White paper on OpenMinTed Community Requirements

December 1, 2016

Author(s):

Agroknow	Akrivi Katifori, Panagiotis Zervas, Nikos Manouselis
GESIS	Mandy Neumann, Peter Mutschke
INRA	Sophie Aubin, Mouhamadou Ba, Philippe Bessières, Robert Bossy, Estelle Chaix, Louise Deléger, Patricia Geretto, Claire Nédellec
ARC	Natalia Manola
OU	Nancy Pontica, Lucas Anastasiou
Frontiers	Frederick Fenter
EMBL - EBI	Christoph Steinbeck, C.J. Rupp
CNIO	Martin Krallinger
EPFL	Renaud Richardet
UNIMAN	Matt Shardlow

The objective of this document is to summarize the Community Requirements Analysis report delivered in D4.2, which recorded and presented the different needs of the participating user communities, identifying commonalities and differences. The deliverable presents a set of the most representative stakeholder personas and concludes with concrete requirements both for the OpenMinTeD platform and the foreseen use case applications.

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

Table of Contents

Table of Figures

Index 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).

Publishable Summary

This paper presents a summary of the results recorded during the requirements elicitation phase of OpenMinTeD in deliverable D4.2 Community requirements analysis report. The deliverable is the second deliverable of Work Package 4 titled "Community Driven Requirements and Evaluation" and aims to provide an overview of the requirements relevant to Text and Data Mining (TDM) from research communities that have been identified as potential end users of the OpenMinTeD project services. The first step in the process of the OpenMinTeD project developing its TDM-powered services, as well as its services to end users and e-infrastructure, is the identification of the actual needs of the communities that are currently in need of these services. As described in the following sections, the requirements have been elicited through targeted online surveys, interviews, focus groups and workshops that engaged different stakeholder representatives, including text mining researchers as well as various end user stakeholder types and communities, based on the methodology described in D4.1 Requirements methodology deliverable.

1. Introduction

The aim of the OpenMinTeD project is to enable the development of an infrastructure that fosters and facilitates the use of text and data mining (TDM) technologies in the field of scientific publications but not limited to it, by two main user categories: application domain users and text-mining experts. OpenMinTeD aims to take advantage of existing tools and text mining platforms, facilitating access to them through the appropriate registries, and enabling or enhancing their interoperability through an interoperability layer based on existing standards. OpenMinTeD supports awareness of the benefits and training of text mining users and developers alike and demonstrates the merits of the approach through a number of use cases identified by scholars and experts from different scientific areas. It brings together different types of stakeholders, including content providers and scientific communities, text mining and infrastructure builders, legal experts, data and computing centers, industrial players and SMEs.

OpenMinTeD involves researchers from four scientific communities ranging from scholarly communication (OpenAIRE, UK/CORE, LIBER, Frontiers), to Life sciences, focusing on biochemistry (EMBL-EBI) and neuroinformatics (Human Brain Project), Agriculture (INRA, Agroknow/FAO) and Social sciences (GESIS) domains. In the context of WP04 Community Driven Requirements and Evaluation the project aims to (i) gather requirements and chart the respective fields as to TDM usage and practices as well as tools, resources and standards used, (ii) define prototype applications that serve the corresponding scientific communities via the OpenMinTeD infrastructure, and (iii) evaluate these applications in relation to the infrastructure. Deliverable D4.2 Community Requirements Analysis Report focuses on the first point and includes the analysis of each of the four user communities that will assist in prioritizing the importance of the identified needs and defining the main project stakeholder personas. It will also produce a concrete requirements list to feed the functional specifications process as well as provide a set of main stakeholder personas and their needs in order to guide the design and development of the OpenMinTeD platform and application use cases.

The remainder of the document is structured as follows: Section 2 provides a brief overview of the processes and activities carried out for requirements elicitation. Section 4 provides an overview of the relevant communities involved in OpenMinTeD and their particular needs in relation to TDM. And section 4 highlights the needs of the main identified stakeholders and their representative personas.

2. *Requirements elicitation methodology*

As more details on the requirements elicitation methodology can be found in deliverable D4.1 Requirements Methodology, this section summarizes the evaluation activities that were realized in the context of this methodology. The methodology proposed a general evaluation process to be adapted to the specificities of each community so as to record the needs of each of the communities, as well as a general approach across communities to validate the collected requirements of each of the main stakeholder groups. More specifically, the requirements elicitation process was completed in three main phases:

Phase 1 included a **preliminary survey** to identify a first set of needs, implicit or explicit, as expressed by the end users of TDM services, and to record a preliminary outlook of the users towards issues like access to content, licensing, storage, etc., in relation to TDM. These surveys were created based on a basic survey outline and adapted by each Community leader to the needs of each community. They were distributed as on-line surveys to potential stakeholders of TDM services.

Phase 2 included a set of activities with the objective to:

1. **Map each Community** and identify main stakeholders. This activity was initiated during the User Workshop that took place in Athens in February 2016, consolidated during the period immediately after the meeting and served as the basis of the Community reports.
2. **Enrich recorded needs** through focus groups and interviews with representatives of the identified stakeholders within and outside the project consortium. This activity took place at the Community level by the consortium partners of each Community.
3. **Compile the "community reports"** essentially recording the outlook of each Community towards TDM issues through the perspectives of the main project partners in each one.
4. **Record text mining researcher needs.** The main stakeholder group of OpenMinTeD was targeted through focus group sessions and interviews.

Phase 3 included the **validation of requirements** through surveys and interviews and through a continuous validation process with the project partners and the creation of the **main stakeholder personas**. A questionnaire per main stakeholder type was created and distributed as an on-line survey or used to guide structured interviews. In parallel, a working list of requirements was available in order for all consortium partners to review and comment on and to guide early on the functional specifications process (Deliverable D4.3).

Through the requirements elicitation process a set of representative personas has been identified. A persona is a representation of the goals and behavior of a hypothesized group of users, personified through a fictional, representative of this group. Personas are a strong tool to support a deeper understanding of the users throughout the whole design, development and testing cycles of a software product.

3. *Overview of the communities*

This section presents the four communities that have been examined within OpenMinTeD to record challenges and needs in terms of Text and Data mining.

3.1 Scholarly communication

Scholarly communication progresses on research and the sharing of information that results from that research. The cycle (Figure 1) starts with a question or an idea that is pursued and discussed leading to research. The next step may take the form of informal communication such as talking to colleagues sharing the same interests and sharing emails. These may result to a presentation, a poster, a panel presentation at a professional meeting. At the end of this process, information sharing becomes more formalized as researchers publish in a journal article.

