned for the next period

- Further elaborate the OpenMinTeD use of Web Annotation and UIMA XMI for knowledge level interoperability.
- Examine and specify inclusion of knowledge resources in the OpenMinTeD registry.
- Engage with use-case partners to further investigate the use of knowledge resources in our existing use cases, and come up with a set of recommendations in conjunction with use case partners.

2.2.5 State of operations

WG2 made progress in the prior period. Internal working group teleconferences have been less frequent than the monthly plan, however, this was compensated for during a three-day working meeting in Manchester in December 2016 with stakeholders from WG2, WG4 and WP6.

[1] https://www.w3.org/annotation/
[2] http://www.omg.org/spec/XMI/
[3] https://uima.apache.org/d/uimaj-2.9.0/references.html#ugr.ref.xmi
[4] https://www.fosteropenscience.eu/content/text-mining-interoperability-knowledge-level

2.3 WG3

2.3.1 Mission statement

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

2.3.2 Mode of operation

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

2.3.3 Progress in the current period

The current period has been dedicated the following actions:

- **Licenses and rights statements:** Following the tasks planned in D5.2, the WG3 "Licenses and rights statements" working plan has been implemented according to the intense discussion with internal and external experts. As a result, a flow chart (version 0.1) has been drafted to help defining an OpenMinTeD rights statements list through a graphical illustration of the main legal dynamics that apply to text and data mining. With a similar aim, a legal annotation experiment has been initiated, with 22 licenses being annotated to date (including, but not limited to, the Creative Commons set of licences Version 4.0, Apache License Version 2.0 and a few Terms of Services (ToS)), and the related guidelines have been drafted with WG1 to instruct independent annotators.
- **Compatibility matrix:** Additionally, the WG3 Compatibility matrix has undergone further expansion and has been refined and adjusted to meet the comments of both internal and external experts. The matrix (in its version 1.0) has been preliminary reviewed by one of the WG3 internal legal expert and it has now been submitted for a more formal external review. The working document in question (version 0.1) was illustrated during the WG3 webinar "Text and Data Mining interoperability at the legal level: rights, exceptions and licenses"[1] held on the 23rd of November 2016.
- **Concrete requirements:** In line with WP5 Interoperability Specifications, the embodiment of abstract requirements has begun and the WG3 concrete requirements for legal interoperability are being outlined, while the working group is also specifically contributing to the WG1 concrete licensing requirements.
- **Glossary:** The legal glossary has also been expanded and it is now integrated with the wider Glossary[2] disseminated by WG1.
- **Dissemination and community engagement:** In collaboration with task T3.3 (WP3) a set of legal Frequent Asked Questions (FAQs) has been provided (version 1.0), to be used not only for training purposes but also to serve as policies for OpenMinTeD: it will be used to train researchers/users, but it also aims at being a support for the policies that need to be drafted to use the OMTD platform.

2.3.4 Tasks planned for the next period

- Following the mission of the WG to explain and simplify the legal aspects of text and data

[1] https://www.fosteropenscience.eu/content/text-and-data-mining-interoperability-legal-level-rights-exceptions-and-licences

[2] http://vocabularies.openminted.eu/glossary/OMTDglossary

mining, but also to promote licensing compatibility, WG3 plans submitting the Compatibility Matrix and the workflow for external review (first circulating it among WG3's external experts). The next step is to implement a tool based on the graphical compatibility illustrations for automated selection and elicitation of licensing information.

- Further to the tasks described above, in the short term WG3 aims at completing the WG3 concrete requirements and circulating them for review at the internal and external level; completing the FAQs list in progress, with the purpose of achieving a better and larger coverage; following up with the Annotation experiment annotating a number of licences (providing at least two independent annotations for each licence and ToS considered) to draw and refine the OpenMinTeD rights statements list that is indeed machine readable.

2.3.5 State of operations

The activities of the WG are aligned with the overall working plan. The interaction with other WPs, in particular contributions to the related reports spur the work within the WG.
Due to the moderate decrement of external expert participation in WG3 con-calls, the WG additionally more strongly reaches out to individual experts, e.g. by setting up more tailored conference calls according to different time zones and contacting individually experts that are not able to participate to the main calls, in order to verify and corroborate the legal findings and working documents so far devised.

2.4 WG4

2.4.1 Mission statement

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

2.4.2 Mode of operation

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

2.4.3 Progress in current period

- **Interoperability requirements:** work on the interoperability requirements has continued. Existing requirements have been partially refined or deprecated as some were found to be rather functional requirements e.g. for the OpenMinTeD Registry Service or OpenMinTeD Workflow Service. Statistics on the number of finalised, draft and deprecated requirements are available in Tables 1, 2 and 3 below. New concrete requirements were defined such that most of the abstract requirements are now paired with at least one concrete requirement explaining a specific approach or technology supported by OpenMinTeD that can be used to meet the abstract requirement.
- **OpenMinTeD Script:** in order to discover and investigate concrete interoperability issues when mixing TDM/NLP components from different sources and platforms, we implemented OpenMinTeD Script as a sandbox environment. It provides a domain specific language for defining workflows and provides adapters that integrate whole sets of components, e.g. all GATE components, all DKPro Core components, all Argo components, etc. instead of integrating each component separately. This revealed new/hidden interoperability issues, e.g. related to differences in the way annotations are modelled in different frameworks, and led to concrete

discussions as to how to overcome them. We expect that parts of the implementation may continue to be evolved as part of the OpenMinTeD Workflow Service, e.g. to serve as an adapter between the Galaxy Workflow Editor and the underlying NLP/TDM components.

- **Discussions about the need for a type system:** the need for endorsing existing type systems or creating a new type system to promote interoperability between components was discussed extensively. As a result, it was decided that OpenMinTeD will not prescribe, endorse, or create a new type system at this time and that instead every component provider may use whatever type system they want. This is consistent with the feedback we received from external experts during WG4 presentation at the OpenMinTeD Interoperability Webinar Series Fall 2016 (cf. Appendix 7.3). They recommended that OpenMinTeD should first focus on making components run on the platform and later return to the details of component interoperability within a workflow. However, if it was decided that OpenMinTeD would support specifically two data formats for annotations: the XML Metadata Interchange[1] (XMI) format, specifically the representation of a UIMA CAS[2], will be used to encode annotations on text in particular when exchanging data between components within a workflow; the RDF-based WebAnnotation[3] standard will be used to make annotations produced by OpenMinTeD workflows accessible to third parties and to encode annotations above the text level, e.g. on document/collection level. First prototypic examples[4] of how data in both of these formats may be encoded have been created for a use-case that annotates funding information in publications.
- **Dissemination and community engagement:** The webinar "A minimalist approach to workflow interoperability"[5] has been presented on November 16 as part of the OpenMinTeD Fall 2016 Webinar Series on Interoperability (see Appendix 7.3).

2.4.4 *Tasks planned for next period*

- Further concretisation and refinement of the interoperability requirements.
- Elaboration and facilitation of the process for packaging and deploying components and the associated requirements. Also, contributing the outcome to an upcoming Interoperability Guidelines deliverable.
- Returning to investigate the need of endorsing or creating a specific type system or set of type systems in OpenMinTeD to facilitate interoperability between components within a workflow.
- Create a Web Services framework for the integration of remote UIMA components into a workflow. Provide wrappers at both client and server side, to facilitate integration.
- Decide upon a protocol for writing Web Annotations. Implement a component to do this on the OpenMinTeD platform.

[1] http://www.omg.org/spec/XMI/
[2] https://uima.apache.org/d/uimaj-2.9.0/references.html#ugr.ref.xmi
[3] https://www.w3.org/annotation/
[4] https://github.com/openminted/interop-examples
[5] https://www.fosteropenscience.eu/content/minimalist-approach-workflow-interoperability

2.4.5 State of operations

- WG4 is performing strongly with regular conference calls well attended by OpenMinTeD partners.
- Communication with external experts happens largely through individual members of the WG. External experts do not participate directly in the conference calls, but they do sometimes contribute to the discussion forums.
- Given the strongly growing number of requirements (specifically from WG1 and WG4), an approach needs to be defined how to organize and/or refine these to make them better accessible from a user's perspective.

2.5 Publications

This section lists publications relevant to this deliverable from project partners within the report period.

