FutureTDM

Explore . Analyse . Improve

REDUCING BARRIERS AND INCREASING UPTAKE OF TEXT AND DATA MINING FOR RESEARCH ENVIRONMENTS USING A COLLABORATIVE KNOWLEDGE AND OPEN INFORMATION APPROACH

Deliverable D4.1

European Landscape of TDM Applications Report

Project

Acronym:	**FutureTDM**
Title:	Reducing Barriers and Increasing Uptake of Text and Data Mining for Research Environments using a Collaborative Knowledge and Open Information Approach
Coordinator:	SYNYO GmbH
Reference:	665940
Type:	Collaborative project
Programme:	HORIZON 2020
Theme:	GARRI-3-2014 - Scientific Information in the Digital Age: Text and Data Mining (TDM)
Start:	01. September, 2015
Duration:	24 months
Website:	http://www.futuretdm.eu/
E-Mail:	office@futuretdm.eu
Consortium:	**SYNYO GmbH**, Research & Development Department, Austria, (SYNYO) **Stichting LIBER**, The Netherlands, (LIBER) **Open Knowledge**, UK, (OK/CM) **Radboud University**, Centre for Language Studies The Netherlands, (RU) **The British Library Board**, UK, (BL) **Universiteit van Amsterdam**, Inst. for Information Law, The Netherlands, (UVA) **Athena Research and Innovation Centre in Information, Communication and Knowledge Technologies,** Inst. for Language and Speech Processing**,** Greece**,** (ARC) **Ubiquity Press Limited**, UK, (UP) **Fundacja Projekt: Polska**, Poland, (FPP)

Deliverable

Number:	**D4.1**
Title:	**European Landscape of TDM Applications Report**
Lead beneficiary:	**ARC**
Work package:	WP4: Fields of Application, Projects, Best Practices and Resources
Dissemination level:	Public (PU)
Nature:	Report (RE)
Due date:	31.05.2016
Submission date:	31.05.2016
Authors:	**Stelios Piperidis, ARC** **Kanella Pouli, ARC** **Maria Gavriilidou, ARC** **Dimitris Galanis, ARC** **Juli Bakagianni ARC**
Contributors:	**Maria Eskevich, RU** **Alessio Bertone, SYNYO**
Review:	**Maria Eskevich, RU** **Alessio Bertone, SYNYO**

Acknowledgement: This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 665940.

Table of Contents

List of Figures

1 Introduction

The growing amount of data, structured or unstructured, in the modern world in addition to the need for making sense of this data has led to the development of tools and technologies enabling it. ***Text and Data Mining (TDM)*** is a set of techniques which detect information out of massive amounts of data and present it in an understandable way. TDM finds applications in all realms of life from farming to decision making in government. The widely used WEKA (Waikato Environment for Knowledge Analysis) system[1] was initially developed to assist mining information from agricultural datasets - for example grading mushrooms and using such grading for quality classification and market pricing [Cunningham & Holmes, 1999][2]. ***TDM*** applications in healthcare aim at analysing patients' data in order to detect disease patterns to enable predictions and provide better healthcare services at reduced cost. As concerns organizations and businesses, the storage and management of data created by users in documents, emails and diverse material on social media (such as posts, likes and emoticons in e.g. product reviews) provide an insight and furthermore the ability to model user behaviour to improve products and services.

An overview of ***TDM*** applications and activities on the global scale[3] shows that all continents are engaged in a race to catch up with technology, with the United States being at the forefront. Each country has been investing in and advancing ***TDM*** in the economic sectors that are mostly serving its specific needs and interests. So, not unexpectedly, India has enhanced knowledge management systems for agricultural data[4], while China has made substantial progress in ***TDM*** concerning traditional Chinese medicine[5].

In this report, after a brief introduction to the different stakeholders and parameters involved in ***TDM***, we focus on ***TDM*** in the European Union and try to paint the landscape of ***TDM*** research, development and applications in a number of areas. We depict this landscape in different economic sectors, scientific areas and domains of activity by exploring data relevant to the available technology and (research) infrastructures, the R&D investment (mostly in terms of funded projects), the research output (in terms of scientific publications produced), the resources and tools available for ***TDM*** as well as the commercial activity (in terms of companies and organizations investing in, using and offering ***TDM*** services). Due to the fact that ***TDM*** is currently a very hot field in terms of research, development and business applications, highly convoluted with the big data hype, and as a result constantly changing, the present report aims at the creation of a landscape which is representative of the status quo but not exhaustive concerning the information provided.

[1] http://www.cs.waikato.ac.nz/~ml/weka/index.html

[2] Cunningham S.J. & Holmes G. (1999). Developing innovative applications in agriculture using data mining. In *SEARCC'99 Conference Proceedings*.

[3] ICADMA 2016: 18th International Conference on Advanced Data Mining and Applications [online]. Available at: https://www.waset.org/conference/2016/08/barcelona/ICADMA [Accessed on 20.03.16]

[4] Indian Council of Agricultural Research (2010). *Directorate of Knowledge Management in Agriculture* [online]. Available at: http://www.icar.org.in/en/information-resources.htm [Accessed on 20.03.16]

[5] Xuezhong Zhou, Yonghong Peng & Baoyan Liu (2010). Text mining for traditional Chinese medical knowledge discovery: a survey. *Journal of Biomedical Informatics*.

2 Text and Data Mining Overview

This section presents an overview of the ***Text and Data Mining*** domain at the international level. After briefly discussing the reason for the emergence and the rapid evolution of the field (section 2.1) and defining the key terms used throughout this document (section 2.2), it proceeds to discuss types of data available, data providers and distributors, worldwide; finally, it concludes with a presentation of tools and technologies for ***Text and Data Mining*** available and used in the world (section 2.3), covering open-source and commercial, research and proprietary.

2.1 The data deluge

- The number of wireless sensors and actuators worldwide has exceeded 24 million, presenting an increase of 553% between 2011 and 2016[6].
- By 2020 there will be more than 16 zettabytes of useful data (16 Trillion GB)[7].
- YouTube claims to upload 24 hours of video every minute, making the site a hugely significant data aggregator[8].
- "Every second, on average, around 6,000 tweets are tweeted on Twitter, which corresponds to over 350,000 tweets sent per minute, **500 million tweets per day** and around 200 billion tweets per year"[9].
- **74,200,000 pages** exist on Facebook, with **7 million apps and websites** integrated with Facebook on 30/5/2016.[10]
- Over 1 billion websites and 3,36 billion internet users, on 11 May 2016[11].
- On average a new scientific article is being published every 30 seconds[12].
- 60 000 publications on a single gene, p53, in the literature[13].

The above are just a few examples indicating the exponential growth of information in the digital world. Progressing in this era of ***Big Data*** poses the demand for high level techniques in order to be able to deal with and utilize the exponentially growing information covered in massive volumes of data coming both from the present and the past, as is the case of digitized cultural information. Such methods, techniques and technologies are considered indispensable in order to

[6] Hatler, M., Gurganious, D. and Chi, C. (2012). Industrial wireless sensor networks: A market dynamics report. ON World. San Diego, CA, USA.

[7] Turner V., Gantz J.F. , Reinsel D., and Minton St. (2014). *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things*. Available at: https://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf

[8] https://youtube.googleblog.com/2010/03/oops-pow-surprise24-hours-of-video-all.html

[9] http://www.internetlivestats.com/twitter-statistics/

[10] http://www.statisticbrain.com/facebook-statistics/

[11] http://www.internetlivestats.com/

[12] Spangler Sc. et al. (2014). Automated hypothesis generation based on mining scientific literature. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge Discovery and Data Mining* (KDD '14). ACM, New York, NY, USA, 1877-1886.

[13] Hager, K.M. & Gu, W. (2014). Understanding the non-canonical pathways involved in p53-mediated tumor suppression. *Carcinogenesis*, vol. 35(4), pp.740–746.

- **discover hidden information, detect patterns and trends, decode information, distinguish deceptive from non-deceptive information**,

so as to enable

- **identification of gaps and/or new requirements, new adapted and personalised actions, improved and data-aware decision making, improved products and services**, as well as to **support transparency and democracy in modern society**.

The information retrieved and the new extracted knowledge have huge impact on human lives through the creation of new ideas, services and products, thus optimising the quality of life and fostering economic development.

The set of methods, techniques and technologies supporting these goals collectively constitute *Text and Data Mining*.

2.2 What is Text and Data Mining

Text and Data Mining refers to a set of computational processes that aim to automatically extract and relate information, as well as discover patterns, from data of different types and media. Such data should be in digital form, readable and processable by computing machines, as well as legally available[14]. The terms ***Text Mining*** and ***Data Mining*** have been independently used in the previous years before being compiled into one single term. ***Data Analysis*** has been suggested as a wider term encompassing ***Text Mining***, ***Data Mining*** and ***Text and Data Mining***[15]. The main difference between traditional ***Data Analysis*** and ***TDM*** is the point of reference: ***Data Analysis*** starts from a hypothesis which is tested against data, while ***TDM*** starts from the data and discovers patterns. This is the reason why the term ***Text and Data Mining*** has been considered slightly misleading: it is the *patterns* what ***TDM*** aims at, not the *data*![16] Thus, ***Data Mining*** should have been more appropriately named ***Knowledge Mining from Data***, which is unfortunately somewhat long.

The nature of data being processed has determined the terminology used: ***Text Mining*** is the analysis of textual data, as well as all other forms of data converted to text (e.g. audio transcripts), while ***Data Mining*** started from mining databases and evolved to encompass mining all forms through which information can be transmitted: sounds, videos, images, graphs, numbers, chemical compounds, likes, clicks, etc. All these types of data are the object of ***Data Science***, an interdisciplinary scientific area including ***Statistics*****,** different types of ***Analytics*** and, of course, ***TDM*****.** These scientific terms are intertwined and in some cases are used interchangeably due to the fact that they denote procedures during which the same material (data) and sometimes the same or related techniques (e.g. ***Machine Learning, ML***) are used for different purposes.

[14] Legal availability of data and content is briefly addressed in this report. For more information on legal issues, please refer to Deliverables D3.2 Collection of TDM regulations and barriers (currently not public) and D3.3 Baseline report of policies and barriers of TDM in Europe.

[15] Triaille J.P., de Meeûs d'Argenteuil J. & de Francquen A. (2014) *Study on the legal framework of Text and Data mining (TDM)*, European Commission. Available at: http://ec.europa.eu/internal_market/copyright/docs/studies/1403_study2_en.pdf

[16] Han, J., Kamber, M., Pei, J. (2001). *Data mining: concepts and techniques*. Morgan Kaufmann. p.5.

Furthermore, the fast proliferation and uptake of ***TDM*** in almost all human activities have given rise to additional terminological proliferation. ***TDM*** is usually codified as ***Business Intelligence*** or ***Competitive Intelligence*** when used to solve business problems; as ***Research Analytics*** when it is used to monitor and measure research activities and scientific output; ***Learning Analytics*** or ***Educational Data Mining*** when it refers to the measurement, collection and analysis of data about learners in different contexts.

The terminological conundrum referred to above is illustrated in two graphs which compare the popularity of some of the above terms, as produced by the Google search engine (Google Trends[17]). Given that this tool allows the comparison of 5 terms maximally, the terms were split in two groups in order to view their evolution over time, the first one comprising the terms ***Data Mining, Text Mining, Text and Data Mining*** and ***Data Science*** (**Figure 1**), while the second group comprises the terms ***Data Mining, Data Science, Machine Learning, Data Analytics*** and ***Business Intelligence*** (Figure 2). Both groups were compared **from 2004 till present** (as regards the time frame) and **in all categories** (as regards domain of use).