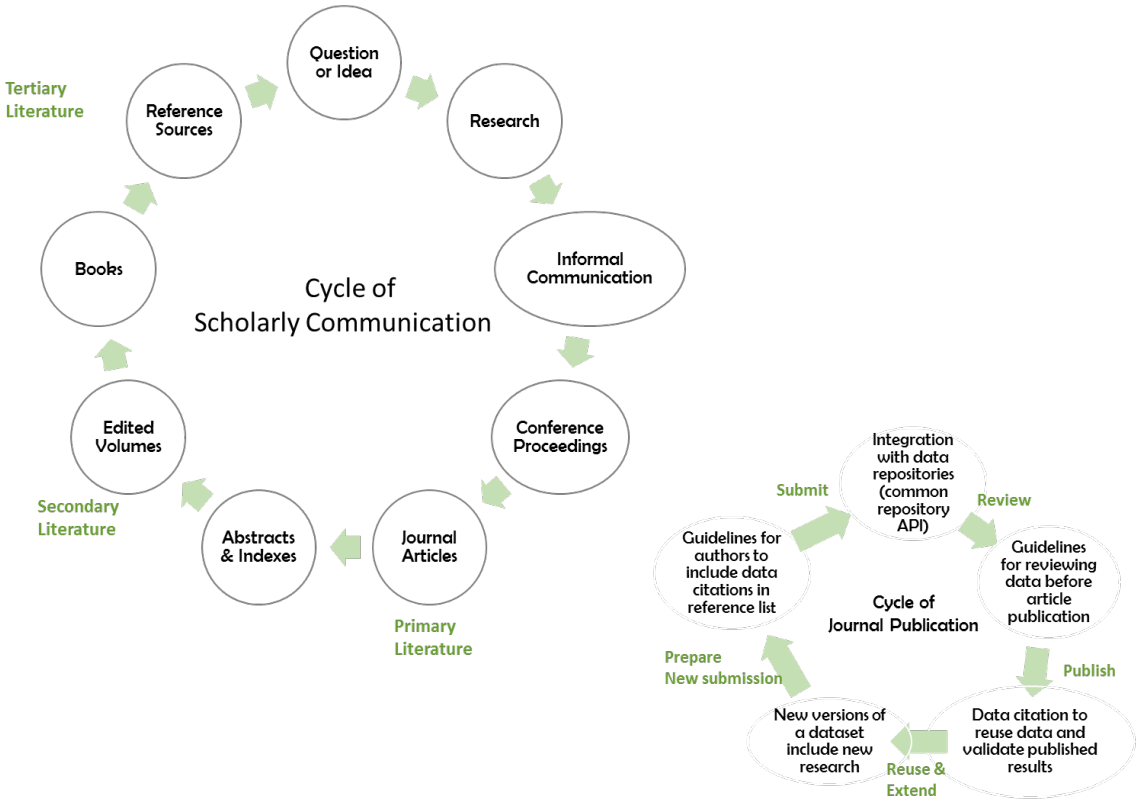


Figure 1 Scholarly production is cyclical in nature.

In Scholarly Communication, bottlenecks arise when information is required at every stage. Traditional structures for literature archiving, searching and discovery are archaic in the face of 2M articles published every year. Reviewing, indexing, alerts become exponentially more tedious and complicated. Licensing and Intellectual Property rights and restrictions, while protecting the original authors, impede the information sharing and progression into the next cycle.

Text and Data Mining innovations should make these cycles work more effectively. Text and Data Mining tools should provide all stakeholders with tools that work in two directions:

- Provide advanced search and retrieval mechanisms for items (data / articles / etc.) that directly touch upon their interests and expertise
- Allow them to leverage large data sets of articles to draw general conclusions (meta-studies, epidemiological studies, etc.)

At the moment, most content providers and aggregators are informed about the benefits of TDM for a more effective content provision and explore different ways to integrate TDM tools and services in their workflows.

The Scholarly communications community is composed of the following main stakeholders:

Policy makers: These define the research landscape by introducing policies for funded research and include public and private funding bodies that sponsor research projects.

Content providers: These include:

a) **Publishers**: there are mainly two categories of publishers, the proprietary/traditional/subscription based/toll access publishers, whose main income derives from a subscription based model, and the pure open access publishers who operate under various business models.

b) **Repositories**: these can be mainly divided further into two categories, the institutional repositories, which collect the scientific outputs of one institution (i.e. academic, research, etc.) and the subject repositories, which collect scientific outputs in specific field(s).

c) **Libraries**: There are a variety of different libraries, such as academic, research, special (corporate, legal, industry, etc.), public, online. All these can serve as a content provider for TDM purposes.

d) **Scholarly communication databases**: These can be bibliographic and citation databases, or other subscription based databases that offer access to peer-reviewed scientific content.

Aggregators: Aggregators or harvesters, like CORE or OpenAIRE collect in one searching point (peer-reviewed and non-peer-reviewed) scientific information from a variety of data providers. Aggregators serve an important role in TDM for two reasons. Firstly, the majority of the harvested content can be used for TDM purposes. When we refer to an aggregator of open access content, like CORE, then this task is made easier, due to the open access status of the aggregated content and their liberal licensing conditions. Secondly, aggregators can provide content to other services, such as the OpenAIRE project, which aims at the widest dissemination of the European Research.

Researcher Networks: The research networking sites, which host in their databases collections of scientific outputs, are becoming lately increasingly popular. The use of these outputs from text and data miners is challenging though, since these sites do not comply with interoperability standards, most of the times have proprietary systems, they do not follow metadata and other standards and do not always make their content available for TDM.

3.2 Life Sciences

The biggest challenges facing society today are biological: security of food supply, water integrity, controlling disease outbreaks in humans and plants, and caring for an ageing population, to name just a few.

The life sciences encompass the study of all aspects of living things, and have developed within diverse specialised communities over the past century. With the emergence of data-driven science, these communities are coming closer together, developing shared approaches and becoming more interdisciplinary.

The life sciences domain is in a state of 'data deluge' with thousands of new publications each week. Even in a sub-field a researcher may have to read and process dozens of relevant publications each day in order to keep up to date with the field. Clearly, this is an intractable problem by human means. As a solution, many researchers are turning to TDM for information retrieval and extraction to help process the large volumes of data. Fortunately, the techniques are advanced and capable of satisfying many of the community's needs.