- R. Eckart de Castilho, 2016. **Automatic Analysis of Flaws in Pre-Trained NLP Models**. In Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI3nOIAF2) at COLING 2016, p. 19-27, Osaka, Japan
  Relation to OpenMinTeD: The paper describes an approach of analysing pre-trained models for common types of flaws that can have a negative impact on the performance of analysis workflows making use of them. Providing easy access to pre-trained models is one of the added values of platforms such as OpenMinTeD. The proposed approach could be integrated into OpenMinTeD to provide users with useful information about the quality of pre-trained offered through the platform.

3. *Requirements*

This section outlines the structure of requirements and presents the requirements generated so far.

3.1 Transition from abstract to concrete requirements

After defining mainly abstract requirements during the preparation of D5.2, for D5.3 we have focussed on concretizing the requirements. Where possible, the concrete requirements make use of existing standards, best-practices, and implementations. However, the elicitation process also points out precisely in which areas currently no accepted standards, best-practices, and implementations exist. These include for example a categorization scheme for components and packaging of components and models. Such areas need special coverage in the Interoperability Guidelines (D5.5 and D5.6) as OpenMinTeD needs to define best-practices here.

3.2 Requirements overview

This section provides an overview of the interoperability requirements that have been generated so far. A total of 111 requirements have been generated by the WGs, many of which are applicable across the WGs (WG1: 40, WG2: 27, WG3: 25, WG4: 44). These can be broken down by status:

- 53 requirements in status “draft” (Table 1)
- 31 requirements in status “final” (Table 2)
- 27 requirements in status “deprecated” (Table 3)

As mentioned before, a focus in this period was the creation of concrete requirements that describe specific OpenMinTeD-supported ways of implementing the abstract requirements. Thus, there are now 42 concrete requirements and 69 abstract requirements. Of these requirements, 51 are recommendations, 53 are mandatory, and a few are optional (6).
In this document, we only provide the requirements overview with their short titles. The full requirements specification is provided as an attachment to this document and is also publicly hosted on our GitHub repository. We recommend browsing the requirements as hosted on GitHub[1] as this is a highly-cross-referenced hypertext.

Table 1 - Requirements in status "draft"

REQ	Requirement	Concreteness	Strength	WGs
5	Components must 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 must declare whether they can be scaled within a workflow	abstract	mandatory	WG4
13	Citation information for component should	abstract	recommended	WG1, WG4

[1] https://openminted.github.io/releases/omtd-share/2.0.0/

	be included in the metadata			
51	License should be attached	abstract	recommended	WG3
53	Licensor must be entitled to grant license	abstract	recommended	WG3
54	Licensees should remain with a copy of the license	abstract	recommended	WG3
55	Standard licenses should be used	abstract	recommended	WG3
56	License should be machine readable	abstract	recommended	WG3
57	License 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 license grant	abstract	recommended	WG3
63	License must qualify for Open Access rights	abstract	recommended	WG3
64	License must qualify for Open Access uses	abstract	recommended	WG3
65	License must qualify for Open Access must not restrict use in any way	abstract	recommended	WG3
66	License must qualify for Open Access may include attribution requirements	abstract	recommended	WG3
73	Stick to widely used data compression formats	concrete	best practice	WG4
74	Machine-readable metadata for UIMA components	concrete	mandatory	WG1, WG4
75	Embedding UIMA component metadata into the source code	concrete	mandatory	WG1, WG4
76	Separating UIMA metadata from the component	concrete	mandatory	WG1, WG4
78	Specifying input and output types of UIMA components	concrete	mandatory	WG1, WG4
79	Documentation of UIMA components	concrete	mandatory	WG1, WG4
81	Embedding GATE component metadata into the source code	concrete	mandatory	WG1, WG4
83	Documentation of GATE components	concrete	mandatory	WG1, WG4
84	Separating GATE metadata from the component	concrete	mandatory	WG1, WG4
88	Embedding output format in UIMA component metadata	concrete	mandatory	WG1, WG4
89	Version documentation in parallel with component/resource	concrete	recommended	WG1, WG2, WG3, WG4

90	Components must be assigned at least one category from the OMTD-SHARE controlled vocabulary for component types	concrete	mandatory	WG1, WG4
91	Encoding citable publications (for scholarly attribution) in resource metadata records	concrete	recommended	WG1, WG2, WG4
92	Including license text in resource packages	concrete	recommended	WG1, WG3, WG4
93	Provide identifiers for knowledge resource elements	concrete	recommended	WG2
94	Data Category Linking Vocabulary	concrete	recommended	WG2
95	The KR format should be in a standard format such as XML, JSON-LD or RDF/XML.	concrete	recommended	WG2
96	Unique identifiers and versions for components using Maven	concrete	mandatory	WG1, WG4
97	Declaring scale out capability in UIMA	concrete	mandatory	WG4
98	Publishing components via software repositories (Maven, Docker)	concrete	mandatory	WG4
99	Encoding in the metadata a direct access link for content resources	concrete	mandatory	WG1
100	Providing access to content resources (sharing/exposing and transferring)	concrete	mandatory	WG1
101	Making models and annotation resources accessible as entities distinct from the components they are compatible with	concrete	recommended	WG1, WG2, WG4
102	Adding version information in the metadata descriptions of all resources	concrete	mandatory	WG1, WG2, WG3, WG4
103	Specifying access mode of resources and encoding it in the metadata descriptions	concrete	mandatory	WG1, WG2, WG4
104	Encoding funding information in the metadata descriptions of all resources	concrete	recommended	WG1, WG2, WG4
105	Encoding of format in the metadata description of content resources	concrete	mandatory	WG1
106	Encoding licensing terms in the metadata description of the resource	concrete	mandatory	WG1, WG3
107	Encoding metadata on domain/subject/ classification for all resources when applicable	concrete	recommended	WG1, WG2
108	Encoding language information in the metadata of content resources	concrete	mandatory	WG1, WG2
109	Encoding statistical information in the content resources	concrete	mandatory	WG1, WG2
110	Assigning a unique persistent identifier for all resources	concrete	mandatory	WG1, WG2, WG3

111	Annotation schema dependencies for UIMA components using Maven	concrete	mandatory	WG4

Table 2 - Requirements in status "final"

REQ	Requirement	Concreteness	Strength	WGs
1	Components must be described by machine-readable metadata	abstract	mandatory	WG4
2	Component metadata have to be embedded into the component source code	abstract	mandatory	WG4
3	Component metadata must be separable from the component	abstract	mandatory	WG4
4	URL to actual content must be discoverable	abstract	mandatory	WG1, WG2, WG3
7	Components must have a fully qualified name that follows the Java class naming conventions	concrete	mandatory	WG4
8	Components must associate themselves with categories defined by the OpenMinTeD project	abstract	mandatory	WG4
9	Components must declare their annotation schema dependencies	abstract	mandatory	WG4
12	Components should provide documentation describing their functionality	abstract	recommended	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 parametrizable options of the components should be identified and documented	abstract	recommended	WG4
26	It should be possible to determine the source of an annotation/assigned category	abstract	recommended	WG4
28	Processing components should be downloadable	abstract	recommended	WG4
33	Licensing information must be included in the metadata	abstract	mandatory	WG1, WG3
34	Licensing information should be expressed in a machine-readable form	abstract	recommended	WG1, WG3
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	abstract	recommended	WG1

	metadata of the resource			
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
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
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
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-LD or RDF/XML.	abstract	recommended	WG2

Table 3 – Requirements in status "deprecated"

