OpenMinTeD: A Platform Facilitating Text Mining of Scholarly Content

Penny Labropoulou[♣], Dimitrios Galanis[♣], Antonis Lempesis[♣], Mark A. Greenwood[♠], Petr Knoth[♢], Richard Eckart de Castilho[♡], Stavros Sachtouris[‡], Byron Georgantopoulos[‡], Lucas Anastasiou[♢], Stefania Martziou[♣], Katerina Gkirtzou[♣], Natalia Manola[♣], Stelios Piperidis[♣]

[♣]Athena Research and Innovation Center in Information, Communication and Knowledge Technologies
[♠]The University of Sheffield, Sheffield, UK
[♢]Knowledge Media institute, The Open University, UK
[♡]Technische Universität Darmstadt, Department of Computer Science, UKP Lab
[‡]Greek Research and Technology Network (GRNET)

Abstract

The OpenMinTeD platform aims to bring full text Open Access scholarly content from a wide range of providers together with Text and Data Mining (TDM) tools from various Natural Language Processing frameworks and TDM developers in an integrated environment. In this way, it supports users who want to mine scientific literature with easy access to relevant content and allows running scalable TDM workflows in the cloud.

Keywords: text mining, open access, corpora, natural language processing, scholarly content

1. Introduction

Recent years have been witness to a huge upsurge in the quantities of digital research data and scientific literature being produced[1], offering new insights and opportunities for improved understanding. Researchers, however, find it hard, if not impossible, to keep up with the growing volume of relevant material. In this context, Text and Data Mining (TDM), "*the discovery by computer of new, previously unknown information, by automatically extracting and relating information from different (. . .) resources, to reveal otherwise hidden meanings*" (Hearst, 1999), has emerged as an indispensable tool for harnessing the power of structured and unstructured content and data. By analysing content and data at multiple levels and in several dimensions, TDM discovers hidden and new knowledge that researchers can further explore.

TDM is, however, faced with its own challenges. Firstly, access to the scientific literature is made difficult due to relevant works often being scattered over many sources (institutional/thematic repositories, publishers' sites, libraries etc.). In addition, running TDM workflows requires a significant amount of *plumbing* due to the lack of interoperability between widely used TDM tools.

OpenMinTeD (OMTD) aspires to become an infrastructure that fosters and facilitates the use of TDM technologies over scientific publications and beyond. More specifically, it aims to bring together, on the same platform, (a) Open Access (OA) content from a wide range of data sources and (b) TDM tools from existing Natural Language Processing (NLP) frameworks and TDM developers.

This combination allows us to serve two different communities. Firstly, it acts as a portal through which **text mining experts** can offer their components and applications to a wider audience as well as building new applications with components from various sources. Secondly it supports **users with no or little prior text mining experience** to run ready-to-use workflows on scholarly publications in order to mine information which they would not otherwise be able to access.

The remainder of the paper introduces the OMTD platform. Specifically, Section 2. provides an overview of its functionalities and Section 3. goes into the technical details (architecture and main modules). The metadata schema is presented in Section 4.; Section 5. provides information on the distribution and documentation of the OMTD software. Finally, Section 6. puts the platform in the wider landscape of similar platforms, and we conclude with a summary of the major achievements.

2. Overview of the OMTD platform

The OMTD platform (Figure 1) acts as a facilitator of TDM focusing on scholarly content. TDM involves a wide range of resources and OMTD tries to bring them together[2] all in one place, make them interoperable, and offer them to its end-users ready to be deployed:

- **content resources** to be mined, i.e. research papers published in conference proceedings, academic journals, etc.;
- **TDM software**, which in OMTD is split into "*components*", i.e. pieces of software that perform basic tasks (e.g. sentence splitting, tokenization, part-of-speech tagging) and can be combined to build workflows, and "*applications*", which are ready to be used by end-users

[1] According to the STM Report 2015, the global research community generates ∼2.5 million new scholarly articles per year in English only.

[2] It should be noted that resources must be registered in OMTD only if they can be accessed and deployed in the context of a TDM processing operation in the OMTD framework. It's not the intention of OMTD to function as a catalogue of information *about* TDM resources.