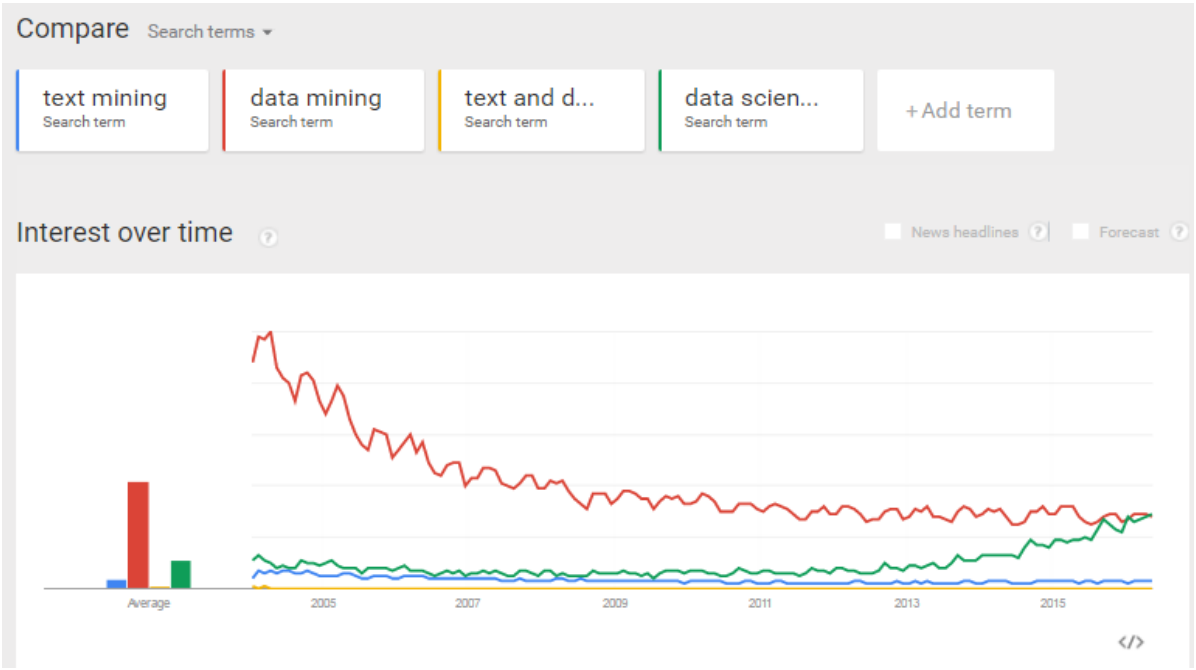


Figure 1. Comparison of first group of terms[18] [19]

Judging from its very low frequency, the terms ***Text Mining*** and ***Text and Data Mining*** are practically unknown (note that the query included all categories, not only science), while the term ***Data Mining*** started with a significant frequency in 2004, which decreased over the years to become almost equal with ***Data Science*** in 2015.

[17] https://www.google.com/trends/

[18] https://www.google.com/trends/explore#q=text%20mining%2C%20data%20mining%2C%20text%20and%20data%20mining%2C%20data%20science&cmpt=q&tz=Etc%2FGMT-3

[19] Unless a source reference is mentioned, all figures included in the present report are produced by the authors and can be reused under the CC-BY 4.0 license.

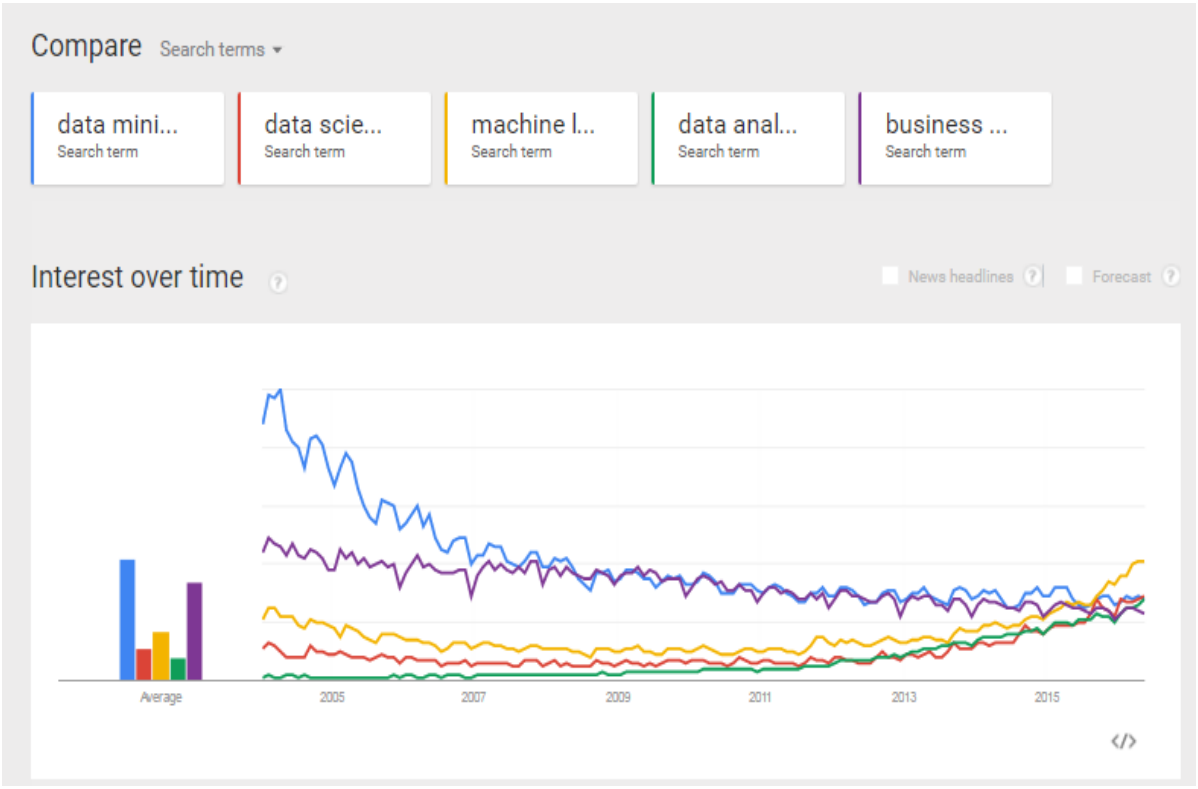


Figure 2. Comparison of second group of terms[20]

A similar pattern can be seen in the second group of terms: starting from significantly different points, with ***Data Mining*** being the most used and ***Data Analytics*** the less used, all terms seem to converge in 2015, while ***Machine Learning*** shows a rising trend.

It becomes evident that ***TDM*** is not a homogeneous, self-contained, scientific domain, but rather a diverse and complex set of methods and technologies deployed in the framework of diverse disciplines and business activities. The common core prerequisite of the ***TDM*** ecosystem is data and content, on which we focus in the next section.

2.3 Data, Tools and Technologies for mining

A huge amount of data is generated daily in small, medium and large organisations and enterprises from all human activities and all business sectors. This data is not homogeneous in nature. A key high-level dichotomy divides it into structured and unstructured.

Structured data is data that normally resides in fixed fields within a record or files, such as data in spreadsheets and/or relational databases. Structured data presupposes a data model defining *i.a.* what fields of data exist and their relations. ***Unstructured data*** refers to data that does not reside in fixed fields and cannot be so easily classified and "understood" by machines, such as *i.a.* files with running text, audio, image and video data, graphics, WebPages and emails, blog posts and tweets.

[20] https://www.google.com/trends/explore#q=data%20mining%2C%20data%20science%2C%20machine%20learning%2C%20data%20analytics%2C%20business%20intelligence&cmpt=q&tz=Etc%2FGMT-3

In-between stands ***semi-structured data***, that is, data that although does not conform to rigidly defined text fields, but contains tags and other mark-up features that annotate the data to a certain extent.

Depending on the type of data, different techniques are being deployed to extract information and knowledge from them, with each type posing its own challenges.

Structured data is mostly the object of processing by ***Data Mining*** techniques, which aim at extracting patterns by combining statistics and statistical analysis with machine learning and database management. Usually it includes: ***association rule learning*** (to discover relationships between variables in large databases, e.g. in recommendation systems and applications), ***cluster identification and analysis*** (to segregate data into smaller groups based on their common characteristics, e.g. TV-viewers or web application users for targeted marketing), ***classification*** (to identify the predefined category(-ies) in which particular data belong, e.g. mobile subscribers that will stop their subscription) and ***regression*** (to estimate how the value of a dependent variable will change when an independent variable changes, usually used to predict e.g. growth of an entity based on a number of macro/micro economic parameters). Structured ***Data Mining*** techniques already enjoy a long tradition, e.g. in financial modelling or meteorological forecasting, and is now adapting itself to manage the big volume of structured data being generated daily.

Unstructured data can be textual/audio/video data. Textual, transcribed audio data and multimedia data converted into textual representations are mostly the object of processing by **Text Mining** techniques. These techniques aim at analysing and annotating unstructured data, usually by ***linguistic annotation*** at multiple levels, both externally (e.g. assigning a domain classification label to a document) and internally (e.g. annotating spans of text as referring to concept(s) in an ontology or terminological database, or as referring to a named entity, or as being the subject/agent of verb/predicate). ***Text Mining*** (usually) involves ***Natural Language Processing (NLP)*** techniques with the ultimate goal to **turn text into structured data for further analysis**[21]**.** ***Visual Data Mining (VDM)*** is the exploration of very large datasets combining traditional mining methods and information visualization techniques within the broader field of ***computer vision*** for information extraction and structuring. ***Visual analytics*** belongs to the fields of information visualization and scientific visualization, focuses on analytical reasoning facilitated by interactive **visual** interfaces and is defined as the intertwined use of automatic and visual methods[22].

2.3.1 Sources of data for mining

In the last decade or so, with the technological advances in storage miniaturization, all **major industries** worldwide started to store different types of data. In a report of May 2011, the McKinsey Global Institute[23] estimated that by 2009, nearly all sectors in the US economy had at least an average of 200 terabytes of stored data per company, and that many sectors had more

[21] See, for example, Simpson et al (2012) for a survey of biomedical text mining.

[22] Keim, D., Kohlhammer, J., Ellis, G. & Mansmann, Fl. (Eds.). *(2010). Mastering the Information Age: Solving Problems with Visual Analytics*, Eurographics Association.

[23] Manyika J., Chui M., Brown B., Bughin J., Dobbs R., Roxburgh Ch., Byers H.A. (2011). *Big Data: The next frontier for innovation, competition and productivity.* McKinsey Global Institute.

than 1 petabyte in mean stored data per company. In the same report they showcase the amount and types of data produced by different industry sectors, as shown in Figure 3.

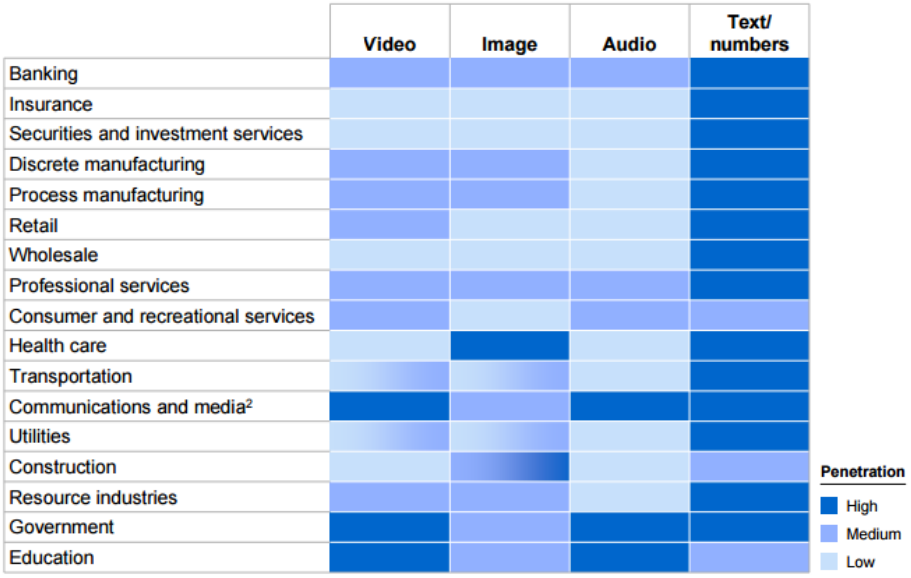


Figure 3. The types of data generated and stored by sector[24]

This data is mostly proprietary and is already undergoing massive mining processes in order to transform it into (business) intelligence that leads to improved organization knowledge management, customer relationship, products and services. Social media companies (Facebook, Twitter, LinkedIn, etc.) and their respective platforms, also, constitute major data holders and aggregators, the data of which has attracted lots of attention by both commercial and research actors aiming at a wide range of data and textual content mining applications.