REQ	Requirement	Concreteness	Strength	WGs
14	Components must maintain License information	abstract	mandatory	WG4
15	Human readable information should be provided by each resource	abstract	recommended	WG1, WG4
18	Workflows should be described using an uniform language	abstract	recommended	WG4
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 License Aggregation in Workflows	abstract	recommended	WG3, WG4
24	Using/treating workflows as components	abstract	mandatory	WG4
25	Incorporation of multiple resources in parallel	abstract	recommended	WG4
27	Components should handle failures gracefully	abstract	recommended	WG4
29	The actual content of all content resources must be discoverable	abstract	mandatory	WG1, WG2, WG3
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
35	All resources must include a unique persistent identifier	abstract	mandatory	WG1, WG2, WG3, WG4
40	Component metadata must include standardised categories/tags that make them easy to discover	abstract	mandatory	WG1, WG4
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
48	All resource metadata records must include a reference to the metadata schema used for their description	abstract	mandatory	WG1, WG2, WG3, WG4
49	Metadata of tools should contain information about the models available for them	abstract	recommended	WG1, WG4
52	License information must be in metadata	abstract	recommended	WG1, WG3
77	Unique identifiers and versions for UIMA components	concrete	mandatory	WG1, WG4
80	Common elements to represent/describe an executable workflow	concrete	recommended	WG4
82	Unique identifiers and versions for GATE components	concrete	mandatory	WG1, WG4
85	Unique identifier and version for components in the OMTD-SHARE schema	concrete	mandatory	WG1, WG4
86	Attaching format properties to the description	concrete	recommended	WG1, WG4

	of the inputs and outputs that use file			
87	Embedding language capability in UIMA component metadata	concrete	mandatory	WG1, WG4

4. Compliance

In the previous section, we discussed the requirements for interoperability that WGs in OpenMinTeD have identified so far. But unless relevant products are compliant with these, the requirements are ineffective. In 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.
The compliance assessment provided here differs from the infrastructure evaluation performed by Task 4.4. T4.4 evaluates the platform services and interoperability specification from the perspective of applications that implement user-oriented scenarios. The assessment here is performed from the perspective of frameworks and resources that are relevant to the platform as a whole, not to specific user-oriented scenarios.

4.1 Compliance levels

Requirement compliance levels

- **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.

4.2 Relevant products

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 they have developed and are, therefore, providing metadata themselves (Frontiers, Alvis, Argo/U-Compare, DKPro Core, ILSP)
- Resources that they are already using 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, licenses)
- Resources that are being collected from the original data providers who also supply the metadata descriptions (CORE, OpenAIRE).

The list of assessed relevant products has not changed since D5.2. Table 4 lists the products and the number of assessed requirements. Additionally, it shows the number of requirements that had been reported as assessed previously in D5.2. Due to the addition of new requirements and the deprecation of existing requirements, the number can have either increased or decreased.

Table 4 - Assessed products and consulted sources

Product	Assessed requirements (excl. deprecated)	Source
ARGO	44 (was 35)	http://argo.nactem.ac.uk/
AGROVOC	26 (was 16)	http://aims.fao.org/aos/agrovoc/void.ttl
Alvis	37 (was 36)	http://www.quaero.org/module_technologique/alvis-nlp-alvis-natural-language-processing/
CLARIN CR	8 (was 5)	https://www.clarin.eu/ccr
CORE	18 (was 11)	https://core.ac.uk
DKPro Core	46 (was 36)	https://dkpro.github.io/dkpro-core/documentation/
Frontiers	18 (was 11)	http://home.frontiersin.org/about/author-guidelines
GATE	36 (was 35)	https://gate.ac.uk/sale/tao/split.html
ILSP	44 (was 13)	https://inventory.clarin.gr/
JATS	13 (was 15)	http://jats.nlm.nih.gov/about.html
LAPPS Grid	25 (was 15)	http://vocab.lappsgrid.org/
Licences	4 (was 6)	variety of standard licences, such as CC and FOSS
OLiA	25 (was 15)	http://acoli.cs.uni-frankfurt.de/resources/olia/
Ontolex	8 (was 5)	https://www.w3.org/community/ontolex/
OpenAIRE	18 (was 11)	https://guidelines.openaire.eu/en/latest/
TheSOZ	26 (was 16)	http://lod.gesis.org/thesoz/de.html
Apache UIMA	0 (was 1)	https://uima.apache.org
schema.org	8 (was 5)	http://schema.org

5. 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, to elaborate existing requirements and to generate new requirements.
The actions described in this section describe mainly what needs to be done in order to increase the compliance of the relevant products with the OpenMinTeD interoperability specification, i.e. actions that need to be performed by partners or third-parties responsible for these relevant products. Actions that relate directly to the respective WGs are listed in Section 2 under the "planned actions" for each WG.

5.1 WG1

For the compliance actions, we have looked at the 26 concrete requirements that involve directly or indirectly WG1. The recommended actions have to do with the provision of OTMD-SHARE metadata records for each product, at least at the minimal level, given that this includes all the mandatory and recommended metadata elements identified by WG1.
Different types of actions are required depending on

- whether the metadata records are provided by the resource creators themselves straight into OpenMinTeD or imported via other actors;
- whether the metadata information already exists in some other format and needs to be converted or whether it is not yet recorded.

It should be noted that some of these actions are already planned by the resource providers to be accomplished within the project: for instance, for Alvis the creation of the OMTD-SHARE formal metadata records has already started, Frontiers is planning their metadata according to the recommendations, Argo and GATE are changing some of their implementation (e.g. embedding component metadata in the source code, separating metadata from code).
In other cases, the compliance cannot be achieved because of different implementation strategies, e.g. Argo (REQ-74 machine-readable metadata, #documentation of UIMA components), ARGO and GATE (REQ-101 separating components from ancillary resources).
More specifically, the following actions must be pursued:

- **Minimal OMTD-SHARE metadata records:** For resources that have no formal metadata descriptions (Alvis, JATS, OLiA, LAPPS exchange vocabulary), providers (or partners assigned with this task) have to create appropriate metadata records compatible with the OMTD-SHARE schema, at least at the minimal level.
- **Convert existing metadata records to OMTD-SHARE:** For resources that already have metadata descriptions in some other format, providers (or partners assigned with this task) must convert their metadata descriptions to the OMTD-SHARE schema; this includes:
    - components that have embedded metadata in the source code and/or in the form of Maven POM XML;
    - knowledge resources that come from the Linked Data community with their own metadata elements;
    - scholarly publications with metadata in various formats harvested by major aggregators who are converting these first into their own metadata schema (and enriching them

with information with their own tools); part of this work is carried out in Task 5.5 where the data connector for digesting research articles from various repositories and publishers' web sites into OpenMinTeD is built.

- **Complete metadata records:** For specific metadata elements that miss from the original metadata descriptions:
  - When the resources are described by the providers themselves, the information must be provided either in the metadata descriptors used by each provider and then converted to the corresponding OMTD-SHARE elements and values or directly added to the OMTD-SHARE metadata records; for instance:
    - Argo: requirements REQ-78 input-output types, REQ-88 output format, REQ-90 component type, REQ-91 citable publications, REQ-92 and REQ-106 license;
    - ILSP: REQ-88 output format, REQ-90 component type, REQ-91 citable publications, REQ-92 and REQ-106 license;
    - DKPro Core: REQ-90 component type, REQ-91 citable publications, REQ-92 and REQ-106 license (for models);
    - GATE: REQ-90 for component type, REQ-91 citable publications, REQ-92 and REQ-106 license.
  - When the resources are described by other actors, the problem is that this information is not always discoverable; thus, the relevant actions are (a) to include the information if available and (b) promote the inclusion of this information (especially if the metadata element is mandatory) to the relevant communities; for instance:
    - all knowledge resources: REQ-91 citable publications, REQ-92 license, REQ-104 funding information;
    - OpenAIRE and CORE: REQ-99 for direct access link, REQ-100 access to content, REQ-104 funding information, REQ-105 format, REQ-106 license.
- **Other:** Other actions that are indirectly relevant to the provision of metadata records:
  - Separate metadata from source code (GATE, REQ-84);
  - Ensure that components are identified as separate entities in the MAVEN repository (GATE, Alvis and ILSP, REQ-96);
  - Separate components from ancillary resources (e.g. ML models) and provide separate descriptions for them (REQ-101);
  - Provide documentation for the metadata in the format required by each framework (GATE, REQ-83; Argo and ILSP REQ-79);
  - Adopt and promote standards and controlled vocabularies for specific metadata elements (e.g. language codes REQ-75 and REQ-108, version REQ-102, license REQ-106).

5.2 WG2

WG2 is concerned with how knowledge resources comply with the OpenMinTed requirements. There are 13 concrete requirements that involve WG2 and knowledge resources, which can be split into three groups:

- **Metadata related requirements:**
  - REQ-91, REQ-102, REQ-103, REQ-104, REQ-107, REQ-108, REQ-109
- **Linked data related requirements:**

- REQ-93, REQ-94, REQ-95, REQ-110
- **Resource delivery related requirements:**
  - REQ-89, REQ-101

The remainder of this section considers the WG2 requirements within each of these three groups.