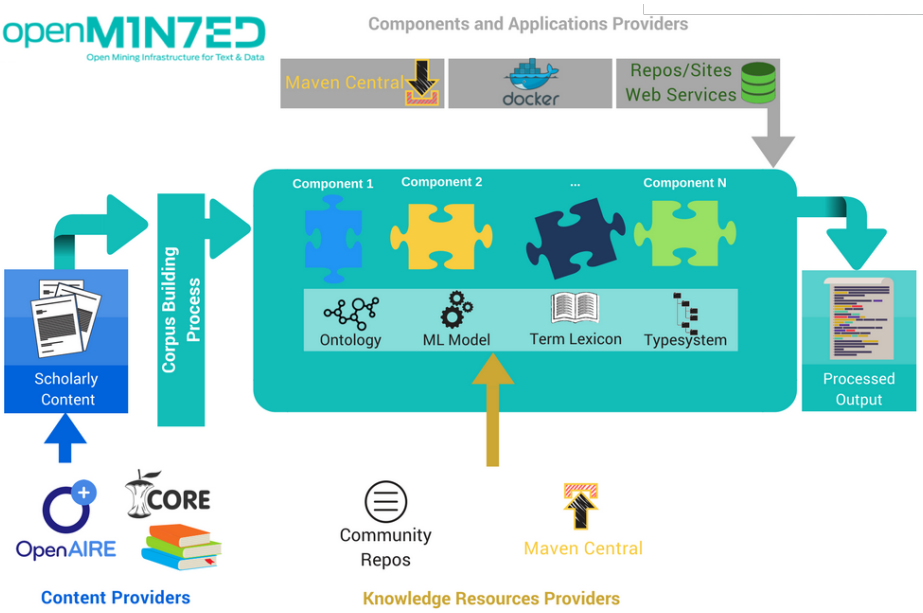


Figure 1: The OpenMinTeD platform

(e.g. for named entity recognition, citations extraction, extraction of relations between scientific entities, etc.)[3];

- **ancillary knowledge resources** used for the operation of the software (e.g. Machine Learning models, lexica, terminologies or ontologies used for annotating the resources to be mined, typesystems, annotation schemas, linguistic tagsets, etc.).

These resources are available from a lot of different types of providers: scientific articles are mainly found at publishers' sites, journal sites, libraries, or academic, university, and institutional repositories, etc., while software and ancillary resources are usually hosted at thematic/institutional/community portals, dedicated software repositories, etc. As a consequence, they abide by different technical specifications. For instance, scientific papers are distributed in different formats: text, PDF, Word documents, XML files complying to different schemas, etc.; even when the same standard is used, there can be different implementations or styles associated with them (e.g. the PMC style in the case of the JATS or NLP Journal Publishing Tag Suite[4]). Moreover, all these resources are documented in different ways and at varying levels of granularity.
Thus, in order to achieve its goals, OMTD must first of all attain (a) technical ***interoperability*** between resources that need to be combined, and (b) ***harmonization*** between resource metadata descriptions of the same type. The first point affects mainly the combination of components coming from different frameworks as well as the combination of the input content with the processing components/applications. For instance, if a research article is in a PDF format, the TDM application that will be used for its processing must be able to handle PDF files. To ensure interoperability between resources of the same type as well as across them, OMTD has set up a set of procedures, protocols, and *technical specifications* that must be followed by providers when registering their resources in the platform.[5] To avoid a proliferation of standards, OMTD has focused on re-using existing standards and technologies where possible and to fill interoperability gaps where necessary. Regarding the second point, OMTD's main contribution was the design and implementation of the OMTD-SHARE metadata schema (presented in more detail in Section 4.).
The OMTD platform supports the **registration** and **storage of the resources and their descriptions** in a multitude of ways in order to accommodate the different needs and practices of providers and the peculiarities of each resource type. Thus, scientific articles are mainly imported into OMTD in large volumes through content providers (cf. Section 3.2.) while runnable software is added by registered individuals, a procedure devised in order to ensure security and compliance with the technical specifications (cf. Section 3.3.).