The Life sciences community is particularly heterogeneous with diverse needs intersecting between sub-communities. The metabolomics specialist and the functional neuroscientist will have some core interests in common - however their specialist needs will differ. Whilst the core areas of the life sciences are well resourced for TDM, the specialisms on the fringe may suffer doubly from a lack of sufficient resources and TDM knowledge. Exceptions exist in some areas, for example the BLUIMA package provides UIMA-based interoperable TDM tools for neuroscience.

Curation and annotation are two typical tasks in this domain which both require highly specialised knowledge of the specific sub-field. Whereas some areas may use crowdsourcing for these tasks, this is very difficult in the life sciences as the crowd will not have the correct background knowledge to make an informed decision. This makes these tasks very expensive as we must employ highly educated people to do very painstaking tasks. Automation of curation, and maybe also annotation, would alleviate the cost to the researcher and reduce the amount of time a curator spends looking through irrelevant documentation.

Life sciences include the following stakeholders with interest in TDM:

Database curators carrying out literature curation, including model organism databases (MOD) like TAIR, MGI, RGD, WormBase, FlyBase, MaizeGDB; Functional genome annotation databases (e.g. GOA), proteomics databases (BioGRID, IntAct, MINT), comparative toxicogenomics (CTD).

Experimental biomedical and basic science researchers: (1) to improve the interpretation and design of experimental research by improving the access to previously published information on the studied bioentities; (2) using literature mining and knowledge discovery software for the generation of new hypotheses that will be experimentally validated.

Clinicians: improve the information access for evidence based clinical practice using text mining technologies and biomedical semantic search engines

Chemists: systematic access to chemical information (structure associated chemical entities) described in the literature and patents.

Biocurators and Bioinformaticians: Text-mining assisted curation results are useful as Gold Standard validation sets for predictive bioinformatics results.

Pharma Industry: Drug discovery and target selection, identifying adverse drug effects, competitive intelligence and knowledge management

Publishers: semantic annotation of online publications, structured digital abstracts

Scientific papers authors: author derived annotations (assisted completion of structured digital abstracts)

Patients: improved search engines, especially important for the detection of cases of similar rare disease cases and personalized medicine (automatic detection of mutations descriptions published in the literature)

Computer scientists: useful training data provided by BioCreative to improve the performance of cutting edge statistical machine learning algorithms and feature selection/exploration

3.3 Agriculture / Biodiversity

The Agricultural/Biodiversity community is quite broad and diverse since there are many disciplines covered under the label "Agricultural Science". In general, agricultural science is a broad multidisciplinary field of biology that encompasses the parts of exact, natural, economic and social sciences that are used in the practice and understanding of agriculture. The most important disciplines are agricultural and food sciences, forestry, environment and natural resources, ecology and nutrition among others. Therefore, the Agriculture/Biodiversity Community includes stakeholders from the above mentioned disciplines, disciplines that usually tend to interact and they constitute the agricultural ecosystem. All these disciplines face many challenges when it comes to extracting and combine meaningful information from disperse sources especially concerning human and animal health, disease outbreaks, environmental issues etc. OpenMinTeD approached this community through the perspectives of different sub-communities.

3.3.1 The AGRIS sub-community

The Agris ecosystem comprises all stakeholders involved in the agricultural research and practices. It includes all Agricultural Science and Technology information from research being conducted in various areas of the agricultural science, like horticulture, viticulture, apiculture, farming, agricultural economics etc.

A major stakeholder group is that of researchers who are carrying out research in major domains of agriculture, such as sustainable and organic agriculture, horticulture, viticulture, plant breeding and plant pathology, plant physiology, domains who are closely related to the domains of food science, such as food security, food safety and nutrition. These researchers may be working in public or private institutions and companies, universities, research institutes and their research data outcomes are of great importance to the whole community. Any support given to the researchers (in the form of applications, services, tools etc.) that will help them do their job more efficiently in order for them to provide meaningful conclusions and solutions.

The research data outcomes are inter-connected with the other stakeholders of the community who are in the position to take advantage of these outcomes and serve other target groups and users of the community, such as companies etc. in the most efficient way.

3.3.2 The Food Safety sub-community

Food safety is considered as a global public good and a complex problem at the same time. Globalization of the food supply has resulted in food safety risks being widely extended beyond domestic borders.

Foodborne disease outbreaks are common in both the developed and developing worlds and have serious implications for public health and trade.

The Food ecosystem or Food case comprises all stakeholders involved in the food research and practices. It includes Food Science and Technology information from research being conducted in the areas of food security, food safety, food alerts etc. Researchers in the food sub-community who are working in the domains of food security, food safety and nutrition, either in research institutions, universities or public or private organizations, are looking for ways to connect the different models of risk assessment that are developed at the moment. Also, they are trying to collect all data and models and more importantly, to extract the various pathogens from within the publications for example. All this information that will be extracted could speed up the food outbreaks cause identification and could solve at the same time the data exchange problem with local authorities, companies and other organizations.

3.3.3 Microbiology Biodiversity Researcher Sub-community

All stakeholders involved in microorganism studies, either researchers, industrials or service providers, advocate a formal and unified representation of microorganism biotope information (habitats, properties of the habitat, interaction of the species with its environment (phenotypes and molecules)) from various sources, e.g. literature and databases. This need is present both in applied microbiology (e.g. food industry, health science or waste treatment) and in fundamental research (e.g. metagenomics studies, microbial biogeography and phylo ecology).

One of the main needs of the microbial biodiversity community is the completion of the knowledge described in databases with knowledge from the literature and other databases on this subject, as well as their comparison for further analysis. While this is a significant community need, there is no resource that centralizes the knowledge on microorganism habitats. Most of the information is expressed in free text, and is not easily used because it is not standardized. The information is described in scientific papers, free text fields of databases, international microorganism culture collections and biodiversity surveys.

Information content analysis and standardization calls for information extraction tools that can automatically analyze descriptions of microorganism biotopes so that biotope descriptions originating from different experiments can be compared at a large scale. Here analysis means not only the extraction of the relevant spans of text, but also the normalization or categorization with reference resources (e.g. taxonomy of organisms, ontology of habitat, ontology of phenotypes, ontology of physico-chemical properties, etc.).