Metadata related requirements

Compliance is patchy, and although some resources meet many of these requirements (Agrovoc and TheSoz) none manage to fully comply with all. (e.g. REQ-91 “include citable publications for attribution” is not being met by any knowledge resources. It is suggested that two actions can increase compliance:

- Write to these resource providers outlining how they can comply with OpenMinTeD, if they so wish;
- Create metadata records for knowledge resources within the project.

Linked data related requirements

Compliance with these requirements is good for the selected resources, in part because selection of resources was biased to linked data resources. Two issues need resolving, as follow:

- JATS fails to meet requirements REQ-94 “linking to other resources” and REQ-95 “standard format”. It is suggested we investigate why JATS does not meet these.
- REQ-94 “Knowledge resource authors should provide linkage between their own resource and others” is not met by JATS (to be actioned as above) and not met by CLARIN CCR. This needs to be checked and discussion instigated with CLARIN.

Resource delivery related requirements

- REQ-101, “Making models and annotation resources accessible as entities distinct from the components they are compatible with” is really a requirement on components, not knowledge resources. All knowledge resources are accessible independently, and so in a sense already comply with this. It is not clear, however, whether the requirement means that all resources should be stored in Maven repositories. This needs to be checked.
- REQ-89, “Version documentation in parallel with component/resource”, is not met by any knowledge resources. It needs to be checked how versioning applies to knowledge resources in general.

5.3 WG3

In the broader framework of the OpenMinTeD Interoperability specifications, the concrete requirements considered by WG3 that comprise legal interoperability are identified as it follows:

- REQ-89 (Version documentation in parallel with component/resource) which, when applied to licensing, it recommends providing the specific version of the licence, although this may result in providing a reference to "any later version" of a given licence.
- REQ-92 (Including license text in resource packages), which aims at recommending the inclusion or attachment of the full legal document containing the licensing terms and embodied in the

license text (generally a license.txt file) to a given resource within the whole packaged resource, but only some components actually contain it.

- REQ-102 (Adding version information in the metadata descriptions of all resources) that, applied to licensing, mandates the exact indication of the licence version the version of the licence in its metadata, generally consisting in provision of progressive numbers and overall featuring most of licences.
- REQ-106 (Encoding licensing terms in the metadata description of the resource), which mandates indicating the licensing terms that govern a given resource, only partially met by products that provide it in licence.txt files or free-text, but fully met by products that include it in void.ttl files or make otherwise possible to extract such information from other files.
- REQ-110 (Assigning a unique persistent identifier for all resources) that, applied to licensing, mandates assigning unique and persistent identifiers such as URIs to licences, especially when the license documentation cannot be attached to the resource and yet the license must be publicly available. This practice of ensuring the license is permanently available may have become a standard, but there is no certainty that it will be with any licence.

In terms of their compliance, the above requirements are met only to a limited extent, with some of them being more challenging than others:

- In particular, REQ-102 and REQ-110 do not pose particular challenges since they are implemented by most licensing products, but not all.
- Similarly, provided that any licence information should point to the exact licence version, implementation of REQ-89 may be reached by a reference to any later versions.
- On the contrary, REQ-92 and REQ-106 appear more difficult to implement, in part given the more demanding feature of practically ensuring the extraction or indication of licensing terms for every resource. Particularly when more than one licence applies, both version of the licensed product and version of licence should be properly labelled.

Fort the next phase, focusing on the actions that could be performed to ensure better compliance, it is worth considering:

- With regard to the concrete requirements listed above, especially when these are mandatory, it becomes essential to investigate to what extent the difficulty of complying depends on technical or practical strands, e.g. with reference to REQ-106.

At the same time, in order to improve the overall compliance, further deprecation of some requirements may be considered, similarly to the previous deprecation in terms of strength (from mandatory to recommended) that has been already made necessary with abstract requirements.

5.4 WG4

There are 23 concrete requirements concerning WG4 (17 mandatory and 6 recommended) and the relevant products are already fully compliant with 4 of these. The majority of these requirements relate to the metadata of resources.

Firstly, there are the requirements that involve adding missing-but-known metadata to resources or minor refactoring of existing metadata. These should be fairly trivial for the resource providers to implement, and are all considered mandatory.

- Alvis: REQ-96, REQ-102, REQ-111
- Argo: REQ-74, REQ-75, REQ-78, REQ-79, REQ-88, REQ-102, REQ-103
- GATE: REQ-83, REQ-84, REQ-96
- ILSP: REQ-74, REQ-75, REC-79, REQ-88, REQ-96, REQ-102

None of the resource providers currently have a mechanism for including the component type(s) within its metadata (REQ-90). For compliance, providers will either have to manually add this information into the OMTD-SHARE metadata records or produce a new mechanism from which it can be extracted.

There are a number of recommended requirements regarding metadata that may not always be applicable or simple not known by the resource providers (e.g. funding information, citable publications). It is quite likely that some of these requirements will not be implemented (e.g. there are no plans for the introduction of funding information by any of the providers).

- Alvis: REQ-91, REQ-92, REQ-104
- Argo: REQ-91, REQ-92, REQ-104
- DKPro Core: REQ-91, REQ-92, REQ-104
- GATE: REQ-91, REQ-92, REQ-104
- ILSP: REQ-91, REQ-92, REQ-104

Other requirements simply involve the refactoring of existing resources. These again are all recommended, and it's possible that not all of them will be adopted (e.g. GATE will never support REQ-17 - Components should be stateless).

- Alvis: REQ-17, REQ-89, REQ-101
- Argo: REQ-17, REQ-89, REQ-101
- DKPro Core: REQ-17, REQ-101
- GATE: REQ-89, REQ-101
- ILSP: REQ-89, REQ-101

Finally, the adoption of Maven by resource providers to fulfil a number of other requirements (e.g. REQ-96, REQ-111) means that REQ-98 will require no other work beyond simply publishing resources to Maven Central, using the guide referred to in the requirement.

6. *List of attachments*

Updated for this deliverable

- Detailed Interoperability Specification v1.1.0 (updated for this deliverable)
  - https://openminted.github.io/releases/interop-spec/1.1.0/openminted-interoperability-spec/

No change since D5.3

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

7. Appendix

7.1 WG3 Flow chart v0.1

The Flowchart was drafted to graphically represent the contents of the WG3 Pilot Table, now considered the main reference to draft the prototype/pilot for an OMTD RS list. Among other purposes, it aims at illustrating the flowing process of limitations and restrictions that may apply to the mining activity (e.g. NC, ND), based on a number of variables such as (a) who is mining - whether an individual or a legal entity (e.g. researcher affiliated with the University); (b) the purpose of the activity; (c) the potential secondary use of a given resource. The resulting flowing process should lead to a final action/decision that helps to clarify if and to what extent a limitation applies.
Ideally, there would be multiple flow charts, rather than one single flowchart, which would each address different limitations and end with separate actions/decisions. The outcome of the flow chart will be defined (e.g. is this use NC/ND?) consequent action/decision will be highlighted. Yet, one final flow chart may be drawn to connect the individual flow charts and lead to a common action/decision.
Later on, the tool yEd Graph Editor will be used to draw the flow charts (only test-driven online).