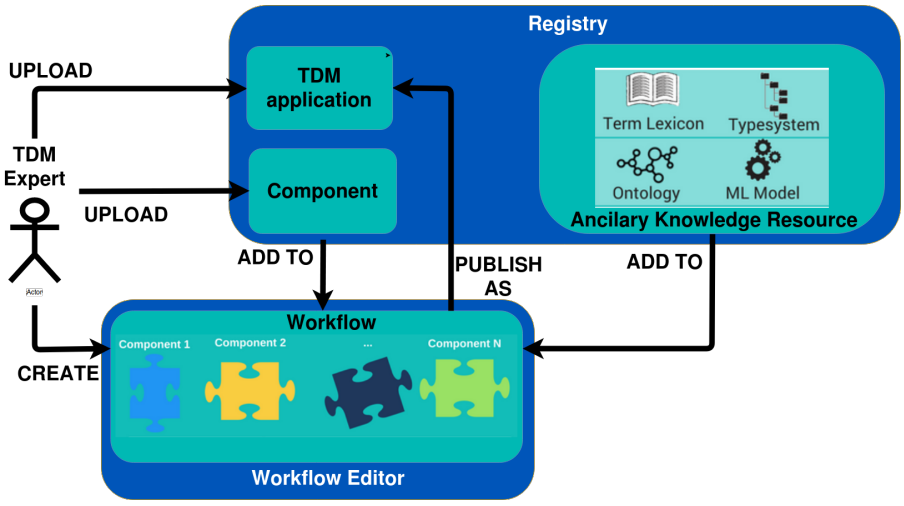
The main services offered to the end-users by the platform are the **creation** and **execution of runnable workflows** cf. Section 3.3.). Given the degree of their complexity, they are addressing two different types of users (Figure 2):

- **knowledgeable/expert TDM users** can (a) publish their software components or end-user applications to the OMTD platform, (b) mix and match components (even with components from other providers) into workflows, (c) parametrize them with the appropriate ancillary knowledge resources required for their operation, (d) test them with real content, (e) adapt them to new domains and tasks, and (f) share the newly created end-user applications with other OMTD users.

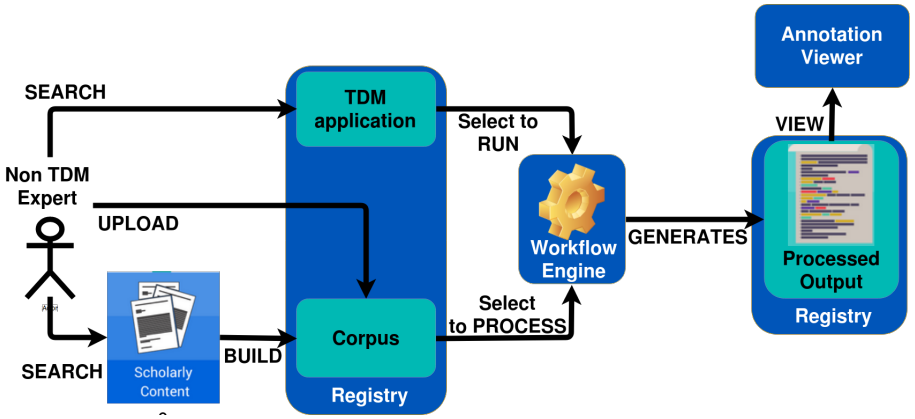
[3] Applications are usually workflows but software providers may opt to also upload one-step applications.

[4] Cf. https://www.ncbi.nlm.nih.gov/pmc/pub/filespec-xml/, note of requirement 2.

[5] https://guidelines.openminted.eu/



(a) Knowledgeable/Expert TDM users



(b) Users with no or little technical knowledge on TDM

Figure 2: Using the OMTD platform

- **researchers with little or no TDM experience** can (a) search for ready-to-use TDM applications, (b) select publications from those already registered in OMTD or upload their own corpora, (c) process the selected content with the TDM applications of their choice in a powerful cloud computing environment, (d) view the results, in the case of annotated corpora, and (e) finally download and publish the results of the processing.

3. Architecture and main modules

The OMTD platform is based on a modular and flexible architecture (Figure 3) that allows to easily integrate different implementations of the same module.

The backbone of the platform is the ***enabling layer***, containing the services that are responsible for the management of the platform or providing functionalities used by every other service in the platform, such as the AAI service providing the authentication and authorization functionality, the Cloud service that manages the hardware resources and executes the components composing the workflows, etc.

The ***data layer*** is the back-end of the system and contains the services that manage the TDM resources (content, software and ancillary knoweldge resources). These services are responsible for storing the metadata descriptions of all resources (Metadata registry), storing the actual data of the entities that are stored locally in the platform (Store service), and importing the resources from external sources (Content service and TDM connector).