3.3.4 Crop Plant Researcher Sub-community

The Crop Plant community belongs to the Plant Biology community that is broad and diverse and ranges from fundamental research on plant biology to seed companies. Improvement of plant breeding in particular is a key to food crises and world hunger in the context of climate change. More generally, improvement of plant species of agronomical interest in the near future has become an international stake because of the increasing demand for feeding a growing world population and to mitigate the reduction of industrial resources (oil especially).

The end-user objectives of the Plant use case is thus the design of formal biological models for gene regulation networks in order to better understand the components and their interactions and to identify biological parameters of interest. The plant community expressed great interest in information extraction among other TDM tasks. The main motivation seems to extend the analysis possibilities already offered by datasets and corresponding tools and platforms. Information extracted from texts is considered an additional, complementary type of data. In that context, annotation resources need to be designed considering the already widespread reference vocabularies with some adaptation required to fit TDM tasks.

Agriculture/Biodiversity includes the following stakeholders with interest in TDM:

- Food safety officers and agencies, water quality managers
- Information managers
- Researchers
- Breeders
- Bioinformatics application providers
- Veterinarians, epidemiologists

3.4 Social Sciences

The Social Sciences Community is rather diverse, as many disciplines are covered under the label "Social Sciences". In general, social sciences are sciences that study the human society from different perspectives such as societal structures and dynamics and the processes and relationships between individual and society. Important disciplines are sociology and political sciences, but also economics, communication studies, psychology, education, anthropology and others.

A significant sub-community within Social Sciences is empirical social sciences. The term "empirical research" refers to the research performed using empirical evidence collected through direct and indirect observation or experience. On the basis of a theoretical framework the scientist creates an empirical concept for measuring his/her research question in an appropriate way. The empirical evidence (the data record of one's direct observations or experiences) can be analysed quantitatively or qualitatively. Through quantifying the evidence or making sense of it in qualitative form, a researcher can answer empirical questions, which should be clearly defined and answerable with the evidence collected (usually called *data*) and use these answers to form publications with a direct relation to the data.

Today, TDM seems to be insufficiently established in the social sciences, which may be explained by the fact that the focus in education/teaching lies on methodological education, and that quality control plays a major role in this field of study. So, social scientists don't want to rely completely on complex text mining workflows they don't understand as it is not realistic to expect them to have deeper knowledge in computer science or natural language processing. They would rather want to use some simple text mining tools to help them detect patterns in big amounts of textual data, link them to other relevant items, get a new perspective on the texts, maybe generate new hypotheses from patterns etc. But this is always followed by close reading of the texts to perform a manual qualitative analysis.

As the complexity of software tools related to text mining is steadily growing (from simple counting of words in the beginning up to detection of argumentative structures nowadays), the publication output of

text mining related research in the social sciences also grows. So it seems that this method of blended reading (or mixed methods) gains popularity here. Thus there is potential to win even more social scientists as potential users of an integrated text mining platform. The essential steps where the researchers could rely on computer assistance are the collection of texts to define corpora to work with, and frequency and co-occurrence analyses (because they are intuitive and easily interpretable). More complex steps pose a problem in the social sciences in terms of their demand for transparency and reproducibility.

The Social sciences community includes the following stakeholders with interest in TDM:

- Social scientists seeking for information relevant to their research
- Content and service providers
- Social science researchers who need to analyse textual data for their research

4. Main stakeholders and their needs

Figure 2 presents an overview of the main stakeholder groups in relation with TDM and the OpenMinTeD objectives. The remainder of this section focuses on the TDM related challenges and needs of each stakeholder group.

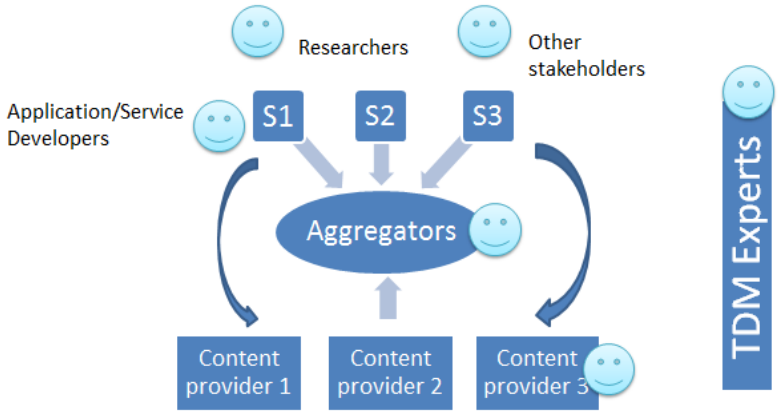


Figure 2 Main stakeholders related to the OpenMinTeD objectives

4.1 Text-mining researchers – TDM Experts

This group includes TDM experts who focus their scientific research on automated ways to derive high-quality information from text. These may include researchers of different levels, more experienced, senior ones or even PhD candidates.

4.1.1 Current practice and challenges

Text mining researchers employ a variety of textual content in the context of evaluating their research products, algorithms, tools and services. Publications, social media records and news are the most prominent ones. They also employ content metadata, to either facilitate the selection of most relevant documents or improve their text mining results.

The survey results indicate that although about 50% of the text mining researchers are familiar with licensing-related terms the other 50% is only familiar with the term “license” without any further details.

Text mining researchers consider several TDM related challenges as very important. These are related to retrieval of useful content, tools and services for their research, evaluation and benchmarking of their own research outcomes, as well as making them available to their community. Issues of licensing and retrieval of open source material seem to be prominent challenges, along with locating and retrieving relevant corpora. Leveraging various, occasionally conflicting, language resources on the same domain also often seems to be an issue. More specifically:

TDM research state of the art and public awareness.

In text mining research the notion of "off the shelf" product is still not fully established. Users of TDM software need to manually re-examine, curate and improve results of text mining processes. The tools in

many cases are error prone, noisy or incomplete and revisions are necessary. Furthermore, there are a lot of "small" tools that are research products and are marginally useful to both other researchers and the Industry. SMEs and the Industry in general cannot re-use them in a profitable way and it is generally easier to re-develop tools than locate and re-use them. On the other hand, the public has increasingly higher expectations as a result of emerging commercial TDM applications and end-to-end high quality; robust and effective, results are needed. This point is crucial for raising public awareness to TDM and the OpenMinTeD platform could be a means to that effect.

Use of standards and frameworks.