Significant amounts of data are generated by the **public sector bodies**. A significant percentage, estimated to be close to 90% of government data is now in digital form, and following recent policies of the last 15 years, a vast majority is open for reuse. Following the public sector data policies of the European Commission, Member States of the EU have set up Open Data Portals[25] at the national level. National portals make available structured and unstructured datasets, in the national language, mostly under a permissive licence (CC-BY or CC-0), a national open license, or without any explicit terms but just, by virtue of being released by a public sector body, are considered to fall within the scope of the Public Sector Information (PSI) directive of the EU[26]. Data in the national Open Data Portals is described by metadata, mostly complying with the DCAT W3C recommendation, and is harvested by the EU Open Data Portal[27], which together with the EU Institutions own data, makes currently available 449,464 datasets [accessed in May 2016].

[24] Source: Executive summary, http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation
[25] https://ec.europa.eu/digital-single-market/en/open-data-portals
[26] http://www.epsiplatform.eu/category/keywords/psi-directive
[27] http://www.europeandataportal.eu/data/en/dataset

A third big strand of major data generators and providers comprises **research organisations, universities and research centres**, as well as **scientific publishing houses**. They usually make research content (mostly in terms of scientific publications and reports) available through national institutional and individual organization-based repositories and web-sites, publishers repositories and web-sites, and of course libraries that have been steadily moving from the traditional physical bookshelves to digital access of different forms. In addition to repositories, aggregators, harvest specific types of data from multiple sources, notably including repositories, in order to make them searchable in a uniform way. Aggregators, such as OpenAIRE[28] and CORE[29], have taken on an infrastructural role and mission, fostering open access, already undertaking mining initiatives to enrich the harvested content, and offering value added services. OpenAIRE has, in addition, started acting as research data repository and it also collaborates with Zenodo[30] repository hosted by CERN (a more detailed discussion of these issues is in section 3, which focuses on Europe). In the US we can refer to Earthcube[31], a network for data sharing relevant to geoscience, funded by the National Science Foundation[32].

GoogleScholar[33] by Google and Microsoft Academic Search[34] by Microsoft offer services around scientific publications and related research information. Semantic Scholar[35], a recent development launched in November 2015, focuses only on open access publications in the field of computer science, offering semantically enhanced search. Web of Science[36] by Thomson Reuters Institute of Scientific Information[37] and SciVerse Scopus[38] by Elsevier are amongst the largest abstract and citation databases of peer-reviewed research literature.

2.3.2 Tools, techniques and technologies for Data Mining

The diverse types of content mentioned above are typically processed for mining purposes using software pipelines that implement:

- ***Information retrieval***: to select the set of data / content relevant to a particular query
- ***Linguistic analysis***: to annotate data with morphological, syntactic and semantic/pragmatic information[39]
- ***Information extraction***: to identify entities, relations between entities and events in which these entities participate, thus rendering the text in a rigidly structured form

[28] https://www.openaire.eu/
[29] https://core.ac.uk/
[30] https://zenodo.org/
[31] http://earthcube.org/
[32] http://www.nsf.gov/
[33] https://scholar.google.com
[34] http://academic.research.microsoft.com
[35] https://www.semanticscholar.org
[36] http://wokinfo.com
[37] http://ip-science.thomsonreuters.com
[38] http://www.scopus.com
[39] This type of analysis is not applicable to non textual data such as numerical information from temperature measurements or clicks.

- ***Data Mining*** that leads to knowledge discovery: to extract additional information, relations, trends, predictions, etc.

A range of tools are currently in use worldwide to implement these processing stages. A quite large number of tools are open-source (e.g. Lucene/Solr for retrieval, UIMA/GATE compliant tools, as well as framework agnostic language processing web services for linguistic analysis and information extraction), while there are also proprietary frameworks catering for ***Text and Data Mining*** either across sectors or specializing in vertical applications (i.e. applications that support a specific business area or process). Examples of such commercial offerings, with a focus on non-EU players in this section[40], include:

- *Across sectors*
    - *Microsoft SQL Analysis Server, Oracle Data Mining* with Oracle Advanced Analytics and Oracle Data Mining component SQL functions*, Open Calais,* based on Natural Language Processing and Machine Learning, with a lead in ***Text Mining***, mainly for usage across sectors;

- *Vertical applications*
    - *IBM SPSS Predictive Analysis-IBM Watson* and the acquired *AlchemyAPI*, for the medical sector (Medical trials and treatments), financial and energy sectors, retail and education;
    - SAS Enterprise Miner with its descriptive and predictive modelling toolkit, for a range of vertical sectors including finance and risk management and fraud detection, retail, telecommunications, health and insurance, education and high-tech manufacturing;
    - *Think Enterprise Data Mining* for telecommunications.

All the products and services offered by the companies mentioned above make use of a wide array of language (reference) datasets, like ontologies, thesauri, terminological databases, raw and annotated domain specific and/or general language corpora, for indexing/classification of data as well as for the annotation of entities extracted and linking these entities to the reference datasets. Examples of such reference data include the MeSH (Medical Subject Headings) thesaurus[41] or the UNIPROT[42] knowledge base for protein sequence and functional information, among others, if the vertical application domain treated refers to the health and medical sector.

In tandem, tools like part-of-speech taggers for annotating each word in a document with its part-of-speech, and syntactic analysers or parsers, for segmenting sentences into syntactic units

[40] For a brief account of EU commercial players in the field of Text and Data mining, see section 3.3 of the current report.
[41] https://www.nlm.nih.gov/mesh/
[42] http://www.uniprot.org/

complete the set of language resources necessary for carrying out content analysis with a view to mining new information from content.

Language resources, datasets and language processing tools, are made available and accessible through the web pages of the developing organisations, or through specific data centres, like the Linguistic Data Consortium (LDC)[43] in the US. Language processing software is made available either in downloadable form or in the form of web services accessible through e.g. infrastructural initiatives like the LAPPS GRID[44] in the US or the Language Grid[45] in Asia.

[43] https://www.ldc.upenn.edu/
[44] http://www.lappsgrid.org/
[45] http://langrid.org/en/index.html

3 Text and Data Mining In Europe

Text and Data Mining has already been established internationally as a method (with the relevant techniques, tools and technologies) for understanding and adding value to data, as proved by the increasing number of publications on the field, and also patents. According to [Filippov, 2014][46], this growth is driven by the USA and Asia (mostly China), whereas Europe does not follow so closely. Europe's research community as well as the industry are well aware of the role of ***Data Analytics*** and ***Text and Data Mining***, both as users and as researchers of the specific field, either developers of such technologies or data providers, curators, or distributors.

In order to detect and present the European landscape of ***TDM*** we have tried to address the following questions both macroscopically and microscopically (per economic sector):

- What is the content ***TDM*** is aiming at? What are the particular characteristics of the data?
- What are the existing ***TDM*** tools and technologies available to all economic sectors?
- What is the scientific production of ***TDM***?

In the first part of this section (sections 3.1-3.3) we refer to data (including research publications) for ***TDM***, tools and technologies for ***TDM*** (including a discussion of Europe's digital readiness) and, finally, to commercial activity around ***TDM***, with reference to European technologies mainly, but not exclusively.

In the second part of this section (section 3.4) we zoom in on parts of this landscape by focusing on particular areas/economic sectors and present a few indicative measures of publications and research projects carried out by research and commercial organisations in Europe with EU funding.

3.1 Data for mining in Europe

3.1.1 Overview of data sources

As is the case worldwide, data of interest for mining also in Europe reside in:

- **The commercial-private sector**: the current practice for companies providing services to the public, for example EDFENERGY[47], UK's producer of low-carbon electricity or the Athens Water Supply and Sewerage Company (EYDAP)[48], is to collect data, mine it and use the results for its strategic planning, to optimise internal processes, to economise on use of resources of all types, to perform predictions etc., in order to offer improved services and to increase profits. Besides companies that collect data internally for purposes as the ones specified above, there are companies whose business objective is data collection; they

[46] Filippov, S. (2014). *Mapping Text and Data Mining in Academic and Research Communities in Europe.* Brussels: Lisbon Council. Available at: http://www.lisboncouncil.net/component/publication/publication/109-mapping-text-and-data-mining-in-academic-and-research-communities-in-europe.html

[47] https://www.edfenergy.com/

[48] https://www.eydap.gr/en/

record data of all types in order to develop, train and evaluate their ***Data Mining*** and ***Analytics Products***, or to exploit it for Marketing Research or for consultancy services, or simply to sell the data (raw or analysed) to interested parties. Medical and personal data, climate measurements, surveillance video and pictures, scientific data and publications, business transactions, measurements from satellites, software engineering data, reports and memos, e-mail messages and social media texts, all constitute data collected, stored, curated and maintained as valuable assets in the ***Data Mining*** market.

- **The public sector**: the public sector is one of the major producers of data (structured and unstructured). A major movement for open data in the European Union (EU) was observed as a result of a first directive issued in 2003 on the reuse of Public Sector Information[49], as well as its updates in 2013[50]. This Directive (known as "the PSI Directive") provides a common legal framework for government-held data (public sector information). It focuses on the economic aspects of reuse of information rather than on the access of citizens to information, and encourages the Member States to make as much information available for reuse as possible. It addresses material held by public sector bodies in the Member States, at national, regional and local levels, such as ministries, state agencies, municipalities, as well as organizations funded for the most part by or under the control of public authorities. Since 2013 content held by museums, libraries and archives falls within the scope of application of this Directive as well, which covers data such as written texts, databases, audio files and film fragments; it does not apply to the educational, scientific, and broadcasting sectors.

  This Directive has given a tremendous impetus to the trend for open data, appropriately organised and stored and available for re-use. Open data portals are maintained by national governments[51] as well as by the EU's Open Data Portal[52]. They contain data (or links to data) which are open for use, made available by authorities, public bodies and organisations. Data included vary from language data such as Translation Memories and Thesauri to Digital Maps to Producer prices in industry, domestic market at a monthly rate, etc.

  The degree of openness and the rights of use differ; users might be allowed to simply download a dataset, or access the data through an interface or even need to sign an (open) licence agreement. In the EU Open Data portal, which is the single point of access to data produced by the institutions and other bodies of the European Union, the data contained is free to use, reuse, link and redistribute for commercial or non-commercial purposes.
- **The scientific research domain**: available data for ***TDM*** are research publications and recently also research data (e.g. measurements of experiments and studies, annotated data in linguistic research, medical, social, historical, space data etc. However, the availability of the actual scientific data is often hampered by legal, ethical or technical restrictions;

[49] http://www.epsiplatform.eu/content/eu-psi-directive-200398ec
[50] https://ec.europa.eu/digital-single-market/en/european-legislation-reuse-public-sector-information
[51] For a list of government portals, see https://open-data.europa.eu/en/about
[52] https://open-data.europa.eu/en/data

therefore, valuable sources concerning the data are related publications. Scientific publications are deposited at **repositories**, wherefrom they can be retrieved and reused.

- **Institutional repositories** are digital archives that collect, preserve and disseminate the intellectual output of mainly academic and research institutions. Until recently the institutional repositories stored publications, reports, journal articles, books etc. Current practice includes the storage and curation of research data as such; this is used (and re-used) for mining, training or evaluation of tools and also re-purposed, i.e. data produced in the framework of a research field might be reused in another scientific domain with different goals; for instance, social data produced within a social sciences experiment can be valuable for linguistic research. This led to the development of **interdisciplinary data repositories**, providing valuable insights on data concerning interdisciplinary research.
- **Aggregators** and **Research Infrastructures**: **Aggregators** harvest existing repositories, providing unique points of access to a great variety of repositories and their data.
  The recent past has witnessed the development of **Research Infrastructures (RIs)**, an important pillar in EU research[53]. **Research infrastructures**, provide facilities, resources and related services used by the scientific community to conduct top-level research in their respective fields, ranging from social sciences to astronomy, genomics to nanotechnologies[54]. The RIs consist of repositories and aggregators, coupled with user services for the facilitation of data identification, access and processing; crucial are the RIs' standardization efforts concerning the adoption of common practices across scientific domains. The importance of RIs lies not only in the expertise of human resource or the technology and tools developed and used but, most notably, in the volume of data collected from experiments, measurements and observations.