The ***service layer*** contains the services that provide the OMTD functionalities to the end-users, and, finally, the ***user interface layer*** acts as the gate to the platform for the users. The following subsections present the main OMTD functionalities and the modules that are involved in them.

3.1. User management

Access to the OMTD catalogue of TDM resources for browse and search is open to all individuals. But further use of the platform requires their registration for security and monitoring reasons, as in the case of publishing new resources. In addition, logging into the OMTD platform enables users to enjoy further functionalities such as their personal workspace, where they can store private corpora and TDM components/applications. In this way, they can, for instance, experiment with the creation of new applications until they are satisfied with their performance and decide to make them publicly available.

The ***authentication and authorization (AAI) service*** is achieved through communication with external authenti-

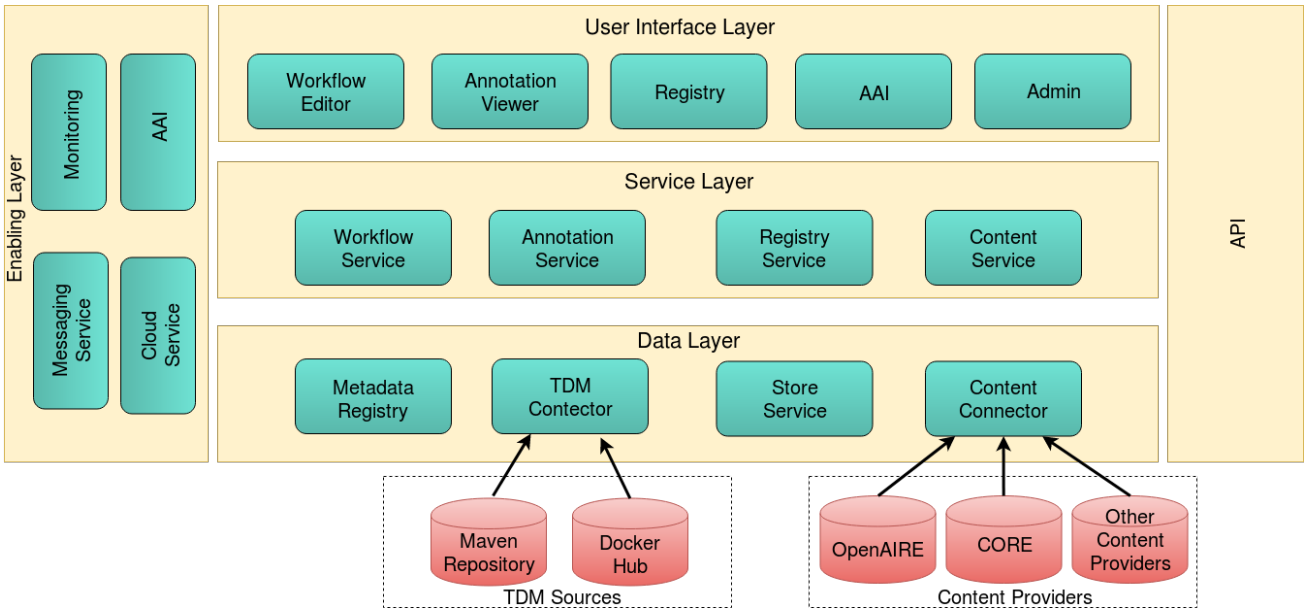


Figure 3: The OpenMinTeD architecture

cation sources and identity providers. This separation between the internal OMTD services and the external authentication sources and identity providers facilitates the service development and removes the requirement from the services themselves to operate their own IdP Discovery Service, a common requirement for services supporting federated access. The AAI service interacts with a number of external authentication providers: (a) the eduGain identity provider[6], that allows users to authenticate using their academic credentials (e.g. universities, research institutes); (b) the ORCID service[7], that allows users involved in research, scholarship, and innovation to authenticate using their ORCID identifier, and (c) various social identity providers, such as Facebook, Google, and LinkedIn, which empowers authentication of users with their social media accounts.

3.2. Content management

The unit of work in OMTD is the "*corpus*" rather than the "*document*". Users in OMTD come to mine large sets of documents rather than finding a specific publication in order to read it. To facilitate this, OMTD has set up a mechanism which provides access to scholarly and scientific content from a wide range of sources and enables users to select among them the ones that interest them for mining.