Working only with existing standards like UIMA and Gate and web services leaves out the majority of TDM research work. Industry frameworks are not sufficiently investigated and taken into account in current TDM research practice.

Component packaging and interoperability.

Research centers in general do not have the policy or mentality to publish code developed by their researchers, as commercialization and product development is a very low priority to them. Interoperability cannot be ensured at the moment as the text mining scientific community does not have the experience to "package" components, in contrast with the Industry. Most TDM researchers remain skeptical towards the feasibility of a fully interoperable solution for a workflow editor.

Incentives towards interoperability.

Researchers at the moment lack a strong incentive to commit to interoperability frameworks, such as visibility of the researcher's work, ease of use by others and being able to use other tools like their own available through the platform. OMTD could be a potential means to disseminate these incentives.

Content and language resources registry and licensing

Content is crucial as it is necessary to perform large scale experiments. However, licensing is an important issue: Can the results after mining on this content be used? What would be the special attributions needed? Licensing issues tend to be ignored within the research domain but are crucial for the Industry. The same applies for language resources: Again it is very important, especially for SMEs which need to use them also for commercial purposes.

4.1.2 TDM related needs

TDM researchers considered several incentives for adopting the OMTD platform in their work practice, including the visibility of their created tools and services and the dissemination of their work, along with the ability to easily combine their work with others' and build composite workflows, also for evaluation purposes, are very strong incentives. More specifically, the envisioned platform is expected to support their work in several ways:

OMTD platform as a dissemination tool, for scientific and commercial purposes.

OMTD can be used a strong dissemination tool of the tools and services created by text mining researchers. A platform offering access to a registry of software related to TDM can act as a common point for the researchers to gain visibility.

OMTD platform can support TDM research.

Text mining researchers agreed that a common registry of TDM tools and services offered by OMTD would be valuable as a tool to preview and use other researchers' software. It is often the case that researchers and TDM service developers need to:

- Access components useful for their TDM workflows
- Review the state of the art on specific aspects of TDM to see if their research is redundant
- Compare components, to prove that theirs is better or to choose the best solution

Managing TDM workflows.

There is a strong need for the researchers to be able to manage TDM workflows through a platform that ensures reproducibility and long term availability of results. Visibility of the whole processing of the workflows is considered very important, with the possibility to look into the intermediate results to identify and remedy possible issues. The workflow editor offered should provide comprehensive workflow management functionalities with emphasis on ease of use and a graphical user interface that will guide even less experienced users in need of TDM services.

Access to textual content and TDM resources.

Researchers would like to be able, through the same platform, access textual content and linguistic resources (reference corpora, lexica, ontologies, etc.) to support their research. There is a need to support the potential evolution of resources, through automatic and manual curation. There is thus an implicit need for managing these resources, including issues of versioning, confidence, providence, etc. Comparison of such annotation is crucial, for evaluation purposes and to resolve possible conflicts.

Hardware resources and storage.

Researchers would like to have access to resources where they could run their workflows and store the produced processed data and resources.

This persona has been created to represent the text mining researchers' needs and objectives.

Table 1 Marios Antoniou – Representative of the Text mining researcher persona

 **Quote:** "*My research would be much easier if I could have direct access to the state of the art in*	**The Text-mining researcher** Name: **Marios Antoniou** Age: 28 Location: Athens, Greece Technical comfort: experienced software developer Profession: Text mining PhD researcher

tools, content, resources."	
Back story Marios is a computer scientist who has recently started his PhD in the Department of Informatics and Telecommunications of the University of Athens. His general domain of research is TDM and, more specifically, entity extraction from publications. He is investigating the state of the art while at the same time preparing experiments to test algorithms he has been developing in collaboration with other researchers in his team.	
Motivations Marios needs to be able to have a comprehensive overview of the state of the art in his field, so as to define more clearly his research domain and select the most cutting edge areas to focus. He is often in need of certain text mining components to combine them with his own to run specific workflows for testing purposes.	
Frustrations Marios finds himself spending a lot of time just to research the existing state of the art. He knows there are several relevant tools and services to his work, through the existing publications he discovers, however is has been proven very difficult to locate the software. Even in the cases when he is able to find the implementations of certain approaches, it is a significant challenge to make them work and evaluate them in the context of the workflows he is testing. Locating and retrieving test corpora and language resources is again an important challenge for him as he is faced with licensing issues and he is frequently being blocked when trying to download document collections from specific publishers.	
Ideal experience Marios would like an easy, direct way to locate and retrieve resources related to his research. He would like to be able, through a single access point, to explore existing tools and services, combine them in workflows and evaluate them, comparing them with his own tools. Similarly, he feels that having direct access to language resources and corpora to use as input to his workflows would be a significant benefit to his work. A single registry of TDM tools and services combined with workflow editing and management functionality will support him in all aspects of his work, including defining his research objectives, examining the state of the art to compare it with his own work, run benchmarks and evaluations of his tools, with relevant document corpora and language resources, and make his work available and visible to the rest of the TDM community.	

4.2 Researchers

Researchers are actively involved in (i.e. conducting) scientific research part of which is the search, retrieval and study of textual information, including scientific publications. They are the end users of the text mining products accessible through the OpenMinTeD platform, the domain researchers who will employ TDM-powered services to retrieve content relevant to their scientific or other interests. In this group we can include other stakeholders who have similar search and retrieval needs, but not necessarily for pure scientific purposes. These could include policy makers, funding bodies, etc.

4.2.1 Current practice and challenges

Researchers employ a variety of textual content, according to their specific domain of interest and needs, with publications, social media records and news being the most prominent ones.

Most of the responders (63%) were familiar with the term "license" but had no clear idea of details in this domain, although they sometimes face licensing issues when looking for content.

Researchers considered several TDM related challenges as important to their work:

Search and retrieval of content

Researchers still face issues when searching for relevant content, both in precision and recall. Precision of the results is presented as a main issue. Users often get unusable results due to ambiguities in the search terms.

As one of the responders explained, "*Finding suitable information at all is a challenge. Often the keyword terms used do not produce any results and thus, one has to check all possible synonyms to find anything at all, especially regarding topics on which very little research has been done so far.*"

In some cases, the users need to restrict their search in particular sections of the textual content, and the lack of this functionality can result in lack of precision in the search results.