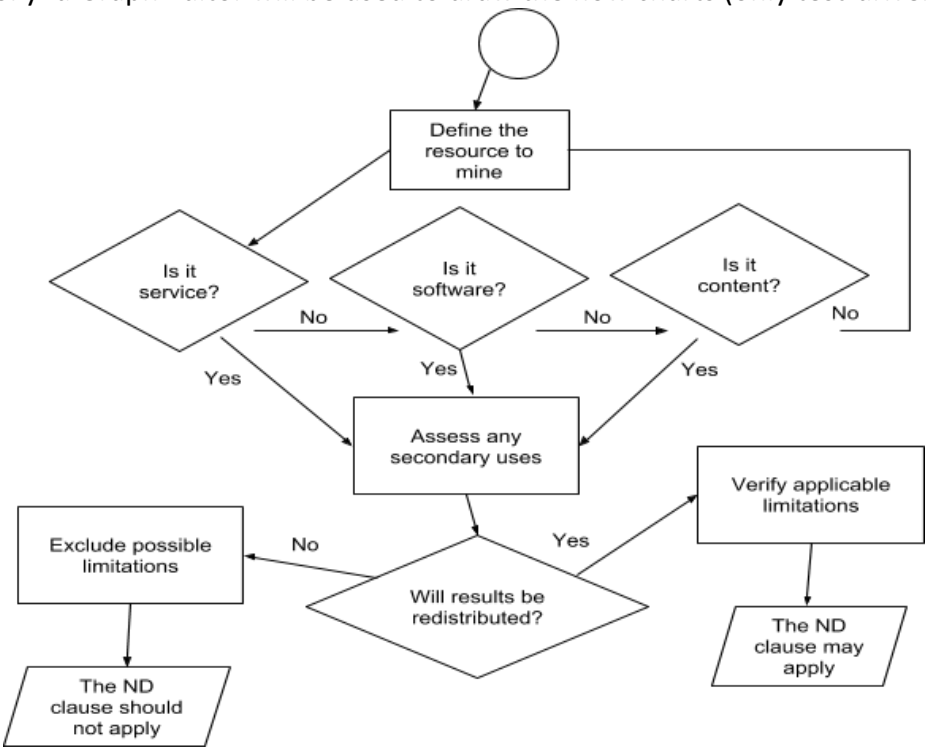

Figure 1: Does the "no derivatives" clause apply?

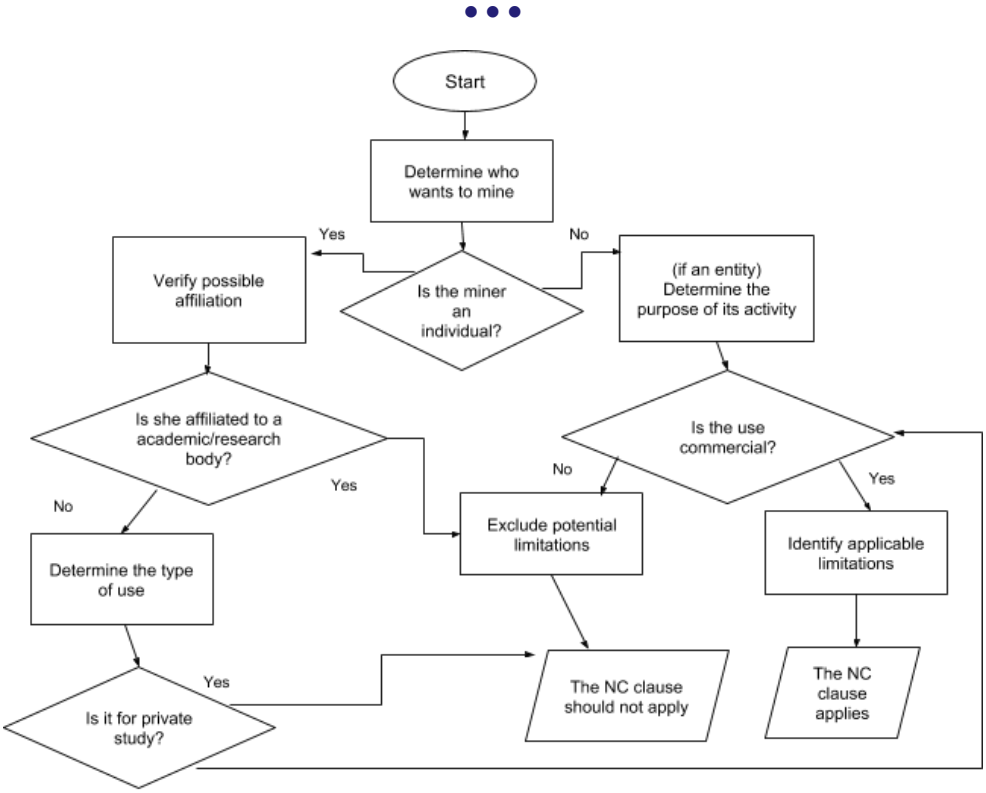

Figure 2: Does the "non-commercial use only" clause apply?

7.2 WG3 Legal FAQs v1.0

Started as a task (T3.3) for WP3 (to add to KB: https://www.fosteropenscience.eu/openminted/text-and-data-mining), the list includes legal FAQs with broader scope - not limited to the use of OMTD platform.

- Who can legally conduct text and data mining on contents protected by copyright or database rights?
    - Anyone that has the right to perform actions for which authorisation is required. Often mere access is not enough as TDM requires to make copies of the resource to analyse. These copies, very often although not always, require the authorisation of the right holder. Authorisation means either an exception or limitation to copyright (such as in the UK the TDM exception for non-commercial purposes (UK CDPA 1988, s. 29A[1]), or in the US fair use (17 U.S. Code § 107[2]). In continental EU, due to the limited availability of exceptions the situation is more uncertain, thus it is always good to see whether the right holder has granted any specific permission to TDM such as by way of a licence (e.g. a CCPL-BY 4.0).
- What restrictions if any apply to open access contents?
    - Open Access has a clear definition based on agreed international standards (Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, Bethesda

[1] http://www.legislation.gov.uk/ukpga/1988/48/section/29A
[2] https://www.law.cornell.edu/uscode/text/17/107

Statement on Open Access Publishing, The Budapest Open Access Initiative). Nevertheless, the term OA refers to a model of share knowledge, it is not a licence in itself. Consequently, it is important to look at the specific licence that govern the resource or content. In principle licence that do not allow commercial uses or the creation of derivative works do not comply with the Open Access definition. However, often, due to lack of specific knowledge (or some other times with the intent to confuse further users), certain resources are contradictory distributed as OA but with a NC licence. This is a very bad practice and should be deprecated. Nevertheless, when this happens, the most prudent approach is to follow what in doubt would be the less risky avenue, and thus the licence (even if it imposes an NC) instead of a general OA declaration.

- When is a valid license required to perform TDM activities?
    - If the mining activity implies the exercise of a copyright or database right, e.g. the right of reproduction or the right of distribution, a permission from the right holder is needed and therefore the requirement for a license exerts, unless an exception is available.
    - On the contrary, if the mining does not imply the exercise of an exclusive right neither relying on a license or an exception is required.
    - This is due to the above described uncertain legal landscape. In principle "the right to read is (or should be) the right to mine". Nevertheless, due to an unclear legal framework that stifle innovation this cannot be said to be yet the case.
- Are there other legal documents that should be considered apart from licenses?
    - Yes. Some resources are available only as services and may not have a license attached, but yet refer to conditions that are specified in the Terms of Service, alias Terms of Use or Terms and conditions (often abbreviated as ToS or ToU). Such terms regulate your use of a service even when the content that you want to mine with that service comes under a different licence. In both cases, it is highly recommended to read them carefully. OpenMinTeD is working on tools that will hopefully make this work easier.
- Shall the licence expressly confer the right to text and data mining?
    - TDM is not usually defined in most legal systems (although it is in a few). If the use implied in the mining activity requires triggers one of the rights seen above (e.g. reproduction, distribution, modification), the licence needs to explicitly grant these permissions to the users. However, unfortunately not all licences terms are perfectly clear and need to be carefully read to determine whether the mining activity is allowed and under what conditions. The main issues here, especially for EU based researchers, is that OA licences have to include in their scope the Sui Generis Database Right. For examples, while CCPLv4 does that, prior versions of CC licences are not so clear on this point and may be problematic[1].
- What restrictions if any apply to CC-licensed contents that one wishes to mine?
    - If the mining activity is undertaken for commercial purposes, resources licensed under CC Non-Commercial (NC) licenses (CC-BY-NC, CC-BY-NC-SA, CC-BY-NC-ND) cannot be mined, while there are no restriction of this kind for resources that are licensed under the other CC licenses that do not have the NC clause.

[1] https://wiki.creativecommons.org/wiki/Version_4

- In addition, if such outputs qualify as adapted or derivative works, resources licensed under a CC no-derivative licenses cannot be redistributed, but can be mined (copied) (CC-BY-ND, CC-BY-NC-ND). This latter restriction only applies in case of sharing the output in question, while making and privately using adaptations is allowed. Also, when sharing the adaptation of resources licensed under Share-alike licenses (CC-BY-SA, CC-BY-NC-SA), the same license must be applied to the adapted output.