More specifically, the platform is connected to a content provider via a ***Content Connector***. The content connector implements an interface to the ***Registry***[8] and ensures that data from the provider is supplied with metadata compatible with the OMTD-SHARE schema. Once this happens, the registration of a content provider is completed, and its content (i.e. PDF, TXT, XML, etc. files) and metadata become available in the Registry. Each connector offers the following functionalities:

- performs mapping from the OMTD-SHARE schema to the external provider's schema and the reverse, allowing the connector to return metadata in a common form;
- provides search functionality by using the proprietary search API of the data provider and returning the results in a common format;
- provides access to the full text of the publications, allowing the construction of new corpora with the criteria set by the user query.

Currently, the platform connects to two content providers:

- **OpenAIRE** – a major aggregator with outreach to many OA repositories, archives and journals.[9]
- **CORE** (Knoth and Zdrahal, 2012) – a global aggregator of OA research papers from institutional and subject repositories as well as publishers[10].

The size of the scholarly content that can be accessed via OMTD is already impressive: the combination of scientific papers from CORE and OpenAIRE, mapped to a common data model (OMTD-SHARE) constitutes, to the best of our knowledge, the world's largest OA dataset freely available for text mining. A breakdown of the content is available in Table 1. Yet, more providers are already joining in, as a result of an Open Call for content providers launched during the project.

End-users have access to the OMTD catalogues of resources separated by resource type. Through the catalogue of publications, a user can use the *faceted search facility* and/or the

[6]https://edugain.org/

[7]https://orcid.org

[8]The definition of the interface is available at OMTD GitHub: https://github.com/openminted/content-connector-api

[9]https://www.openaire.eu/

[10]https://core.ac.uk/

[11]If the content from many different publishers can be harvested via a single interface, we treat them as just one data source. For example, this is the case for all DOAJ journals or content from PMC OA subset.

	OpenAIRE	CORE
Full texts	4.1 million	10.3 million
Abstracts	4.1 million	85.6 million
Metadata records	4.1 million	125.7 million
Full texts data size	4.4 TB	48 TB
Data sources	182	3,673 [11]

Table 1: Number of research papers available from OMTD via the corpus connectors

free text query to select a subset of papers from the respective providers and **build a corpus** that they can then submit for processing with the TMD application of their choice. For instance, asking for "*research articles*" in "*English*" published after "*2012*" containing "*proteins*" and "*genes*" in their metadata. The available facets, which include *language*, *access rights*, *publication type* and *publication year*, are powered by information provided by the data sources in OMTD-SHARE.

The resulting corpus is stored in the OMTD Data Storage (***Store service***). This service has the appropriate REST API for managing (creating, saving, downloading) the data.

Users are likely to create corpora with the underlying resources overlapping, as in the case, for instance, of two users who decide to mine content in the domain of biomedical research with similar selection criteria. This would imply the need to store multiple copies of the same documents in the storage. Given the data size of the collections, this would create unnecessarily high requirements on the storage. To avoid this, the Store service makes sure that each underlying document is deposited only once. This is achieved by making use of a hash based on the document's binary file. A corpus is then defined only as a collection of hashes of documents it contains (rather than the documents themselves).

In addition, users have the choice to **upload their own corpora** directly to the Registry as long as they follow the technical requirements set for all input corpora.[12] This enables them to use the platform for processing private corpora without exposing them to other users.

3.3. TDM software management

In the context of OMTD, a TDM application or component must be published in the OMTD ***Registry*** with a metadata description conforming to the OMTD-SHARE schema and information on the access point from which it can be invoked at the time of execution.

In both cases, the TDM provider must follow a specific set of technical specifications and instructions set by OMTD[13] that seek to ensure robustness and executability. While the OMTD-SHARE schema is generic enough to describe a wide range of TDM applications or components, we currently provide support for easily integrating those that belong to one of the following three categories:

- TDM tools built with either GATE (Cunningham et al., 2013) or UIMA (Ferrucci et al., 2009) – two popular NLP frameworks – and published through Maven Central[14];
- TDM tools packaged as Docker images according to the relevant OMTD specifications and published through Docker Hub[15];
- TDM web services conformant with the relevant OMTD specifications.

Once a TDM application has been registered, it is automatically available for use.