  Research Infrastructures and **aggregators** are domain-specific or interdisciplinary. Indicative cases of interdisciplinary aggregators are:

  OpenAIRE[55], an EU-wide infrastructure, developed a repository facility and scientific data management services, as well as an e-Infrastructure for accessing scientific publications. It is integrated with **Zenodo**[56], a digital repository for everything not served by a dedicated service, that enables researchers, scientists, EU projects and institutions to share, preserve, showcase and share multidisciplinary research results (data and publications) of any size, any format and from any science.

  CORE[57] (COnnecting REpositories)'s mission is to aggregate all open access research outputs from repositories and journals worldwide and make them available to the public. It aims to facilitate free access and reuse of open access research outputs distributed across many systems, providing services for different stakeholders including academics and researchers, repository managers, funders and developers.

[53] See annex A for a detailed listing of EU funded RIs
[54] http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=what
[55] https://www.openaire.eu/
[56] http://zenodo.org/
[57] https://core.ac.uk/

EUDAT[58], the European Data Infrastructure is creating a pan-European infrastructure for e-science to enable European researchers and practitioners from any research discipline to preserve, find, access, and process data in a trusted environment, combining numerous community-specific data repositories with the permanence and persistence of some of Europe's largest scientific data centres.

The existence of many infrastructures and data sharing projects in various domains attest the increasing interest in data sharing activities and evidences the realization by the research community of the fact that data production, collection, curation and maintenance is a task that benefits from openness and sharing.

Additionally, the demand for efficient dealing with the increasing amount of data has put to the spotlight a series of technical issues, such as

- architectural issues, concerning the design of the infrastructures. Infrastructures offer distributed repository networks, IaaS Services (Infrastructure as a Service), huge storage facilities on the cloud, Virtual Machines and Virtual Networks;
- the need for well-designed, well-documented and easy to use workflows that can be executed on data; either raw data generated by sensors and software systems or derived data produced as outcomes of processing;
- the need for interoperability, which can often be a very demanding requirement, especially in the interdisciplinarity framework. Interoperability is the ability of multiple systems to exchange information, integrate components and use mutually intelligible data formats. It is clear that interoperability affects the whole design of a system, especially as regards the mutual integration of resources and tools between systems in view of a broader infrastructure ecosystem; the issue of interoperability is interrelated with the existence and the adoption of common standards and best practices[59]
- as a result of the above, the need for comprehensive but interoperable metadata in appropriate formats (most commonly, RDF[60]).

The map of domain specific aggregators, infrastructures and data sharing projects includes (among others):

Medical/life science domain: Elixir[61], a pan-European research infrastructure that manages and safeguards the massive amounts of data being generated every day by publicly funded research. Two initiatives on the human brain, namely, the **Human Brain Project**[62]**,** an EU initiative for the modelling of the human brain and **Collaboration in European Neurotrauma Effectiveness Research in Traumatic Brain Injury**, an EU project sharing information among 60 hospitals and 38 science infrastructures, collecting patients' data in order to facilitate prediction and medical treatment. **Pharmacog**[63], a pan-European private-public partnership of academic institutions, global pharmaceutical companies and five SMEs focuses on drug discovery for Alzheimer's

[58] www.eudat.eu
[59] Monachini et al (2011).
[60] www.w3.org/RDF
[61] www.elixir-europe.org/
[62] www.humanbrainproject.eu
[63] www.imi.europa.eu/content/pharma-cog

disease. Finally, the **European Medical Information Framework (EMIF)**[64], a project to provide a common architecture for sharing data on 48 million patient records.

Environmental domain: LifeWatch[65] provides the e-Science infrastructure underpinning research into biodiversity and ecosystems across Europe. It collects data that can be cross-referenced with other sources, relating to weather and climate, for example, in order to assess the effect of climate change or agricultural practices on biodiversity.

Space domain: The European Earth observation programme **Copernicus**[66], previously known as GMES (Global Monitoring for Environment and Security), provides environmental data crucial for understanding planet and climate change as well as the influence of human activities in these changes. All of this information is freely available – to public authorities, to scientific and commercial users, and to the general public.

Social Sciences and Humanities domain: CESSDA[67], the Consortium of European Social Science Data Archives, is a legal entity (limited company under Norwegian law), having evolved from an infrastructure of the European Strategy Forum on Research Infrastructures (ESFRI) Roadmap in June 2013. It provides data services to the social sciences, offering coordination of European data service providers and facilitating access to resources relevant to the European social science research agenda. **SHARE ERIC**[68], the Survey of Health, Ageing and Retirement in Europe, is an infrastructure offering data on health, socio-economic status and social and family networks from 20 European countries, available to the entire research community free of charge. The European Social Survey European Research Infrastructure (**ESS ERIC**[69]) offers social data based on a cross-national survey conducted across Europe since 2001; it measures attitudes, beliefs and behaviour patterns of diverse populations in more than thirty nations on social, political and moral issues. The ESS data is available free of charge for non-commercial use. **DARIAH**[70], an infrastructure for arts and humanities scholars, operates through its European-wide network of Virtual Competency Centres, each of them centred on a specific area of expertise; namely, *e-Infrastructure*, *Research and Education Liaison*, *Scholarly Content Management* and *Advocacy, Impact and Outreach*.

Linguistic Infrastructures: Language being the main dimension for ***Text Mining*** across disciplines, European language infrastructures are listed separately. Their significance lies in their outcome: making available methodologies, techniques, tools and services that can be used for mining across sectors (e.g. a part-of-speech tagger for medical texts or a tokenizer and sentence splitter for chemical texts) but also data and other resources for training and evaluating tools and services. **CLARIN**[71], the Common Language Resources and Technology Infrastructure, provides access for scholars in the humanities and social sciences to digital language resources, i.e. it provides links to data (in written, spoken, or multimodal form), and to

[64] http://www.emif.eu/
[65] www.lifewatch.eu
[66] http://www.copernicus.eu/
[67] http://cessda.net/
[68] http://www.share-project.org/
[69] http://www.europeansocialsurvey.org/
[70] http://www.dariah.eu/
[71] www.clarin.eu

processing tools. It acts as an aggregator harvesting metadata on language resources residing in repositories across Europe, covering all European languages. **META-SHARE**[72], is a European infrastructure for sharing and exchanging language data, tools and related web services[73]. It is designed as a network of distributed repositories of language resources (LRs), which makes available datasets documented with a common metadata schema[74] as well as language processing services. It is devoted to the sustainable sharing and dissemination of Language Resources for the Human Language Technologies domain but also for all domains where language plays a critical role.

RIs are a treasury of data; the notion of data has evolved from referring to **a means for doing research** to an **independent entity in its own right**. This evolution of data to an asset for the digital world calls for and underscores the need for:

- open access supported by a comprehensive openness inspired legal framework, and
- interoperability for enhanced data and service discovery and execution, supported by **flexible or adaptable** metadata schemes and formats.

The importance of the availability of data documented by metadata but also the development of relevant tools and the support by computing technologies is underlined in the EU's Strategy Report on Research Infrastructures[75]: "The research process is becoming more and more dependent on the advanced analysis of **large amounts of data**, often acquired in a very short time, and on the availability of effective on-line analysis and high throughput, high performance computing and the strongly emerging shift towards cloud computing. **It is crucial that the data and the necessary contextual information for exploiting the metadata are readily available across the network to remote users.**"

The RIs adhere to the **Open Access policies** following the principles of Open Science[76] in the EU, the Open Access pilot actions[77] launched in both the FP7 and Horizon 2020 EU programmes and the Open Access to scientific information[78] initiative, aiming to maximise the impact of scientific results in the Digital Single Market[79].

3.1.2 Research publications in Europe

According to [Filippov, 2014], the global academic and research community produces over 1,5 million new scientific articles annually, while the total number of articles in circulation as of 2010 was estimated to 50 million. Of these huge numbers only a small percentage concerns ***Text and Data Mining***, as attested by a survey conducted by the same study. In order to quantify the use of ***Text and Data Mining*** as topic of scientific publications (but not as a methodology for other scientific purposes), Filippov consulted ScienceDirect (database operated by Elsevier)

[72] http://www.meta-share.org/
[73] Piperidis (2012)
[74] Gavrilidou et al (2012)
[75] https://ec.europa.eu/research/infrastructures/pdf/esfri-strategy_report_and_roadmap.pdf
[76] http://ec.europa.eu/research/openscience/index.cfm
[77] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf
[78] https://ec.europa.eu/digital-single-market/en/open-access-scientific-information
[79] http://ec.europa.eu/priorities/digital-single-market_en

which contains 11 million articles. Among those, only 1500 articles (approximately) were devoted to ***Text and/or Data Mining***.

This global picture is complemented by a similar survey conducted by [Tsai, 2012][80] , focused on ***Text and Data Mining*** in the social sciences: from 1989 to 2009, 1181 publications were identified having ***Data Mining*** as their topic in the Social Science Citation Index (SSCI) database. From the total 1181 publications, those originating from EU member states amount to 26.3% of the sample.

The CORE database allows full-text mining of research literature, as documented by [Knoth and Zdrahal, 2012]. Deploying the CORE searchable index of full-text documents, Knoth and Herrmannova retrieved articles relevant to ***Text Mining***: on a sample of 2 million full text documents of publications, they identified those containing the term ***Text Mining*** anywhere within the title, abstract or full-text of the paper, extracted the country of affiliation of the authors using the suffix information from email addresses and proceeded to analyse the data from various perspectives.

As regards the European scientific production, the majority of the publications retrieved contain at least one European author. From these publications, the vast majority of authors come from the UK (Figure 4). While the breakdown seems to be consistent with the expected behaviour and also with the observations of the META-NET White Papers (see section 3.2.3), the fact that the UK appears to produce most of papers on ***Text Mining*** in the EU can be accredited to the better coverage (in the CORE database) of the UK compared to the rest of the EU. Additionally, it could be argued (no support data is available, though) that the change in UK's copyright law, effective from June 2014, has alleviated legal constraints for ***Text and Data Mining***, thus resulting in greater numbers of research publications in the country. This change specifies that "copying content from online journals or other texts for the purposes of non-commercial research is no longer an infringement of UK copyright laws providing copiers have lawful access to that content and they, generally, make "a sufficient acknowledgement" of the original work"[81].

[80] Hsu-Hao Tsai. (2012). Global Data Mining: An Empirical Study of Current Trends, Future Forecasts and Technology Diffusions. *Expert Systems with Applications*., 39(9), pp. 8172–8181.