- Who can benefit from a TDM exception?
  - If a TDM exception applies, the beneficiaries are usually indicated in the exception clause.
  - In the example of the UK, under section 29A of the Copyright, Designs and Patents Act 1988 (CDPA), a person who has lawful access to the work can benefit from the exception. However, the explicit limitation to the sole purpose of research for a non-commercial purpose clearly narrows down the list of potential beneficiaries.
  - The French exception, embodied in article 38 of the French Digital Republic law (2016-1321 of 7 October 2016) similarly refers to those who have lawful access to the work and likewise it explicitly limits the purpose to non-commercial research thus circumscribe the number of beneficiaries. In addition, the exemption covers only the mining of resources included in or associated with a scientific publication for the needs of research, which further narrow its scope compared to the UK case. However, it is worth considering that the succeeding decree necessary to make article 38 fully applicable has been recently rejected by the French *Conseil d'Etat*, which at present makes the provision in question not in force.
- If the TDM exception applies in a given jurisdiction, e.g. the UK, can other non-UK based researchers collaborating with UK colleagues benefit from the same exception?
  - No. In the example of the UK, only UK-based researchers benefit from the exception. However, if their research partners are affiliated to a UK institution, the text and data mining may be only performed by their UK-based colleagues. If no affiliation exists, the copyright law of their own jurisdiction applies (JISC 2016[1]).
- Is the acknowledgement of the original source always necessary when performing TDM by benefiting from an applicable exception?
  - It may, depending on the exception clause itself. In the example of the UK, the law requires sufficient acknowledgement of original sources, unless impossible, e.g. in case of a large amount of resources with a multitude of contributors.
  - With regard to the French exception, a succeeding decree will clarify the conditions under which TDM should be conducted and its outputs be eventually conveyed. Nonetheless, the attribution requirement can be already easily foreseen as a direct consequence of a strong moral rights regime.
- Can the original resource (or its copy) and the output of the mining activity be shared or distributed?
  - It depends on whether the licensing terms, or the exception if applicable, allow it or not.
  - In the example of the UK exception, the transfer of the copy to any other person, except where the transfer is authorised by the copyright owner, is not allowed.

[1] https://www.jisc.ac.uk/guides/text-and-data-mining-copyright-exception

- Regarding the French example, the conditions under which the mining outputs may be eventually shared will be determined by a future decree.
- Is TDM for commercial purposes allowed?
  - It depends on whether the licensing terms, or the exception if applicable, allow it or not.
  - Both the UK and the French exceptions unequivocally exclude commercial exploitation.
- Are restrictions on the mining activity by means of technological protection measures (TPMs) allowed when a TDM exception applies?
  - Right holders may impose technical barriers such as TPMs to limit or restrict access to their resources, but these should not prevent TDM as such. Under UK law, in particular, section 296ZE of the CDPA 1988 provides a specific remedy where effective technological measures prevent permitted acts, allowing to issue a notice of complaint to the Secretary of State. In any case, circumvention of such TPMs is never allowed.
- Can the mining activity be restricted by contractual provisions when a TDM exception applies?
  - If any restrictions imposed by right holders by contract result in preventing miners from benefiting from the exception, such contractual terms may be deemed unenforceable, as it is under UK law.

7.3 OpenMinTeD Interoperability Webinar Series Fall 2016

This section contains an excerpt of the internal milestone report for the OpenMinTeD Interoperability Webinar Series Fall 2016 (MS 22) which is relevant to the progress report sections of the interoperability working groups above.

7.3.1 Webinar 1: Achieving interoperability between resources involved in TDM at the level of metadata

7.3.1.1 Abstract

Resources involved in TDM processes include content resources to be mined, TDM tools and services operating on them as well as the reference/ancillary knowledge resources the latter utilize at the time of processing. OpenMinTeD sets out to create an open, service-oriented e-Infrastructure for TDM of scientific and scholarly content. To this end, the OMTD-SHARE metadata schema serves as a facilitator, providing the interoperability bridge between the various resource types, and as an intermediary with the users (TDM developers and end-users) guiding them to locate the resources that best answer their research needs and that can be put together in a TDM workflow that can operate seamlessly. The webinar includes the presentation of the schema, main principles and strategies selected to achieve interoperability, focusing on specific exemplary points, as well as issues that are raised when populating the OpenMinTeD registry with resource metadata from various sources.

7.3.1.2 Topics discussed

Open controlled vocabularies

The first issue raised concerned the handling of "open controlled vocabularies", i.e. vocabularies which are used as recommended values for filling in specific metadata elements and which are continuously updated with new entries; the IANA list of mime-types constitutes such an example. At present, such vocabularies are implemented as enumerations in the XSD, which supports the validation of metadata

records at any stage but creates the need for continuous update of the XSD. The suggestion was made to include these enumerations at the level of the registry, postponing the validation of these elements at the entry of the metadata records into the registry; the benefits of this approach are the stability of the XSD as compared to a better management of values in the registry, but causes a mismatch between metadata records in and out of the registry.

Content from publishers

The discussion around the import of content from publishers into the OpenMinTeD platform and the ways of promoting interoperability among them focused on two points:

- Publishers will be able to register their content through OpenAIRE and CORE and are, thus, advised to follow the respective guidelines, in order to ensure that they are also compatible with OpenMinTeD requirements. At the same time, OpenAIRE and CORE guidelines must endorse recommendations issued by OpenMinTeD for scholarly publications (e.g. mandatory license element encoded according to guidelines, normalized link to the article text, etc.)
- A connector to publishers' system is currently built and will be integrated in the OpenMinTeD platform, which will allow accessing and harvesting open access content from them. The connector takes into account the particularities of each publisher's system, given that interoperability is a desideratum but not yet accomplished, with specific modules built for each system. This approach allows OpenMinTeD to actively get the content from the publishers and harmonize it inside the platform.

Outreach and collaboration with other communities

One of the suggestions made concerned the interaction with groups such as Force11 and other scholarly communications communities, in order to promote recommendations and collaborate on standards.

Unique and persistent identifiers

In relation to the issue of identifiers, the discussion points were the following:

- How and by whom identifiers should be assigned: the advantages and disadvantages of relying on best practices, such as the one used by the FOSS community while registering products in the Maven repository vs. the assignment of unique ID's (DOI's) by a central authority (CrossRef) as in the case of publications have been debated;
- How to ensure uniqueness and persistence: the point on how unique Handle Persistent Identifiers (Handle PIDs) really are, since they are assigned by non-central authorities, thus leading to the same resource receiving new identifiers each time it is registered at a new repository, was raised; the assignment of DOIs also doesn't guarantee persistence, as they are "persistable" but they are not de facto persistent
- Need for such identifiers: all agreed that unique PIDs are needed, especially for attribution and citation purposes as well as for ensuring persistence and enabling repetition of experiments;
- New requirements for identifiers: new types of resources, such as virtual collections (dynamically evolving data collections) and workflows pose new requirements for the assignment of identifiers which need to be considered

- Ways of assigning identifiers: the hashing approach (also presented in one of the papers at the LREC Interoperability workshop) was discussed as one that seems a promising solution also for persistence and beneficial for virtual collections. Further investigations into this approach should be followed.

Subject/Domain classification

The need for making subject/discipline/domain classification as a mandatory feature in the description of scholarly publications was pointed out, supported by the fact that this is the first criterion used to select resources for TDM. The discussion that followed brought about the problems in its encoding:

- Multiple authority lists used in the metadata descriptions of resources;
- Resource providers are reluctant to encode it and, even when they do, inconsistencies are observed;
- Automatic topic classification is a promising solution, but it mainly centers around clustering especially at a lower level, while an upper level of concept classification is also needed;
- Classification taxonomies are changing over time which also affects the quality of encoding of the feature;
- The need of using various features, including the content itself, in classification.

Implementation of the schema in RDF

A question about the choice of XSD vs. RDF for the implementation of the metadata schema gave the occasion to mention the intended use of RDF/OWL for the next phase of the project. The use of SKOS for linking specific elements with elements from related metadata schemas and inside the schema is also in the following plans of the metadata team.