Processing the search results

Having retrieved textual content that is potentially useful for their research objectives, researchers need to analyse the results. The quick identification of the topic of the document is again a challenge, in order to assess its relevance, as well as, correspondingly, the identification of similar documents. A

The majority of the responders feel a strong need for added value services and functionality to support them in searching and processing of the search results. This additional processing functionality would not only help them exclude irrelevant results but also make sense of the relevant ones by quickly detecting patterns in the content itself.

4.2.2 TDM related needs

Researchers highlighted the need for effective search and retrieval in terms of precision and recall of open access full text documents, for visualizations to present text mining results and for query refinement based on relevant results already identified. Specifically, researchers felt that they would benefit significantly from the following TDM related services:

Multi-step search results refinement. Users would like additional search functionality among the results of the initial query, with different criteria. They would like a "*multi-step*" search process to be able to define complex criteria beyond Boolean operators that would make it possible to exclude directly irrelevant results

Using controlled vocabularies. The researchers mentioned the use of controlled vocabularies as a necessity in different contexts. Firstly, they felt that standardized metadata vocabularies are crucial for search and retrieval as they can be used to locate documents which may not include the search term (for example

"health behaviour" but rather a related sub-term ("alcohol consumption"). Vocabularies can also guide authors towards standard terminologies and more effective manual classification of publications. The use of vocabularies can also support effective inter-disciplinary searchers. GESIS provides search term recommender based on controlled vocabulary, but disambiguation of search terms is still an open issue.

Automatic extraction of named entities and terms. Users would find very helpful to be able to extract automatically from the search results specific information. Objects to be identified in texts should include not only Named Entities (e.g. taxa or genes) but also terms (e.g. anatomy parts or phenotypic traits) which can present variation such as syntactic variation (e.g. *length of panicle* vs. *panicle length*).

They would also like to produce statistics on the corpus about the number of occurrences of specific facts, for example number of demonstrations mentioned in a corpus of newspapers articles

Identification of entity relations. An important case of information extraction need expressed by the users was the identification and presentation of relations between different information objects, for example, relations of citations between publications and research data, topic relationships between publications / authors and relations between entities within the same publication.

Resolution of ambiguities. The researchers considered very helpful to their research the resolution of conceptual ambiguities ("Washington" the city and "Washington" the president), but, more importantly the resolution of topic ambiguities. For example, in the case of searches on "youth and violence" they would like to be able to differentiate between results that discuss "violence BY youth", and those that discuss "violence AGAINST youth"

Open access. Researchers feel it is very useful to be able to view clearly in results hit list whether a document is open access or not and what is the exact licensing schema of specific datasets they need to use. Most of the researchers felt that open access is crucial; however, they did not report that licensing issues have been a serious impediment to their work. As one of the researchers commented, "*If I need the particular document or corpus, I will find a way to solve this issue*".

Language. The issue of language was mentioned as important. The researchers would like to be able to automatically search in different languages at the same time using the equivalent translated terms. In the case of research in older documents, the issue of archaic language in the terminology used was mentioned, which needs to be taken into account in text mining tools and services.

Linking publications and datasets. This issue is considered crucial by researchers who work with quantitative datasets (survey results). They would like from the publication a reference and access to the related datasets. And, equally important would be to be able to discover the part of the data relevant to specific variables mentioned in the publications

Reliability and trust. In general, although the researchers expressed a need for text mining functionality to become a part of their research workflow, they equally expressed concerns on the maturity and quality of such services, especially in complex use cases as topic ambiguity and sentiment analysis. They felt that they need an objective and "*correct*" measure of the performance of the automatic text mining process on their particular information need, explicitly present as an indication offered with the result list. Without a way to trust the performance of the TDM services, they will not feel reassured that they did not miss any significant

part of the documents or data due to errors in TDM process. And in this case they would feel the need to manually verify the results. As one of the researchers commented, "*I am sceptical whether the results of the automated processing of texts are so robust that I can present them to my scientific community.*"

A domain dedicated portal. Users want federated access to textual content from various literature databases. This portal would offer adapted search facilities including facets corresponding to fine grained objects of interest for the researcher, e.g. genes, phenotypic traits, markers, etc. The portal should be constantly updated towards its sources.

Shared vocabularies and reference lists. They should be used to annotate texts in order to create bridges between data from databases and data extracted from texts. Unique and persistent identifiers for objects described in annotation resources must be included in annotation sets and TDM outputs. To facilitate their use in other tools, TDM outputs should be represented in open standard formats. Whenever possible, existing expert resources developed for other purposes should be considered for use in TDM workflows, avoiding duplicate efforts and irrelevant results.

TDM based approaches for vocabulary design and adaptation. TDM users of the domain ask that they are made available to the community in order to leverage the production of knowledge. Easy access to extraction tools and resources is necessary as well as training and documentation. TDM output curation and visualization tools including vocabulary editing ones are of major interest to this end.

Generic TDM workflows adaptable to specific needs. Generic workflows should be built and shared so that they can be reused and adapted to specific needs, e.g. same question but different species or plant part.

The persona in Table 2 is representative of the researcher stakeholder type.

Table 2 Marien Van der Beek,The Social Scientist, representative of the researcher stakeholder type

 **Quote**: "*A scientific analysis of the past can help us build a better future.*"	**The Social Scientist** *Name*: **Marien Van der Beek** *Age*: 28 *Location*: Eindhoven, Netherlands *Technical comfort*: mainstream software user *Profession*: History of Technology PhD Student
Back story Marien is a PhD student in the TU/e TechnischeUniversiteit in Eindhoven. She is part of the research team focusing on history of technology and, in particular, the effects of the European railway infrastructure towards Europeanization.	
Motivations Marien is often working with publications on her research interests, as well as technical reports and, very often, surveys on the preferences and uptake of different transportation approaches, over the last years by the public. As she is interested in the European perspective of transportation, she needs to review material in	

different languages. She is fluent in English and German, apart from Dutch, but at times she needs to review relevant documents in different languages and she employs the help of colleagues or translators.

Frustrations

For her it is crucial when searching for content to get as many relevant results as possible and, at the same time, not to get too many irrelevant results that she has to waste time to sort through.

When working with surveys, Marien needs to identify quickly the relevant subparts of datasets that deal with questions she is interested in. Currently, she has to rely on the metadata of the survey for this task, which in the best case contains a list of descriptors on the survey level.

One of Marien's biggest problems is the amount of textual data she has to deal with. As she doesn't have the time to read every single document, e.g. from a corpus of 50 years of newspaper articles, she would need the help of a tool to identify the most interesting texts to select for close reading.