[81] http://www.out-law.com/en/articles/2014/june/researchers-given-data-mining-right-under-new-uk-copyright-laws/

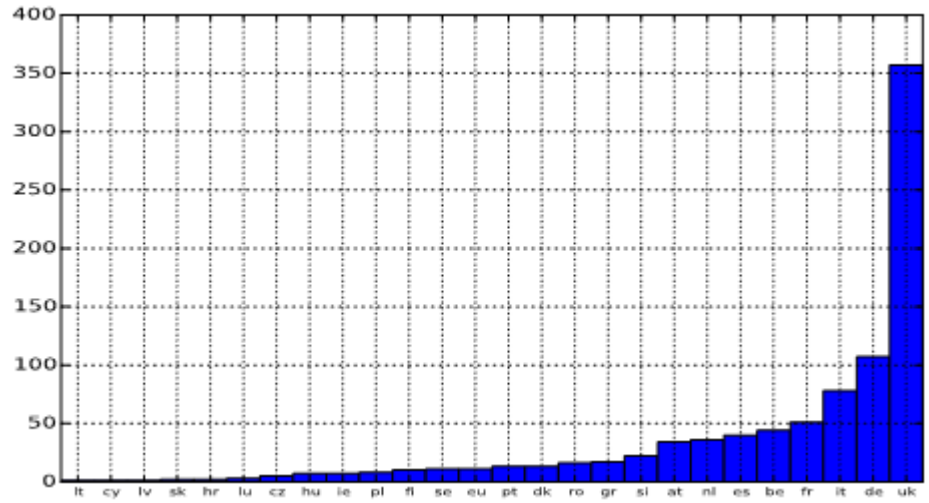


Figure 4. The distribution of papers on Text Mining in the EU per country (from CORE)[82]

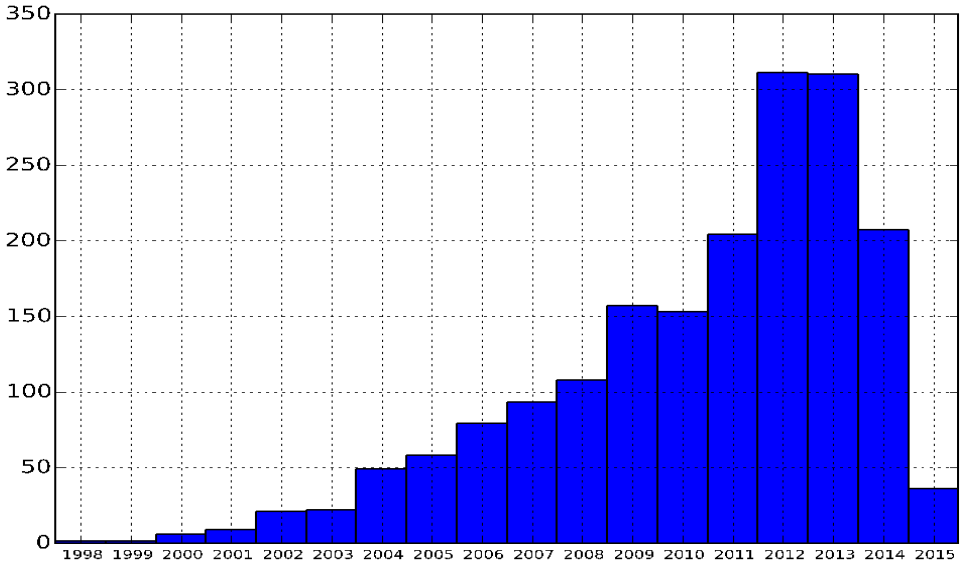


Figure 5. The growth of papers on Text Mining in the CORE dataset by year

The data (Figure 5) indicates that ***Text Mining*** started to be popular in 2001 and has been growing since then **(years 2014-2015 should be treated cautiously, as the data is still incomplete)** at a slightly faster rate than the growth of all research papers (Figure 6).

[82] The authors would like to express their gratitude to Petr Knoth and Drahomira Herrmannova for carrying out these measurements on the CORE dataset.

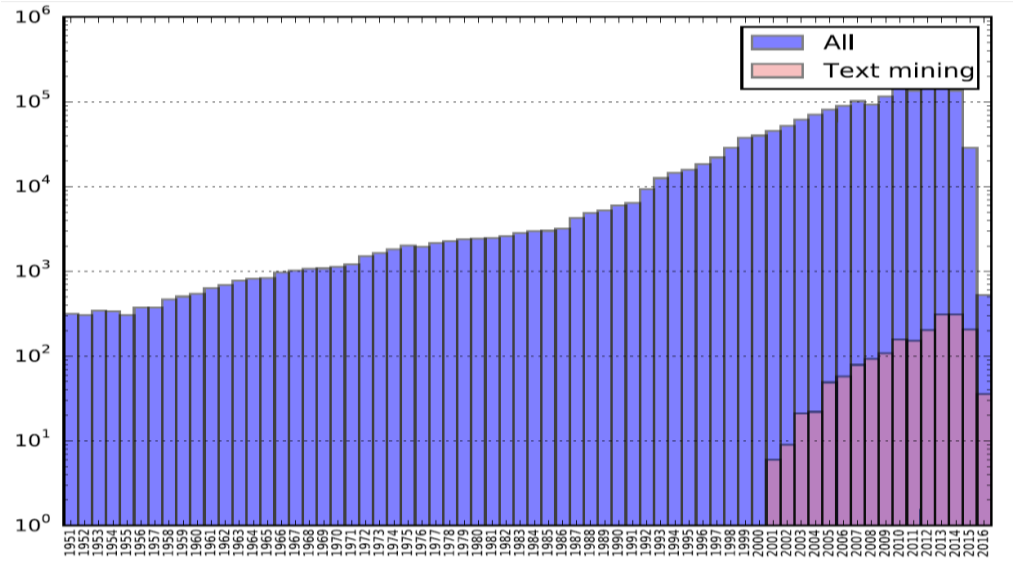


Figure 6. The growth rate of Text Mining publications in comparison with the total in the CORE dataset

3.2 TDM technology in Europe

This section presents tools and services available for ***TDM*** in Europe. The discussion takes into account that a ***TDM*** user can be either a developer (i.e. a skilled programmer) or an end-user, a researcher not necessarily digitally literate. ***TDM*** users may belong to scientific or commercial domains which deal with high volume of data (raw, analysed, textual or in other media, etc.) which demand the use of ***TDM***.

Tools and technologies for ***TDM*** can be language-independent or language-dependent; they can also be specialized to a particular domain. However, not all languages are supported equally well. English claims models and resources for almost all software packages; further well-supported languages include German, French, Spanish, Chinese and Arabic, followed by many languages with limited support (more on this in section 3.2.4).

A large number of tools are open-source; these are, naturally, popular in the research domain. On the other hand, there are also proprietary tools, technologies or frameworks for ***Text and Data Mining***. Both open-source and proprietary make use of (reference) language datasets such as thesauri, terminological and lexical resources, glossaries, ontologies and language corpora, for the tasks of named entity recognition and annotation or for data indexing and classification.

The following sections present an overview of tool and technologies developed and used in Europe for ***Text and Data Mining*** purposes. Given that reference datasets used in ***TDM*** are mostly specialised, domain-specific language resources (e.g. medical term lists, agricultural glossaries, meteorological ontologies etc.), the detailed description of such resources is included in the relevant sections on each sector, where the different domains belong, as defined in the FutureTDM Deliverable D3.1 *Research Report on TDM Landscape in Europe*[83] (see also Annex B of the present report). Thus, the Agrovoc thesaurus which covers the areas of food, agriculture,

[83] http://project.futuretdm.eu/wp-content/uploads/2016/05/D3.1-Research-Report-on-TDM-Landscape-in-Europe-FutureTDM.pdf

fisheries, etc. is listed among the primary sector resources, the ChemSpider database providing information on chemical structures and their properties is listed in the secondary domain and similarly for all included resources (see present report, section 3.5 and onwards).

3.2.1 Available tools for TDM

This section focuses on two types of tools central for ***TDM***, namely **workflows** and **annotation editors**.

Almost all ***Text Mining*** applications are in the form of **workflows** of operational modules, whereby each module's output is the input to the next module. Some modules can be common to many pipelines (e.g. sentence splitting), whereas other modules are task specific. Obviously, these modules need to be interoperable and based on a common structure. An indicative list of workflows for TDM is presented in Figure 7.

Name	Implemented in
Alvis[84]	Java[85]
Apache cTAKES[86]	Java / UIMA[87]
Apache OpenNLP	Java
Apache UIMA[88]	Java
Argo[89]	Java / UIMA
Bluima[90]	Java / UIMA
ClearTK[91]	Java / UIMA
DKPro Core[92]	Java / UIMA
GATE Embedded[93]	Java
Heart of Gold[94]	Java + Python[95]
JCoRe[96]	Java / UIMA
NLTK[97]	Python

Figure 7. Workflows for TDM[98]

The software packages mentioned above, include analytics tools (standalone NLP tools), collections of components, workbenches with graphical user interfaces, as well as interoperability frameworks, though most of them can be multiply classified.

[84] https://migale.jouy.inra.fr/redmine/projects/alvisnlp
[85] https://java.com/en/
[86] http://ctakes.apache.org
[87] https://uima.apache.org/
[88] https://uima.apache.org
[89] http://argo.nactem.ac.uk
[90] https://github.com/BlueBrain/bluima
[91] https://cleartk.github.io/cleartk
[92] https://dkpro.github.io/dkpro-core
[93] https://gate.ac.uk
[94] http://heartofgold.opendfki.de/
[95] https://www.python.org/
[96] http://julielab.github.io
[97] http://www.nltk.org/
[98] OpenMinTeD Deliverable 5.1

Among the interoperability frameworks, the **Apache UIMA** framework and the **GATE** framework appear to be the strongest and most widely used. UIMA[99], IBM's middleware architecture for processing unstructured information, provide powerful search capabilities and a data-driven framework for the development, composition and distributed deployment of analysis engines. GATE, on the other hand, is an open source software that boasts a mature and extensive community of developers, users, educators, students and scientists and provides a process for creating robust and maintainable text processing workflows.

However, it is not the case that the UIMA-based software packages are directly interoperable with each other: the same concepts (e.g., tokens or sentences), have different names and often different properties and relations to each other.

As shown in Figure 7, most of the software is implemented in Java, fact which assists interoperability, both across most hardware and operating system platforms (e.g., Windows, Linux, OS X) and between different software packages. The above list is not exhaustive but it represents the NLP-related software available for ***TDM*** research, either as open source software or via web services.

In ***Text Mining***, **annotation editors** are tools used for editing annotations in text and for their visualization. Annotation editors are central to many ***TDM*** tasks, such as the creation of corpora for system training and evaluation; visualization and error analysis of application output; manual correction of automatically created annotations; and validation of annotations prior to release for use. Tools falling under the heading *annotation editors* may cover various levels, notably text level annotations, document level metadata, and/or annotations using external resources such as ontologies, thesauri, gazetteers etc. Annotation editors typically constrain the end user to create and edit annotations according to some fixed schema or type system. Some editors also allow for ad-hoc, schema-less annotation, and others support ontology-backed schemas. An indicative list of annotation editors is in Figure 8.

Name	Parent framework
Argo (Manual Annotation Editor)[100]	Argo
WebAnno[101]	-
GATE Developer[102] and GATE Teamware[103]	GATE
GATE Teamware[104]	GATE
Brat[105]	-
Alvis AE[106]	Alvis
Egas[107]	-
WordFreak[108]	-

[99] Ferrucci and Lally (2004)
[100] http://argo.nactem.ac.uk/tag/manual-annotation-editor/
[101] https://webanno.github.io/webanno/
[102] https://gate.ac.uk/gate/doc/
[103] https://gate.ac.uk/sale/tao/splitch25.html
[104] https://gate.ac.uk/sale/tao/splitch25.html
[105] http://brat.nlplab.org/
[106] http://www.quaero.org/module_technologique/alvisae-alvis-annotation-editor/
[107] https://demo.bmd-software.com/egas/

Tagtog[109]	-
PubTator[110]	-
MMAX[111]	-
Knowtator[112]	Protege up to 3.5

Figure 8. List of annotation editors[113]

Besides these tools, there exists a large number of tools and resources specifically designed for use within particular domains, which could be re-used, however, in additional domains. An indicative list of such tools and resources is presented in Figure 9.

Name	Initial purpose	Example usage domains
ELKI[114]	research & teaching in Data Mining	cluster benchmarking
Galaxy[115]	genomics research	bioinformatics
Kepler[116]	scientific workflow management system	bioinformatics, data monitoring,
KNIME[117]	Data Mining (pharmaceutical research)	business intelligence, financial data analysis
Pegasus [118]	scientific workflow management system	astronomy, bioinformatics, earthquake science
Pipeline Pilot[119]	pharmaceutical and biotechnology	Chemicals, Energy, Consumer Packaged Goods, Aerospace
Taverna [120]	scientific workflow management system	bioinformatics, astronomy, chemo-informatics, health informatics
Triana[121]	gravitational wave studies	signal processing

Figure 9. List of tools and workflow engines built for specific domains[122]

3.2.2 Tool hosting facilities

The discovery of tools and services is facilitated by **registries**, which maintain relevant metadata for their documentation; registries usually do not host the tools/services themselves. The hosting of actual tools/services is the task of **repositories**, while the interaction between tools/services and resources is the task of **platforms**, which enable running of web services either on data included in the platform or uploaded by the user. The following table presents an

[108] http://wordfreak.sourceforge.net/
[109] https://www.tagtog.net/
[110] http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/index.cgi?user=User284144660
[111] http://mmax2.sourceforge.net/
[112] http://knowtator.sourceforge.net/
[113] OpenMinTeD Deliverable 5.1
[114] http://elki.dbs.ifi.lmu.de/
[115] https://galaxyproject.org/
[116] https://kepler-project.org/
[117] http://www.knime.org/knime
[118] https://pegasus.isi.edu
[119] http://accelrys.com/products/collaborative-science/biovia-pipeline-pilot/
[120] http://www.taverna.org.uk/
[121] http://www.trianacode.org
[122] OpenMinTeD Deliverable 5.1

indicative list of registries, repositories and platforms for tools and services (Figure 10), most of which are based on open source software.