On the other hand, registered components must get integrated in an application in order to be executed as part of it. In order to be combined in a functional workflow, components need to exchange data with each other. On the OMTD platform, we recommend the use of the UIMA CAS XMI format[16] for this purpose. The combination of components further requires careful checking that the output of a component is indeed a valid input for another component, in terms of various parameters, such as format and annotation information they carry; for instance, a parser may require as input a part-of-speech tagged corpus using a specific type system. This is a complex task carried out by TDM experts who build workflows (and ultimately applications) with the OMTD ***Workflow Editor***, which is based on the Galaxy project (Afgan et al., 2016). The Galaxy workflow editor environment[17] has been embedded in the OMTD platform; it allows users to select components, to compile them into workflows, and to configure them by setting the respective parameters. Saving a workflow in the editor exports it to the OMTD Registry so it can be used to process a corpus of the user's choice.

To enhance robustness and modularity in our setup, a second instance of the Galaxy software is used as the ***Workflow Execution Engine***; i.e. it takes the responsibility to execute a workflow step-by-step.

Each step is actually executed in a ***cloud-enabled cluster*** which is deployed and configured to run Docker containers with the aim to provide portability, scalability, and performance. The cloud aspect of the platform is necessary in order to handle, in a smooth and responsive way, the various loads posed by multiple users requesting TDM resources for large amounts of data at the same time. The cluster is based on the Apache Mesos[18] project (used for resource allocation and negotiation), and Apache Chronos[19] (used as a scheduler). These tools, coupled with Cadvisor[20] (data recording

[12] https://guidelines.openminted.eu/guidelines_for_providers_of_corpora/

[13] Full details on how to describe and package applications and components are available in the OMTD Guidelineshttps://guidelines.openminted.eu/guidelines_for_providers_of_sw_resources/.

[14] http://search.maven.org

[15] https://hub.docker.com

[16] http://docs.oasis-open.org/uima/v1.0/os/uima-spec-os.html#_Toc205201050

[17] For the editor we use a cut-down version of Galaxy with only the functionalities that are required in OMTD. For example, the actions for managing data (e.g. uploading data) have been disabled from the menu since this task is undertaken by the OMTD software.

[18] https://mesos.apache.org

[19] https://mesos.github.io/chronos/

[20] https://github.com/google/cadvisor

and collection at the container level), Prometheus[21] (aggregator), and Grafana[22] (visualization), form a full open-source stack that enables provisioning VMs on the cloud and running TDM components as Docker containers. In OMTD, Galaxy executes workflows using a job runner implementation[23] for Chronos; the runner was developed in the context of OMTD and will be included in the next releases of Galaxy. In OMTD, all components run in Docker containers. For components that are not distributed as Docker images, i.e. UIMA and GATE, as well as for web services, built-in Docker images are used. For example, a generic UIMA Docker image has been created that contains an OMTD UIMA executor that is responsible for downloading the component and its dependencies from Maven Central, and running it with the provided input and output parameters. Similar images have been created for GATE components (also Maven-based) and web services. In the latter case, the image contains a generic client that is used to call any OMTD-compliant Web service. The OMTD platform is currently at the stage of being populated with components and applications offered by the consortium partners and successful applicants of an Open Call issued during the project. At the end of this process, it is expected to host at least 200 components and 15 applications tailored to specific use cases.

3.4. Management of the output results

The output results of a processing procedure (managed by the ***Annotation Service***) are saved in the OMTD ***Data Storage***, from where they can be accessed by the rest of the platform (as also happens with the input corpus).
Auto-generated metadata records, based on those of the input corpora enriched with information from the metadata records of the processing applications, are provided to facilitate the registration process of the new resources. Users can decide to download them or make them publicly available through the platform.
The majority of the TDM applications in OMTD produce as output annotated corpora, but there are also applications with other outputs, such as lists of terms/named entities.
When the output is an annotated corpus formatted with the UIMA CAS XMI format, it can be fed into the ***Annotation Viewer*** that has been implemented and integrated in the platform, thus allowing users to better understand the results.