Video

- https://www.fosteropenscience.eu/content/achieving-interoperability-between-resources-involved-tdm-level-metadata

Indicative pointers to related activities and work

This list is an indicative and non-exhaustive list of the activities and resources mentioned during the workshop.

- OpenAIRE guidelines: https://guidelines.openaire.eu/en/latest/
- CORE: https://core.ac.uk/join and https://core.ac.uk/recommendations
- Pontika, N., Knoth, P., Cancellieri, M. and Pearce, S. (2016) Developing Infrastructure to Support Closer Collaboration of Aggregators with Open Repositories, LIBER Quarterly, 25, 4. https://doi.org/10.18352/lq.10138
- Force11: https://www.force11.org/
- Maven repository for identifiers: https://maven.apache.org/pom.html#Maven_Coordinates
- CrossRef: http://www.crossref.org/
- Handle PIDs: http://www.handle.net/
- McCrae, John P. and Bordea, Georgeta and Buitelaar, Paul, "Linked Data and Text Mining as an Enabler for Reproducible Research". *Proceedings of the Workshop on Cross-Platform Text*

Mining and Natural Language Processing Interoperability (INTEROP 2016) at *LREC 2016*, Portoroz, Slovenia. http://interop2016.github.io/pdf/INTEROP-7.pdf

7.3.2 Webinar 2: A minimalist approach to workflow interoperability

7.3.2.1 Abstract

In this webinar, we introduce OpenMinTeD Script, a minimalistic scripting environment for cross-framework workflows. Bringing together text and data mining tools and services from all over Europe (and beyond) and integrating them in the OpenMinTeD platform to analyze scholarly publications is our designated goal. We recognize existing efforts of making TDM tools and services interoperable, such as e.g. Argo, DKPro Core, GATE, or LAPPS Grid. Thus, instead of integrating individual tools and services, we focus on integrating these already existing sets of components and thus address interoperability at the collection/framework level. Using OpenMinTeD Script as a minimal integration layer, we investigate and improve cross-framework component configuration, component life-cycle, data conversion, and workflow deployment. We report on progress and experiences gained so far and provide a lookout for the upcoming challenges.

7.3.2.2 Topics discussed

Data representation, type systems, and mapping

Part of the discussion revolved around the question of defining a new type system, re-using existing type systems, and user-defined custom type systems. In order to avoid reinventing the wheel and in the light that existing type systems may each specialize in certain areas/domains, it was brought up that the type system to be used by/endorsed by OpenMinTeD could take type definitions from other sources, e.g. existing UIMA type systems, LAPPS Vocabulary, ontologies, etc.

The distinction between schema and meta-model mapping was made during the presentation. The mapping discussion focused largely at the level of schema mapping. Although the Argo Type Mapping component was found to be very flexible, potential problems with recursive structures like syntactic dependency relations and constituency parse trees were discussed.

At the level of meta-model mapping, e.g. differences between the UIMA meta-model and the GATE meta-model were discussed, specifically that the GATE model presently lacks a format concept to represent relations between annotations. It was suggested to extend the type mapping facility such that it can be applied in the same manner to GATE and to UIMA annotations, using the same mapping rule language in both cases.

Different representations of annotations and knowledge were mentioned, e.g. NIF, GATE XML, XMI, RDF or JSON-LD. It was discussed whether these could be alternatives to more text-oriented annotation representations like GATE XML and UIMA XMI. E.g. the NIF specification for representing NLP illustrates the high level of verbosity that needs to be introduced into the RDF representation to allow reasoning over text. E.g. while GATE and UIMA frameworks have built-in support to reason over annotations in terms of spans and offsets, such capabilities are not available in SPARQL. Hence, NIF e.g. needs to explicitly link each token to the previous and following tokens as well as linking tokens and sentences to each other.

Building, deploying and running workflows

It was suggested that immediate issues to be addressed should be primarily packaging, installation and deployment and that type mapping may be a secondary issue.
In this sense, OpenMinTeD Script provides a minimal scripting language for connecting components from one or more source in a very simple environment where interoperability issues can be investigated independently from the rest of OpenMinTeD. In particular, the workflow editor/engine which was not finally selected in the project until very recently. OpenMinTeD Script has been already used up to this point to investigate interoperability issues and build example workflows.
OpenMinTeD Script is implemented as a Groovy-based Domain Specific Language (DSL). The language supports primitives such as *"read"*, *"write"*, and *"apply"*. Type mapping is currently realized through the Argo Type Mapper component that can be added between incompatible components in a workflow. Deeper integration of type mapping into the DSL is being contemplated.
Presently, OpenMinTeD Script supports mainly JVM and Java related: Groovy, Scala, Clojure, Jython, JRuby. However, some initial work has been done in order to integrate components in arbitrary languages through an adapter for remote service invocation, e.g. to invoke ILSP or LAPPS Grid services. Support for dockerized components is also being investigated.
The architecture of OpenMinTeD Script will not change much when adding new adapters or data converters. These are expected to be eventually transferrable from the OpenMinTeD Script interoperability exploration sandbox into the OpenMinTeD production environment.
The question was brought about whether OpenMinTed would provide a visual workflow editor. It was mentioned that OpenMinTeD is looking into choosing such an editor, however, at the time of the webinar, the final decision had not been announced yet. Shortly after the webinar, it was announced that OpenMinTeD would be using the Galaxy workflow editor.

Video

- https://www.fosteropenscience.eu/content/minimalist-approach-workflow-interoperability

Indicative pointers to related activities and work

This list is an indicative and non-exhaustive list of the activities and resources mentioned during the workshop.

- Review third-party frameworks like CLARIN and NLP web services which are not covered by GATE or UIMA components collections
- Analyze the possibility of adding a visual interactive editor for workflows like ARGO and Galaxy
- Study the implementation of a hybrid type system with type definitions from different sources
- Mapping method: still some problems with expressiveness, reutilization and overlaps some solutions provided
- Study the application of semantic technologies
- Provide examples, mappings and use-cases with OpenMinTeD scripting
- Consolidate all above points

7.3.3 Webinar 3: Text and Data Mining interoperability at the legal level: rights, exceptions and licenses

7.3.3.1 Abstract

The webinar will focus on the complex and fragmented EU copyright framework which applies to activities relevant for Text and Data Mining purposes. The general legal landscape will be briefly presented in order to identify limits and opportunities offered by current copyright rules. Regarding the former, the webinar will illustrate which rights (e.g. right of reproduction and right of distribution) can be triggered by TDM activities and what this entails. Regarding the latter, available exceptions and limitations will be analyzed in an attempt to offer an overview of when (and where) an existing copyright exception could cover TDM activities. This part will include the recent draft proposal for a Directive on Copyright in the Digital Single Market and other national initiatives. Finally, copyright licenses and the ways in which the OpenMinTeD project intends to favor legal and metadata interoperability among the many different and often incompatible licenses and terms of use will be discussed.

7.3.3.2 Topics discussed

Risks related to performing TDM activities: copyright infringement

Addressing the first question posed by the audience, with regard to the likelihood of infringing conducts when carrying out text and data mining activities, the discussion focused on whether any text and data mining activity is indeed likely to infringe copyright. To such an extent, both legal and technical aspects of TDM were taken into consideration, where the latter included determining whether the problem is that a copy is made (therefore the right of reproduction is infringed) or that, whatever the method used (for instance by streaming) text and data mining would still be unlawful. As the presenter highlighted, in absence of an applicable copyright exception, this is regrettably the case. Similar conclusions applied when it was considered the instance in which the *sui generis* right for the protection of database was infringed. Therefore, it appeared particularly urging to advocate for a change in the law, especially in terms of supporting a broader exception that would allow TDM in the first place.

Economic aspects of licensing for TDM

Prompted by another question from the audience, participants discussed the uncertainty and unclarity surrounding licensing agreements, including the likelihood of "paying double" when asking a specific permission to perform TDM whereas a license may already grant that right, but the licensee is not aware of. The presenter emphasized how this uncertain and unclear setting is essentially due to the inherent complexity of licensing terms and, as participants confirmed, to the lack of legal training among miners. Keeping up with the economic aspects that surround TDM, discussants also explored the implications of allowing text and data mining for commercial purposes. Acknowledging all these matters, the illustrated WG3 work plan with its related working documents, i.e. the compatibility matrix, the flowchart and the legal annotation experiment, appeared to be addressing the issue properly. Yet, springing from the participants' comments, the most licenses these documents cover, the most successful would be the attempt to make text and data mining easier and law-compliant.