Title	Category	Domain
ALVEO[123]	Platform	general
AnnoMarket[124]	Platform	general
Language Grid[125]	Platform	general
LAPPS Grid[126]	Platform	general
QT21[127]	Platform	general
BioCatalogue[128]	Registry	life sciences
BiodiversityCatalogue[129]	Registry	biodiversity
CLARIN Virtual Language Observatory[130]	Registry/Repository	social sciences and humanities
LRE Map[131]	Registry	general
META-SHARE[132]	Registry/ Repository	general
LINDAT/CLARIN[133]	Registry/Repository	general

Figure 10. Table of tool hosting facilities[134]

3.2.3 The language dimension in text mining

Given the huge amount of data available over the internet and the broad range of ***Text Mining*** applications that have emerged in the past years, one could suppose that the path for text mining would have become very smooth.

By its very nature, ***Text Mining***, or ***Text Analytics*** as it is mostly referred to today, is strongly intertwined with the intricacies of language itself and necessitates the existence of appropriate tools and resources for each natural language. Even language independent tools may depend on language data for their training; some of them make use of lexical/knowledge resources such as lexica, thesauri, ontologies, gazetteers etc.; others are trained on texts of a specific domain, which renders them inadequate for processing texts of a different domain. In this sense,

[123] http://alveo.edu.au
[124] https://annomarket.com
[125] http://langrid.org
[126] http://www.lappsgrid.org
[127] http://www.qt21.eu
[128] https://www.biocatalogue.org/
[129] https://www.biodiversitycatalogue.org/
[130] https://www.clarin.eu/content/virtual-language-observatory
[131] http://www.resourcebook.eu
[132] http://www.meta-share.org
[133] https://lindat.mff.cuni.cz/
[134] OpenMinTeD Deliverable 5.1

language processing tools necessary for text mining face all problems posed by language. Prominent among these, are

- the issue of multilinguality: mining in more than one languages requests a multitude of resources and tools, not all of which are available; and
- the different manifestations of language(s), i.e. sublanguages of specific domains, text types and language registers.

The combination of the two issues constitutes a great challenge as regards the development of resources and tools for ***Text Mining*** taking into account Europe's linguistic diversity. ***Text Mining*** processes not only properly formed text with correct spelling, grammar and elaborate language, but also (mostly, even) ill-formed text, with incomplete sentences, erroneous spelling and grammar, specialized language, jargon etc.

A typical example of the combination of these two issues (multilinguality and special languages) is the case of languages used in the social media. The two images of Figure 11 depict the different languages used for tweeting around Europe[135] from May to October 2011 (left) and the top 10 Twitter languages in London in summer 2012[136] (right), where a total of 66 different languages were tweeted between March and August 2012 (the Olympic Games hosted by London during the summer of 2012 which resulted in a huge concentration of many nationalities in a specific place at a limited time-frame provided excellent circumstances for the study of these issues).

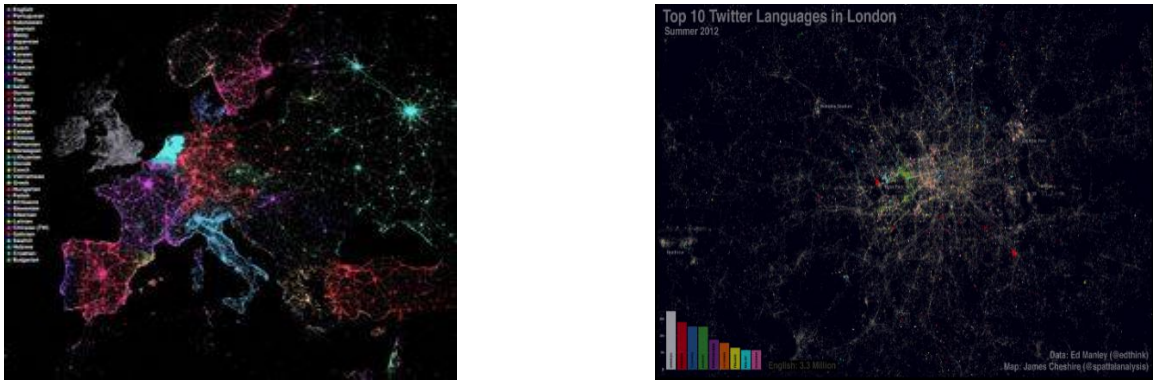


Figure 11. Languages in use on Twitter in Europe (left) and top 10 Twitter languages in London (right)

We get a different picture as regards linguistic diversity in the domain of scientific publications: querying the CORE database (https://core.ac.uk/) for articles on text mining, we observe that the vast majority of papers (almost 90%) are written in English, as shown on the left side of Figure 12. However, a non-negligible amount of scientific publications (especially in domains like Social Sciences, Arts and Humanities) are written in other languages than English.

[135] http://bigthink.com/strange-maps/539-vive-le-tweet-a-map-of-twitters-languages
[136] http://www.theguardian.com/news/datablog/interactive/2012/oct/25/twitter-languages-london-top-ten

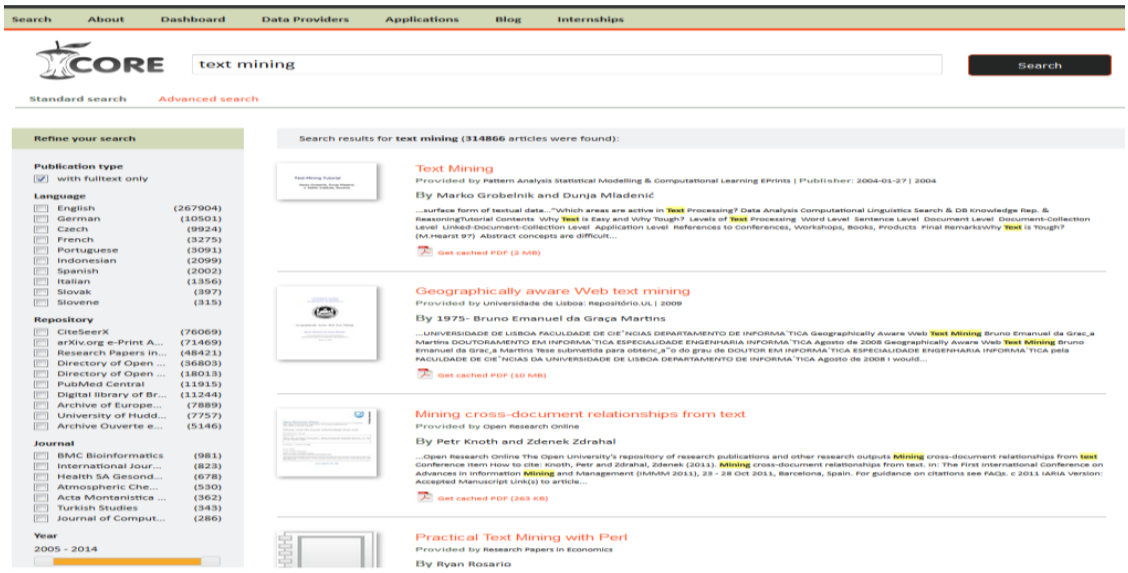


Figure 12. Searching the CORE repository for 'text mining'

Moving further in the realm of ***Text Mining***, natural language processing and computational linguistics, Figure 13 shows the distribution of languages used for actual linguistic research (i.e. languages as objects of research), as attested in the Journal of Computational Linguistics (JCL) and the three major conferences of the area, namely the annual meetings of the Association for Computational Linguistics (ACL), the Conferences on Empirical Methods in Natural Language Processing (EMNLP) and the International Conferences on Computational Linguistics (COLING), in the years 2008-2010.

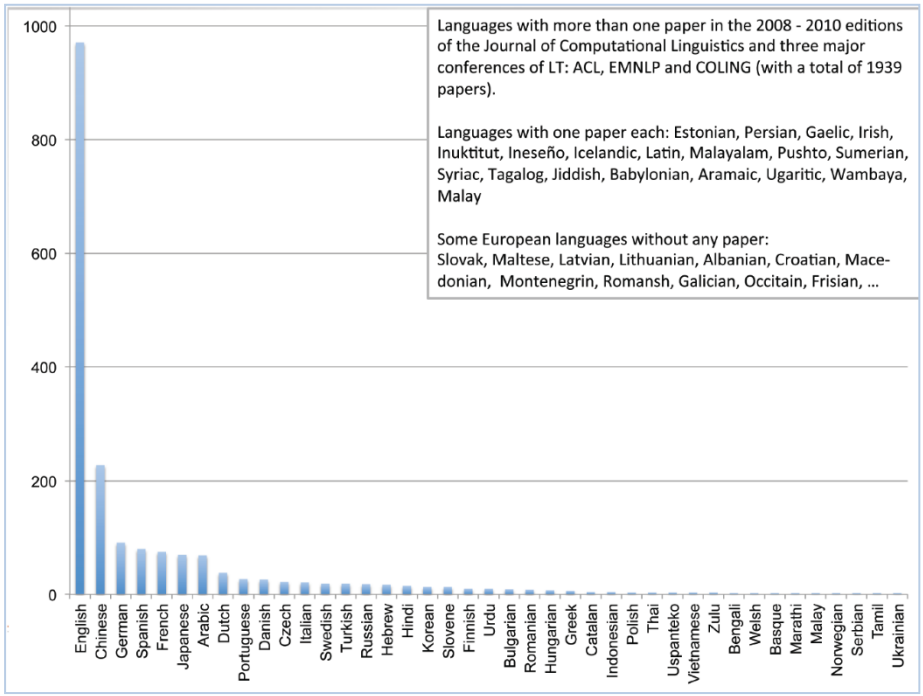


Figure 13. Languages treated in publications in JCL and major LT conferences in 2008-2010

Figure 13 illustrates that research has focused primarily on English, Chinese, German, Spanish and French, while some European languages are either very sparsely covered or not studied at all.

Taking the above into consideration, we pose the question:

How digitally ready are EU languages to support text analytics applications on content written in these languages?

This question is discussed in the following section.

3.2.4 Digital readiness of EU languages

The notion of 'digital readiness' of a language defines the degree at which a natural language is supported as regards ***Language Technology (LT)***; that is, what are the available NLP tools, for which technologies and applications, and how many language resources exist to support technology development for this language.

It is evident that European languages are far from being at the same degree of 'digital readiness'; this came as the outcome of a study conducted in the framework of the META-NET project[137], which is documented in the META-NET White Paper Series[138]. The study, prepared by more than 200 experts and documented in 30 volumes of the META-NET White Paper Series, assessed language technology support for each language in four different areas: machine translation (MT), speech interaction, text analysis and the availability of language resources, in order to measure each language's digital readiness.

	excellent	good	moderate	fragmentary	weak or no support
Machine Translation		English	French, Spanish	Catalan, Dutch, German, Hungarian, Italian, Polish, Romanian	Basque, Bulgarian, Croatian, Czech, Da-nish, Estonian, Finnish, Galician, Greek, Icelandic, Irish, Latvian, Lithuanian, Maltese, Norwegian, Portuguese, Serbian, Slovak, Slovene, Swedish
Text Analysis		English	Dutch, French, German, Italian, Spanish	Basque, Bulgarian, Catalan, Czech, Danish, Finnish, Galician, Greek, Hungarian, Norwegian, Polish, Portuguese, Romanian, Slovak, Slovene, Swedish	Croatian, Estonian, Icelandic, Irish, Latvian, Lithuanian, Maltese, Serbian
Speech		English	Czech, Dutch, Finnish, French, German, Italian, Portuguese, Spanish	Basque, Bulgarian, Catalan, Danish, Estonian, Galician, Greek, Hungarian, Irish, Norwegian, Polish, Serbian, Slovak, Slovene, Swedish	Croatian, Icelandic, Latvian, Lithuanian, Maltese, Romanian
Resources		English	Czech, Dutch, French, German, Hungarian, Italian, Polish, Spanish, Swedish	Basque, Bulgarian, Catalan, Croatian, Danish, Estonian, Finnish, Galician, Greek, Norwegian, Portuguese, Romanian, Serbian, Slovak, Slovene	Icelandic, Irish, Latvian, Lithuanian, Maltese

Figure 14. LT support for EU languages as regards four key areas[139]

As shown in Figure 14, a total of 21 of the 30 languages (70%) were placed in the lowest category, 'support is weak or non-existent' for at least one area by the experts. Several languages, for example, Icelandic, Latvian, Lithuanian and Maltese, receive this lowest score in all four areas. On the other end of the spectrum, no language was considered to have 'excellent support', and only English was assessed as having 'good support', followed by languages such as Dutch, French, German, Italian and Spanish with 'moderate support'. Languages such as Basque,

[137] www.meta-net.eu
[138] http://www.meta-net.eu/whitepapers/overview
[139] http://www.meta-net.eu/whitepapers/overview

Bulgarian, Catalan, Greek, Hungarian and Polish exhibit 'fragmentary support', placing them also in the set of languages which are not yet digitally ready.