4. The OMTD-SHARE metadata schema

One of the main pillars of OMTD is the **OMTD-SHARE schema** which is used for the formal description of all registered resources. It forms an integral part of the platform as it is used for the registration and operation of the resources. It aims to serve as a facilitator, providing the interoperability bridge between the various resource types involved in TDM processes. It also acts as an intermediary with the target audience of TDM developers and consumers. Given that users come from different scientific communities, they often use different terms for similar concepts. Thus, the design of the schema takes into consideration this fact and proposes a *common core vocabulary* for the formal description of resources and their properties, including links to established vocabularies of various communities.
OMTD-SHARE is largely based on the META-SHARE metadata schema (Gavrilidou et al., 2012) which caters for the description of language resources (data and tools/technologies used for their processing). OMTD-SHARE is more restricted as it focuses on textual data, but it also extends the basic schema in order to include concepts specific to TDM and the scholarly content, and to better describe processing workflows.
For the schema, we have adopted the same design principles that were used for the META-SHARE infrastructure and similar initiatives (Piperidis et al., 2015). It, therefore, sets out to document the full lifecycle of a resource, from its inception to its usage, and its relations to other entities (e.g. actors that have created or used the resources, the projects that have funded them, derived versions of the resources, tools used for their processing, etc.). To encode this wealth of information, the schema includes a large number of metadata elements organized in a structure of semantically related modules ("components"), following the Component Metadata Infrastructure (CMDI) paradigm (Broeder et al., 2012). A subset of the metadata elements, especially those related to administrative features (e.g. identification, contact, licensing information, etc.), are common to all types of resources, while other elements, mainly those representing technical features about the contents and format of resources, differ across resource types. Where possible, closed and open controlled vocabularies[24] are preferred over free-text fields in order to foster interoperability. Moreover, certain elements are used as connecting links across resource types: e.g. "*data format*" is used in the description of the technical properties of a document but is also used in the specifications of a component or application, as regards its input or output; thus, for instance, components that take as input PDF files can be matched with documents of PDF format.
To allow for an easier encoding by the providers, a **minimal version** of the schema has been set up by a careful selection of mandatory and strongly recommended elements:

- required for the execution of the processing workflows: distributable form, format (for content and knowledge resources), variant (for models), command, parameters, specifications for the input and output resource (including resource type, language, format, character encoding, and annotation types), dependencies on knowledge resources (e.g. typesystem, ontology, model), implementation framework (for software)
- deemed valuable for discovery purposes: description, languages, domain, keywords, subtype and contents (for knowledge resources), function (for software), annotation type (for corpora and software)
- interesting for documentation: contact details (email

[21] https://prometheus.io/
[22] https://grafana.com
[23] https://docs.galaxyproject.org/en/latest/dev/build_a_job_runner.html
[24] The OMTD-SHARE ontology (http://w3id.org/meta-share/omtd-share/) will formally further this practice. Currently, the ontology caters for TDM operations, data formats, annotation types and TDM methods.

and/or landing page), accompanying documentation (e.g. user manuals, relevant publications, etc.)

- necessary for identification and citation purposes: name, version, identifier, creator, user query (for built corpora)
- providing information on the access legal terms: rights statement, license (name and URL or text)

The OMTD-SHARE schema is defined by a set of XML Schema Definition (XSD) files[25] and each OMTD resource is described with an XML metadata file. Resource providers can upload XML files together with their resources or use the OMTD editor form when registering their resources. Where possible, functionalities for automatically generating or converting existing metadata are provided by OMTD.

5. Distribution and documentation of the OMTD platform

The code of the OMTD platform is open source and available on GitHub[26]. All the software components of the platform are deployed in virtual machines at ~OKEANOS IaaS[27], the cloud infrastructure of GRNET[28], a member of the OMTD consortium; most of the components run within Docker containers (e.g. Registry, Annotation Viewer, all TDM components) which facilitates fast deployment and portability. The entry point of the platform is: https://services.openminted.eu/, from where users can access all services. More details (e.g. deliverables, news, related events) about OMTD can be found at the project's portal[29].

6. Comparison with other platforms

The CLARINO Language Analysis Portal (LAP) (Lapponi et al., 2015) and the Language Application Grid (LAPPS) (Ide et al., 2014) are examples of alternative platforms which offer text analysis capabilities. However, they differ from OMTD in their respective goals as well as in their technical setups. LAPPS and LAP both focus on NLP in general, not on the mining of scientific publications in particular. While OMTD connects to major publication aggregators, LAPPS connects to the Linguistic Data Consortium for corpora and LAP expects users to upload the content they wish to process. OMTD and LAP both dynamically deploy workflows for execution via Galaxy; OMTD uses a Docker-based deployment while LAP deploys to a HPC system. LAPPS makes uses of statically deployed processing (web) services. All three systems use Galaxy as their workflow editor. In contrast to OMTD, the other two platforms do not offer a simple processing UI for novice users. With their workflow editing capabilities and generic NLP components, they usually target expert users.