Ideal experience

When searching for content, Marien wants to have her information need matched as closely as possible. If she enters ambiguous search terms, she wants the system to detect this and present her with the possible alternatives she could choose from.

When working with survey data, Marien would appreciate if publication metadata was equipped with information about and direct links to survey variables that were examined in the publication. Browsing the dataset, she wants to see in what other publications the variables were referred to so that she can go and read what analysis other researchers performed on them. Ideally, an integrated search system would even be able to present her variables from different surveys that match keywords or concepts she is looking for.

When working with large corpora of texts, Marien wishes for a tool that helps her identify documents that are worth a closer investigation. The identification may be based on word frequencies, topics, interesting word co-occurrences etc. As Marien is not at all familiar with TDM algorithms and implementations, she needs to be presented with a measure of correctness, precision of the results through a tool that allows her to perform TDM on her corpus of documents in an intuitive and transparent way.

She would like to visually explore the results in the form of plots and graphs to identify certain interesting structures she would like to inspect closely, maybe performing close reading of interesting (parts of) texts identified.

4.3 Content providers

Content providers include publishers, entities (persons or organizations) that produce and distributes publications such as journals and books in printed or digital form. It also includes librarians, information scientists, knowledge and information managers, content and repository managers and curators, the content specialists, which are responsible for an institution's collections (digital and analogue).

4.3.1 Current practice and challenges

The responders to the OpenMinTeD survey provide access mostly to scientific publications. XML seems to be the most prominent format (59%), whereas 33% of the providers report to be using Dublin Core as a standard and 22% no standard. 56% of the responders offer the document location within its metadata record.

89% of the responders are familiar with licensing issues and terms and consider a particular challenge to select the appropriate licensing scheme for their content to include both modifications and TDM processing. The most preferred licensing policy seemed to be Creative commons (56%). On the question for the existence of provision of the modification of content, all responders replied negatively, except two, who explained that "*in the archival contract there is a clause that states that content may be modified by the archive (necessary for archiving reasons)*".

In a scale of 1 to 5, on how familiar they are with the application of text mining techniques to content, the responders' average response was **2.83** (std 1.34). When asked if they would allow the application of text mining on their content most of the responders (63%) were positive, whereas 26% would probably allow it, under specific conditions.

Content providers view TDM as a potential solution to a set of challenges towards their ultimate objective which is to provide to their users added value services, including unified access to resources of various types and sources and advanced search functionality, specialized on specific parts of the document and taking into account domain-specific vocabularies.

4.3.2 TDM related needs

To promote their objective for advanced and effective search functionalities on their content, providers need to enrich their metadata through appropriate TDM services:

Automated method for enriching existing metadata records

Using the Registry offered by OpenMinTeD will allow them to select the tool that is more appropriate for metadata extraction and using it through the project infrastructure to extract content metadata.

Building domain specific terminology for further text annotation/indexing/search

Content providers need to be able to manage terminology and vocabularies related to their content and enrich them with appropriate semi-automatic methods.

An important point is that the difference in level of expertise of the users should be taken into account in the OMTD platform for the Workflow editor. Reusability of tools and resources is also a key of success to reduce the development costs and technical barriers for non-experts.

Trust in TDM results is an issue that often emerged in discussions with data curators. Confidence metrics should be included in TDM results and/or easily accessible on the platform. Curation tools are also necessary to revise TDM outputs before their integration in other domain applications. When correcting outputs, the possibility of revising annotation resources should also be offered thus creating a virtuous circle.

Table 3 presents the persona created for the Content provider stakeholder.

Table 3 Marie Foss, Data curator - Represents the Content provider stakeholder type

Quote: *"If we want researchers to access our content we need to invest on enriching its metadata."*

The Data Curator

Name: Marie Foss

Age: 48

Location: Lyon

Technical comfort: experienced software user, basic programming skills

Profession: Data curator

Back story

Marie works for several years now for an open access publisher hosting scientific journals and conference proceedings. Her role in the organization is to guide the content curation processes towards metadata enrichment to make the content more searchable and accessible through the search and retrieval functionality offered to researchers in the publisher website.

Motivations

Following the recent developments in the scientific communication domain and the availability of a huge volume of publications on-line, with an increasing part of them available as open access full text, Marie feels that content providers should invest in new and advanced ways of metadata enrichment to support the needs of researchers for more precise and complete result sets in their, often complex and demanding, information needs. Part of the content that her organization manages is scientific publications from the life sciences and agriculture domains, fields where she has noticed that there is complex and rich terminology which is in constant flux. She would like to be able to take advantage of the particularities and updates in each domain in terms of vocabularies and terminology to offer more advanced search and retrieval to their users.

Frustrations

Marie, although not a computer expert originally, she has tried over the years to keep up with the advances in the fields of information search and retrieval in general and TDM in particular. Although she knows that the small IT department of her organization has not the resources to develop and integrate TDM services in their workflows, she constantly strives to find off-the-shelf TDM solutions that could provide input to their workflows in the form of content terminology taxonomies or ontologies or even TDM software to apply on the content. Locating the appropriate solutions for her case, evaluating them and applying them for their content is for her a painstaking process which has not been particularly fruitful so far. Even for her as a manager of a technical team including application developers, it has been a significant challenge to locate, install, adapt and re-use TDM software.

Ideal experience

Marie would ideally wish for a single point of access where she would be able to view information on the latest advances in the field of TDM. She would like to locate easily new solutions, not only software, but also the latest standard ontologies in specific fields. She would then like to be able to test these new approaches on her content, preferably through external services that would not require her to install and import software components into her organization workflows, and then, if the results were found satisfactory, to integrate them into the metadata scheme of their content.

A platform bringing together information on TDM with related software and other resources, along

with an easy way to apply and evaluate tools and services before integrating them in her workflows would be ideal for her.

4.4 Aggregators

The aggregator group of stakeholders includes organizations and companies which develop and support e-infrastructures that provide federated access to a multitude of content providers' material. These often employ software developers who require the integration of TDM components and services to their own, which may form part of the services provided by aggregators or content providers.

4.4.1 Current practice and challenges

Aggregators, like OpenAIRE, CORE and AGRIS, aggregate open access research outputs from repositories and journals worldwide and makes them available through a set of services. Their objective is to process and enrich the aggregated content through advanced search and retrieval functionality.