Text Mining requires large amounts of written or spoken data, which, in the case of languages with relatively few speakers, it is difficult to acquire. Besides the size of the language speaking community, additional factors hampering digital readiness are the complexity of the respective language and the existence of active research in the public or private domain concerning the specific language.

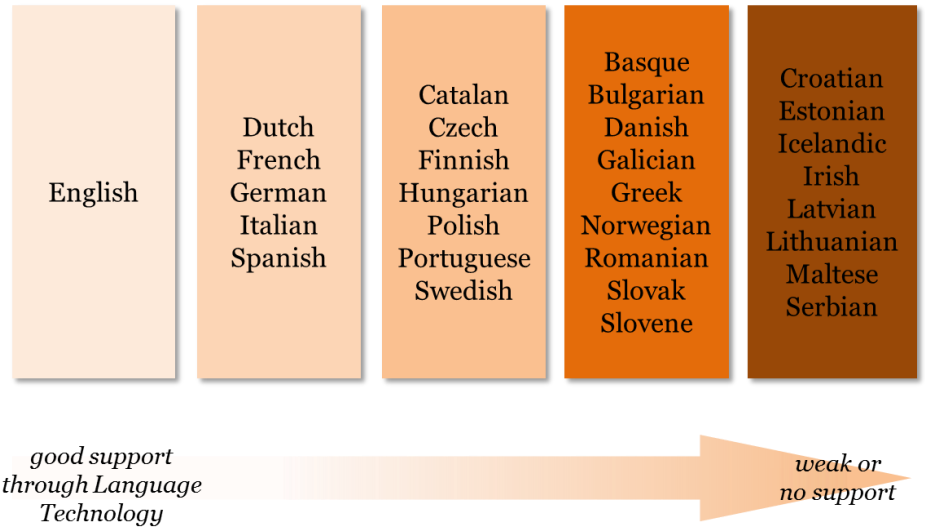


Figure 15. Overall ranking of European languages as regards LT support[140]

The overall ranking of the European languages as regards their support by Language Technology as attested in the META-NET White Paper Series is presented in Figure 15: English benefits from a large speaking community (as a mother tongue/first language and as a second/foreign language), a long tradition of LT, many research teams dealing with it over the world, and, last but not least, abundant funding (also because of US funding). On the other side of the spectrum, reside languages lacking in some (or in all) of these parameters, with the obvious results as regards their digital readiness.

3.3 Commercial activity in Europe around Text and Data Mining

3.3.1 Areas of activity

Text and Data Mining techniques have been used in different business areas, such as business intelligence, knowledge management, customer relationship, marketing, human resource management, security, open-ended survey responses, competitive intelligence or data science in various industry sectors.

The European Data Market study[141] of the European Commission "aims to define, assess and measure the European data economy, supporting the achievement of the Data Value Chain policy of the European Commission." It seeks to present and to support the stakeholders

[140] http://www.meta-net.eu/whitepapers/overview
[141] http://www.datalandscape.eu/

involved in the data market in the EU and to help the transformation of the currently disparate communities to a genuine stakeholders' ecosystem.

In this study, the key components of the data economy active in TDM are presented as regards the number of their members[142] (Figure 16). The **Enabling Players** include Venture Capitalists that provide risk funding to data start-ups; incubators that support their growth; research institutes that enable innovation; training organisations that provide the right skills; and public regulators such as the European Commission that can have a big influence on the development of the sector through for instance its data protection directives. **ICT Enablers** cover all the support tools and technologies (commercial and open source), that act on a more infrastructural level. **Data Marketplaces** include services where data is stored, curated and exchanged. **Analytics** represents arguably the core sector of big data. It includes a wide variety of products such as Analytics platforms, Social analytics, Business Intelligence, Artificial Intelligence, Statistical computing, Machine learning, Visualisation, Unstructured data. **Vertical Applications** includes analytical tools devoted to specific domains such as marketing, legal, government, science, health and finance. **Cross infrastructures** are mainly provided by large vendors and cover many of the functionalities needed in the data market. Finally, the **Data users** include any sector of the economy that can benefit from "data based innovation". It should be noted that many players can be classified into more than one category (i.e. companies that are listed as providing Vertical Applications can also be included in the Analytics category).

Figure 16 represents a snapshot of the field, as captured by this study at the specific time.

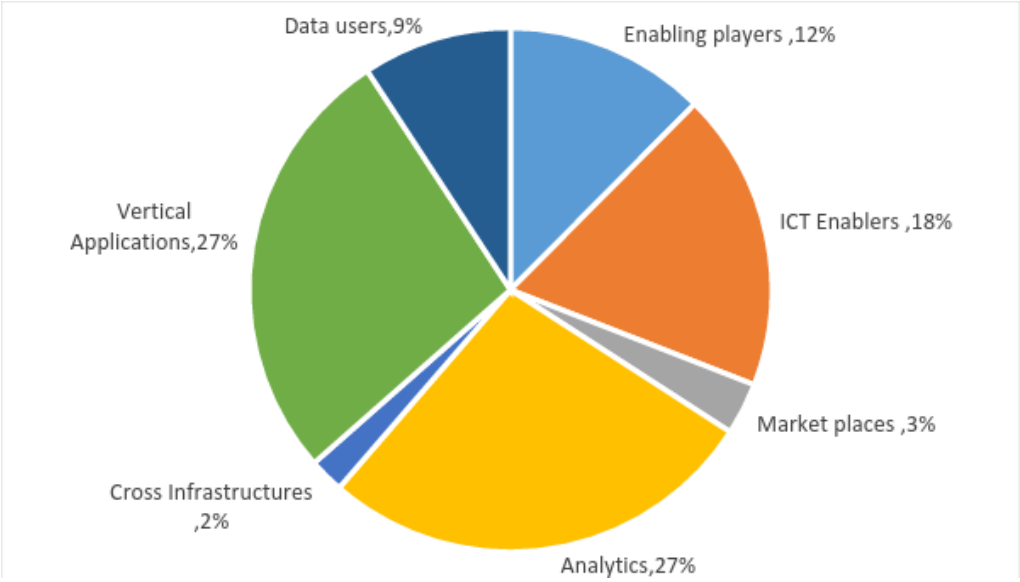


Figure 16. Players active in the Data Market[143]

The larger areas attested (i.e. the categories with the larger number of members) are the areas of **Vertical Applications** and **Analytics**, understandably so, as these two represent the core sector of the big data market.

The market division as regards countries is another perspective on the ***Text and Data Mining***

[142] Source of data: http://www.datalandscape.eu/eu-data-landscape.
[143] Source of data: http://www.datalandscape.eu/eu-data-landscape

market landscape in Europe; UK and Germany hold the leading positions with France, Spain, Italy and Belgium following[144]. It should also be noted that statistics mentioned here refer to companies created and based in Europe, excluding Europe-based departments of international companies.

A valuable insight from this study is the identification of the object of processing on which the various companies put particular emphasis, namely, text or data. Rigid calculations based on the use of the terms **data** and **text** in the description of the objectives of each company indicate that the former accounts for 85% while the latter for 15%. The overwhelming prevalence of data over text can be attributed to the fact that text can be considered as only one of the forms of data, in addition to the trend of coupling text and data in single application offerings, in which **data**, as a more generic term, subsumes **text.**

The tag cloud in Figure 17, that we have created based on the companies' descriptions, highlights the domains of activity of the European private sector in a concise graphical way. In this tag cloud, in absolute accordance with the chart on Figure 17, the most prominent term is ***Data***, while the terms ***Business, Technology, Solutions, Big Data, Information***, and ***Analytics*** are also noticeable.

Figure 17. Tag cloud of EU companies' domains of interest

3.3.2 Skills sought by commercial actors

Indicative of the significance of ***Text Mining*** for the private sector are job advertisements demanding ***Text and Data Mining*** skills. The research was conducted on LinkedIn data per country, focusing on the description of jobs as specified in the advertisement texts [accessed on 19.05.2016]. In order to find the skills requested by companies, a set of key concepts was used starting from ***Text Mining***, ***Data Mining*** and ***Text and Data Mining***, which is the focus of this report, but also extending the research to other related terms. The frequency of terms shown in Figure 18, reveals that the majority of job advertisements in all European countries focuses on

[144] EDM_D6_Interim Report Release October 16 2015_Final, at https://idc-emea.app.box.com/s/k7xv0u3gl6xfvq1rl667xqmw69pzk790 .

Business Intelligence and ***Big Data*** demanding from employees ***Data Analysis*** skills. ***Data Mining*** is also a skill required by companies. However, ***Text Mining*** as a term has a low appearance frequency in job advertisements; this could be explained if we consider the term to be subsumed by the term ***Data Mining***.

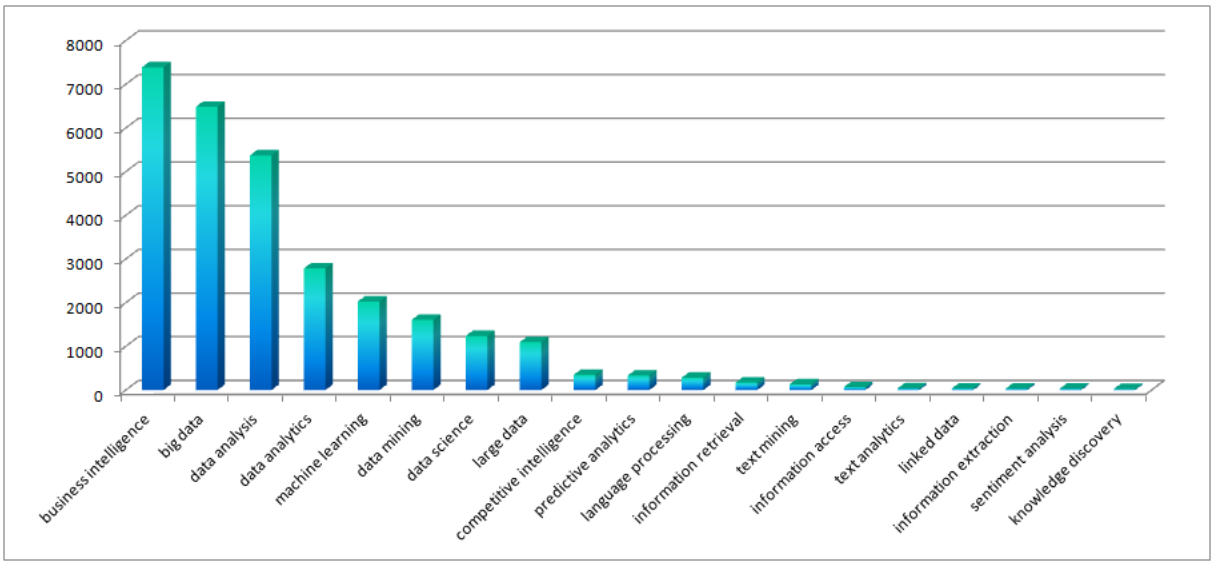


Figure 18. Frequency of key concepts used in job ads in all European countries

As expected (Figure 19) there are more job advertisements related to ***TDM*** in the United Kingdom than any other European country (13,162 job advertisements). Germany comes second, followed by France, the Netherlands and Ireland, all of which have more than 1,000 job advertisements. There are two countries, namely Cyprus and Slovenia, where only 1 job related to ***TDM*** was retrieved.