To the best of our knowledge, OMTD is presently the only platform dedicated specifically to using ready-built TDM applications on scientific publications and providing access to large amounts of such publications by connecting to content aggregators.

7. Conclusions

In this paper we have described the OMTD platform, which purports to foster TDM in the scholarly communication world. Its major achievements can be summarised as:

- bringing content and applications together in the same environment and facilitating access to them,
- enabling researchers to easily select content and run TDM applications on it without complex technical details,
- reducing costs of operation for researchers (no need for storage and high computing),
- defining and putting into practice a set of interoperability specifications and guidelines across TDM components but also with input content and ancillary knowledge resources.

8. Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme (H2020-EINFRA-2014-2) under grant agreement No.654021 (OpenMinTeD). It reflects only the authors' views and the EU is not liable for any use that may be made of the information contained therein. We would like to thank all our colleagues working in the OpenMinTeD project consortium: Athena Research and Innovation Center in Information, Communication and Knowledge Technologies, University of Manchester, Technische Universität Darmstadt, Institut National de la Recherche Agronomigue, European Molecular Biology Laboratory, Agro-Know, Stichting LIBER, University of Amsterdam, Open University UK, École Polytechnique Fédérale De Lausanne, Fundación Centro Nacional de Investigaciones Oncologicas Carlos III, The University of Sheffield, GESIS – Leibniz - Institut für Sozialwissenschaftem, The Greek Research and Technology Network, Frontiers Media SA, University of Glasgow, Barcelona Supercomputing Centre.

9. Bibliographical References

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., Grüning, B., Guerler, A., Hillman-Jackson, J., Von Kuster, G., Rasche, E., Soranzo, N., Turaga, N., Taylor, J., Nekrutenko, A., and Goecks, J. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.*, 44(W1):W3–W10.

Broeder, D., van Uytvanck, D., Gavrilidou, M., Trippel, T., and Windhouwer, M. (2012). Standardizing a component metadata infrastructure. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and*

[25]The set of XSD's is available at: http://github.com/openminted/omtd-share-schema; for an introduction to the schema, see: https://guidelines.openminted.eu/the_omtd-share_metadata_schema.html

[26]https://github.com/openminted

[27]https://okeanos.grnet.gr

[28]Greek Research and Technology Network, http://www.grnet.gr

[29]http://openminted.eu/

Evaluation (LREC'12), Istanbul, Turkey, may. European Language Resources Association (ELRA).

Cunningham, H., Tablan, V., Roberts, A., and Bontcheva, K. (2013). Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS Computational Biology*, 9(2):e1002854, 02.

Ferrucci, D., Lally, A., Verspoor, K., and Nyberg, E. (2009). Unstructured information management architecture (UIMA) version 1.0. OASIS Standard, mar.

Gavrilidou, M., Labropoulou, P., Desipri, E., Piperidis, S., Papageorgiou, H., Monachini, M., Frontini, F., Declerck, T., Francopoulo, G., Arranz, V., and Mapelli, V. (2012). The META-SHARE Metadata Schema for the Description of Language Resources. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC'12)*, Paris, France, may. European Language Resources Association (ELRA), ELRA.

Hearst, M. A. (1999). Untangling text data mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 3–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ide, N., Pustejovsky, J., Cieri, C., Nyberg, E., Wang, D., Suderman, K., Verhagen, M., and Wright, J. (2014). The language application grid. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Knoth, P. and Zdrahal, Z. (2012). Core: Three access levels to underpin open access. *D-Lib Magazine*, 18(11/12), nov.

Lapponi, E., Oepen, S., Skjærholt, A., and Velldal, E. (2015). Lap: The clarino language analysis portal. In *Proceedings of the CLARIN Annual Conference (CAC) 2015*, Wroclaw, Poland, October.

Piperidis, S., Galanis, D., Bakagianni, J., and Sofianopoulos, S. (2015). A data sharing and annotation service infrastructure. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 97–102, Beijing, China, July. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.