After an initial content provider and dataset registration phase, the aggregator harvests the content metadata and performs quality control and enrichment, followed by an indexing phase to make the data available through its search and retrieval functionality.

All interviewed representatives of the aggregators within the project confirmed the importance of TDM for the semantic enrichment phase. They reported using tools ranging from topic/subject identification in the abstract, based on specific domain ontologies (the case of the AGROVOC ontologies in AGRIS) to more complex enrichment processes like project, funding or citation extraction in the case of OpenAIRE.

The semantic enrichment is performed as part of the aggregation workflow of aggregator, either through off-the-shelf tools or with custom components developed in premises by the aggregator development team to tailor the TDM approach to their specific objectives.

Interoperability at the level of metadata harvesting was not reported as an issue, as usually content providers are asked to comply with specific harvesting specifications, in some cases even standards like OAI-PMH or Dublin Core, that can enable the aggregator to perform possible transformations and controls to the harvested metadata in a straightforward manner.

There were two main challenges reported in relation to TDM:

Licensing

At the moment there is not an explicit legal framework or guidelines for the licensing process in relation with the use of neither the metadata not the full text harvested by content providers for the use for text mining. Some aggregators handle this issue through the service agreement with the content provider at the registration level. Others make no explicit mention for TDM during the dataset registration phase but consider implicit the agreement of the content providers for the use of metadata for text mining purposes. As TDM gradually becomes a standard practice as the semantic enrichment phase of aggregator harvesting processes, there is a strong need for a comprehensive, explicit licensing framework to guide the process of licensing for TDM during the aggregation of metadata and content.

Accessing publication full text for TDM

A related issue is full text access for aggregated publications, either for TDM purposes or to provide it directly to the end users of the aggregator services. In some cases, a link to the full text is provided in the metadata of the publication. If this link is not explicitly present in the metadata it can be more complicated to extract and provide. Additionally, it is often the case when the aggregator needs to access the full text for TDM purposes, though, for example, web crawling that they are blocked by the content providers. In this case an appropriate licensing framework is needed to ensure that there is a smooth collaboration of content providers and aggregators.

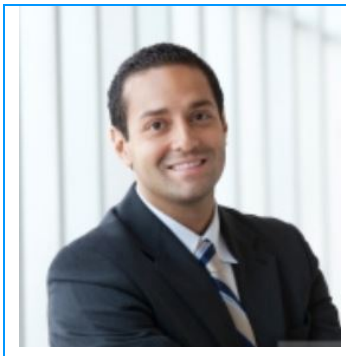
4.4.2 TDM related needs

Aggregators expressed common needs with those of the text mining researchers, in relation with:

- A Registry of text mining tools and services that will support them to easily locate and re-use text-mining software.
- A workflow editor where they can create and manage workflows for specific TDM purposes
- The possibility to run the workflows combined with language resources and enriches the metadata offered by their services.
- The latest information on licensing schemes related to the provision of content in general and TDM issues in particular

The persona in Table 4 has been created to represent the aggregator and application developer needs and objectives.

Table 4 Juan Campos, the Technical Manager, representative of the aggregator and application developer stakeholder types

<table>
<tr>
<td> 
Quote: “Making content from different providers searchable and accessible is by no means an easy task.”</td>
<td>The Technical manager
Name: Juan Campos
Age: 43
Location: Madrid, Spain
Technical comfort: experienced software developer
Profession: Technical manager</td>
</tr>
<tr>
<td colspan="2">Back story
Juan works for an organization which offers federated access to open access publications from a multitude of sources, private and public publishers as well as institutional repositories. He manages a technical team consisting of application developers and content curators with the objective to constantly improve their information system search and retrieval functionality and ensure the smooth and effective inclusion of new content providers to their aggregation services.</td>
</tr>
</table>

Motivations

As the aggregated content is constantly growing, Juan wishes to make sure that researchers accessing his organization's portal will be able to effectively retrieve the content they need. It seems that the quantity of the content available can only become an asset for the users as long as it is available through an easy to use and precise search interface with added value access and retrieval functionality.

Juan tries to stay informed of the new advances in TDM and feels that such solutions would significantly boost the quality of their search and retrieval services. Rich metadata seems to be the key to making the content discoverable and accessible by the end users. For the past few years his organization has invested in including TDM in their metadata management workflows, testing different approaches and components.

Frustrations

Juan and his team have to consolidate a wide variety of content metadata. Providers, according to their policies and needs, register their content with metadata in different formats, following different standards and in different levels of detail. Although Juan's team provides guidelines and has in place specific policies for the registration of a content provider, the quality of imported metadata is not always up to the desired standards expected by his organization so his team strives to enrich it with their internal TDM processes.

However, this is not a trivial task. On one hand, uncertainty on licensing for TDM impedes the mass processing of the content needed for metadata enrichment, as providers' reaction to this vary. Additionally, Juan is not sure what is the best way to get informed and be up to date with the state of the art on TDM and discover new processes for possible inclusion in their workflows.

Ideal experience

Juan would like to have access in a consolidated way to existing TDM approaches, with a direct way to evaluate their effectiveness and suitability for his needs, compare them with other implementations and even download and include them in his organizations workflows without significant effort.

Apart from discovering TDM software, Juan needs to be informed regularly on latest advances on domain-specific vocabularies, ontologies and terminology that can be used to annotate the content and make it accessible to the domain expert in the terminology of the domain.

5. Conclusions

The requirement elicitation activities in the context of the OpenMinTeD project confirm that there is a complex mosaic of stakeholders that have a direct or indirect interest in TDM solutions in relation to their working practices. Guided by the needs of the researchers, which are the end users of the TDM solutions, with constantly more demanding and complex information needs, content providers, aggregators, application developers find themselves in the position to seek out TDM solutions that can be incorporated to their offered services. In this sense they are in need of the research products of text-mining researchers, a stakeholder group who approaches OpenMinTeD with more direct needs, in relation to registering, organizing, evaluating and, thus, making more accessible and available these TDM research products. This documented attempted to summarize these different perspectives and present an overview of the current status, challenges and needs in relation to TDM, which eventually lead to concrete requirements for OpenMinTeD.

This document presented the summary of the results of this process which resulted in deliverable D4.2, with the ultimate objective to (a) offer a concrete functional requirements list to feed the functional specifications to be produced in the context of D4.3 Functional specifications and (b) describe through the definition of stakeholder personas the needs of the main stakeholder types, to be used as user archetypes that will guide all subsequent design and development activities within the project.