Potential interaction of the TDM exception with other copyright exceptions

In response to participants' questioning with regards to whether there might be an interaction between a text and data mining exception and other existing copyright exceptions, it was specifically considered the applicability of the right to quote. In any case, it was clarified that invoking the quoting exception cannot rescind from the conditions imposed by the exception itself, for instance with respect to the quotation amount and the need for acknowledgement. Indeed, apart from any parallelism, the limits of such correlation were also critically observed, particularly when data are concerned. Therefore, it was concluded that despite the closeness of the two exceptions, it is essential to bear in mind each one's conditions and limitations.

The proposed EU text and data mining exception

By the end of the discussion, the European Commission's proposal for a Directive on copyright in the Digital Single Market of 14 September 2016, was more than once mentioned and the pros and cons of the proposed TDM exception in the European Union were considered. In particular, the mandatory feature of this exception across all Member States was highly welcomed by discussants, while its very narrow scope (being, until the contrary, research organizations the sole beneficiaries) appeared to be more questionable. The recent French TDM exception was also mentioned, being its scope also particularly narrow (both in terms of purpose and subject matter). Participants concurred that the EU exception should indeed take a step forward and go beyond the limits so far portrayed. Hence, the issue remains open to further discussion.

Video

- https://www.fosteropenscience.eu/content/text-and-data-mining-interoperability-legal-level-rights-exceptions-and-licences

Indicative pointers to related activities and work

This list is an indicative and non-exhaustive list of the activities and resources mentioned during the workshop.

- Why We Need a Text and Data Mining Exception (But it is Not Enough). Proceedings of the LREC 2016 Workshop "Cross-Platform Text Mining and Natural Language Processing Interoperability". 23 May 2016 – Portorož, Slovenia, at 57, http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-INTEROP_Proceedings.pdf
- Legal Interoperability Issues in the Framework of the OpenMinTeD Project: A Methodological Overview. Proceedings of the LREC 2016 Workshop "Cross-Platform Text Mining and Natural Language Processing Interoperability". 23 May 2016 – Portorož, Slovenia, at 60, http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-INTEROP_Proceedings.pdf
- Envisaging a broader TDM exception to overcome the pitfalls of current copyright law in the EU. OpenMinTeD Blog, July 8 2016, http://openminted.eu/envisaging-broader-eu-tdm-exception-overcome-pitfalls-current-copyright-law/
- OpenMinTeD - How to take the next practical step in TDM. OpenAIRE Newsletter, October 3 2016, https://www.openaire.eu/OpenMinTeD-2
- OpenMinTeD interoperability webinar. CREATe Blog, November 21 2016, http://www.create.ac.uk/blog/2016/11/21/OpenMinTeD-interoperability-webinar/

7.3.4 Webinar 4: Text mining interoperability at the knowledge level

7.3.4.1 Abstract

Different TDM components may define the same linguistic or domain concept by reference to different linguistic, terminological and ontological resources. For example, two components may define part-of-speech by reference to two different knowledge resources. This leads to difficulty in interoperability when mixing such incompatible components in the same workflow. OpenMinTeD is tackling this knowledge level interoperability through the selection and mapping of widely used and de facto knowledge resources, and by creating guidelines for other such mappings. We will present the current state of our selection and mapping, and discuss progress with interoperability experiments that make use of these mappings in real workflows.

7.3.4.2 Topics discussed

Unique IDs at the document level

The use of the Web Annotation standard as a representation of annotations, for interoperability with systems external to OpenMinTeD, was discussed. Such systems were described as being “end-to-end”, as user is concerned with the system as a whole, and not with the internal components of that system.

As a linked data based standard, Web Annotation requires unique identifiers to be assigned for each element of an annotation, e.g. the annotation as a whole, and individual components of the annotation. The issue of who should assign these identifiers was discussed, in particular with respect to some hypothetical OpenMinTeD compliant service provider that may not have the facilities to issue unique identifiers. It was proposed that in this case, the service provider would only need to ensure that an identifier was unique in the context of the annotation document itself. It was also proposed that the OpenMinTeD platform could take such an identifier and represent it in a globally unique way, e.g. by combining it with some other identifier.

The Web Annotation standard also requires the target of an annotation to be provided as a link within the annotation description document. This is a reasonable expectation for documents that have been retrieved from repositories that provide unique identifiers, but needs some thought for cases where an end user is providing unidentified text, e.g. a file stored on their local machine.

A related question was raised, on how we might identify a fragment of text within a document - for example, where an annotation spans specific locations in the text. The fact that Web Annotation has a representation for document fragments was briefly discussed.

Storing annotations in OpenMinTeD

Should annotations be stored in some central repository, or be distributed, and how does this depend on the workflow? Once annotations are created by some service, we have to consider their storage. Annotations may be stored in the OpenMinTeD registry, with storage governed by some policy which will answer questions such as how long data will be retained. If we are to use the registry in this way, we need to consider if the registry should be for the storage of temporary data, as may sometimes be the case with annotations.

Interchange formats: XMI and the Web Annotation standard

In addition to discussing the use of the Web Annotation standard as an output format, the webinar discussed the use of UIMA XMI as an interchange format, and as an alternative output format. At some would we not need to provide some linking or interchanging between the two representations?
In answer to this, it was discussed how Web annotations are intended more as an output format for end-to-end systems, as opposed to a representation for interoperation between internal components of systems. Internally, we intend to use XMI - many of the systems being integrated already support this, or support representations close to this. However, this is not necessary for the end users of end-to-end systems, and the final output in these cases will be Web Annotation.

Using ontologies or external knowledge resources in the simple cases

As a linked data standard, Web Annotation provides type information by referencing external namespaces. For example, annotations created by a service for some domain, may be defined in terms of an ontology that is already in use in that domain. It was discussed how OpenMinTeD may make recommendations for domains in which it already operates, but that there will always be cases where a service provider is providing some service that falls outside of this, perhaps some simple service, or some service from a domain with no developed type system. In these cases, we need to give users the capability to provide a type system of their own.

Balance between custom ontologies and recommended ontologies

Users may not be using some standard ontology for various reasons. For example, there be no such work in their domain; they may not be aware of standards work in their domain (and OpenMinTeD may not be aware of this either); there may be a variety of competing standards in their domain. We should not exclude users because they are not using some standard ontology, or because they are using their own. However - we should be encouraging them to use standard ontologies, and we should be pointing out to them that while they do not adopt semantic standards, they are only achieving syntactic interoperability.

Interoperability at the syntactic level

Following on from the above point, the webinar raised the question of how much useful interaction and interoperability could the OpenMinTeD platform achieve at the syntactic level alone. Examples of where there was benefit from the syntactic level were provided: visualization of results of text data mining output but without knowing anything about the semantic level; tools operating at the syntactic level for sorting, counting, and aggregation at the level of undescribed types and keys and values. Clearly the more semantic level is available, the more possibilities there are for extracting information. We should therefore encourage people in various directions to look at standards for representing their requirements. However, even if people within a particular domain have not got full rich conceptualizations should still be able to integrate their tools in OpenMinTeD.

Video

- https://www.fosteropenscience.eu/content/text-mining-interoperability-knowledge-level

Indicative pointers to related activities and work

This list is an indicative and non-exhaustive list of the activities and resources mentioned during the workshop.

- UIMA XMI: http://www.omg.org/spec/XMI/
- Web Annotation: https://www.w3.org/TR/annotation-model/
- Lemon: http://lemon-model.net
- LAPPS Exchange Vocabulary: http://vocab.lappsgrid.org/
- OpenAIRE mining examples: http://mining.openaire.eu
- Example systems:
    - Lifewatch: https://lm.portal.lifewatchgreece.eu/extract
    - Bio-YODIE: https://cloud.gate.ac.uk
    - Europe PMC: https://europepmc.org/Annotations
    - Archaeology data service: http://ads.ac.uk/nlp/demo.jsf
    - Funding ontology used in examples: http://vocab.ox.ac.uk/projectfunding