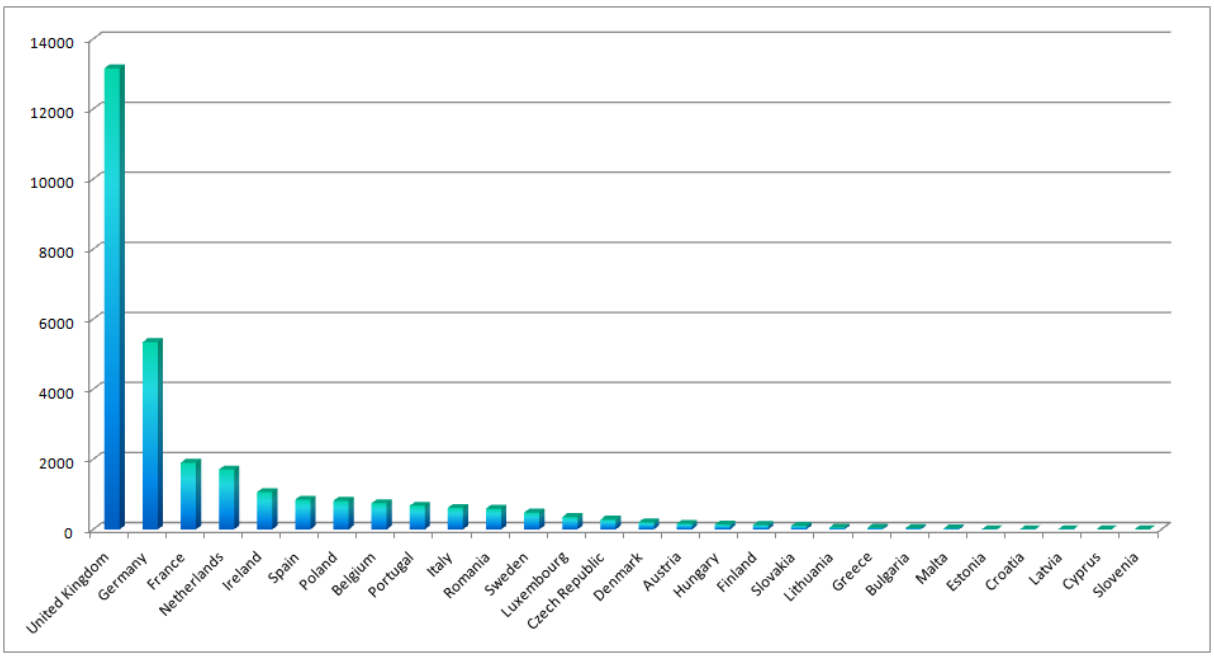


Figure 19. Frequency of job advertisements related to TDM per country

3.4 Research activity, outputs and infrastructures for TDM in Europe

3.4.1 Methodology

In order to depict the landscape of the ***Text and Data Mining*** research in Europe we employed two axes, namely

a. EU funded research projects and infrastructures[145], and
b. scientific articles published in journals and conferences.

These two axes are considered as reflecting the strategic decisions of the EU and the policies adopted as regards the selection of specific research domains for funding, as well as the scientific trends attested in the publications.

3.4.2 EU-funded research projects

This axis of the study concentrated on projects funded within the FP7 and Horizon 2020 framework programmes, as provided by the European Open Data Portal[146]. The time span covered is from 2007 to 2016 (although some of the longer duration projects extend up to 2021).

In order to select only the relevant projects, the concepts and terms mentioned in Section 2.2 (indicatively, ***Text Mining, Data Mining, Analytics***), as well as some of the ***TDM*** application terms were used as query terms[147]. The retrieved European projects were grouped in two big categories:

- projects containing the exact terms[148] considered to be ***TDM* specific**, namely ***Text Mining, Data Mining, Text and Data Mining*****,** and
- projects containing alternative terms[149] considered to be ***TDM* associated**, such as ***Big Data, Data Analysis, Machine Learning*** etc.

The results of the two sets of queries (exact vs related terms) retrieved from a total number of 30,456 projects are represented in Figure 20. A percentage of approximately 3% of the total number of projects are considered to be related to ***TDM***.

[145] Projects which have resulted in the development of infrastructures are presented in the respective sections.
[146] https://open-data.europa.eu/en/data/dataset/cordisFP7projects
[147] For a full list of the terms used and the results of the queries per term, see ANNEX C: Frequency of TDM related terms in FP7 and Horizon 2020 Projects
[148] The exact terms are the three terms presented here in addition to the abbreviation TDM (Text and Data Mining).
[149] The alternative terms are a set of 35 terms. For a full list of all the terms used, please see Annex C.

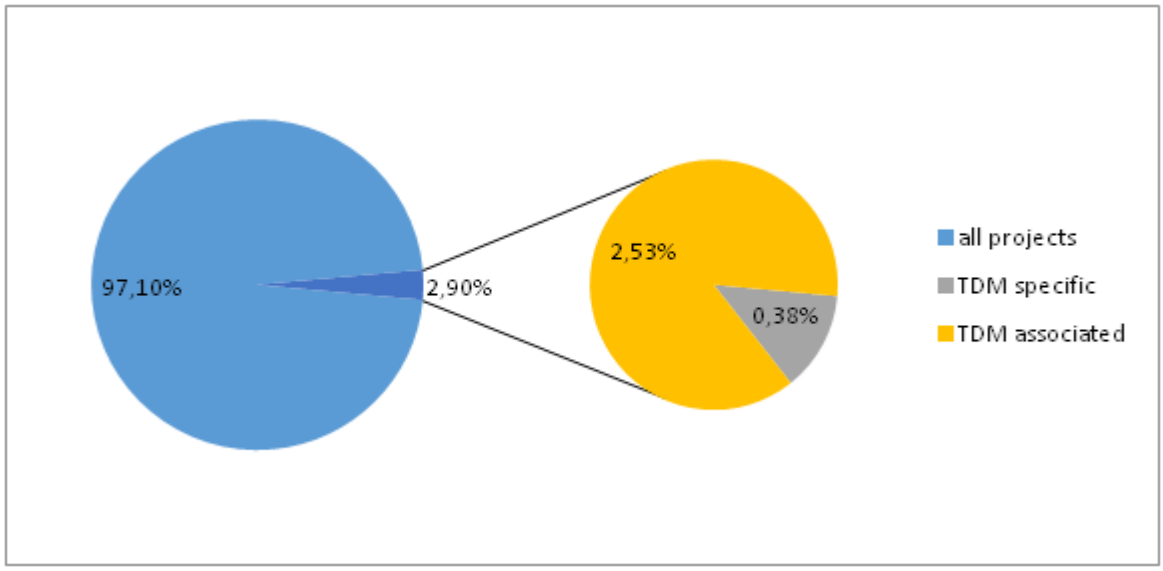


Figure 20. EU FP7 and Horizon 2020 funded projects related to TDM

The tag cloud[150] of Figure 21 illustrates the concepts and terms and their frequency of use in the objectives (and presumably, the conceptualization) of each research project.

Figure 21. Tag cloud of concepts and their frequency in FP7 and Horizon 2020 projects descriptions

The frequency of terms within the FP7 and Horizon 2020 frameworks denotes the research interests and the focus of the European R&D community (public and private, research and commercial) in all domains, concerning ***Text and Data Mining***. The most frequent term in FP7 projects is ***Data Analysis,*** while in the Horizon 2020 projects the prominent term is ***Big Data,*** indicating a shift of interest, which is in accordance with the international trends. The graph in Figure 22 presents the frequency of these terms in FP7 (blue bars) and Horizon 2020 projects (red bars).

[150] The size of each term is analogous to the frequency: the bigger the size, the more frequent the term.

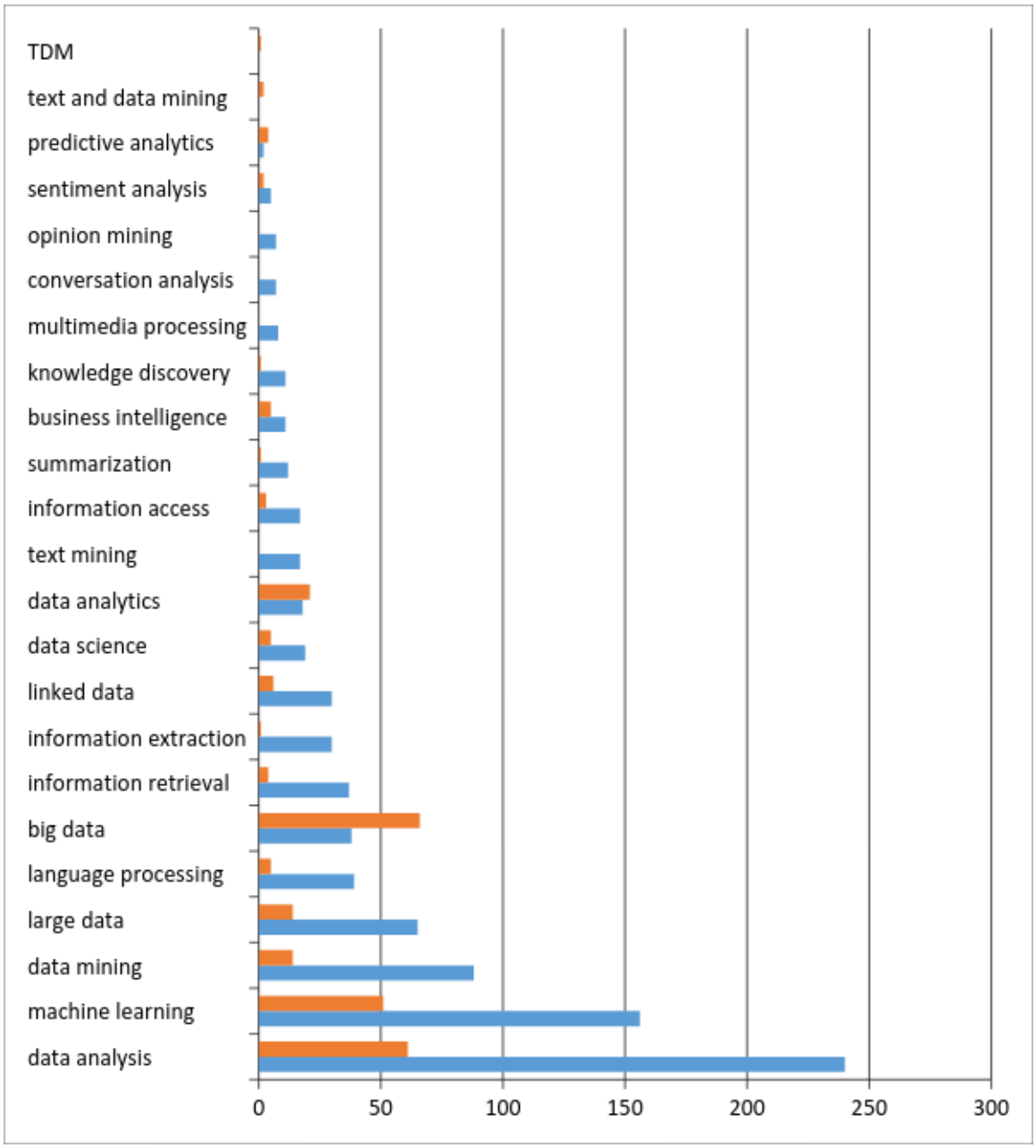


Figure 22. Frequency of terms in FP7 and Horizon 2020 projects

All projects retrieved were further classified to one or more of the economic sectors defined in FutureTDM Deliverable D3.1 Research Report on TDM Landscape in Europe, which organises the various areas of scientific and commercial activity reflecting them onto five economic sectors.

The classification of the projects was based on a mapping of the CORDIS Subject Index Classification Codes and the Horizon 2020 topics to the respective economic sectors. Through this mapping, the projects' classification highlights the economic sector perspective, based on the domain of scientific and commercial activity. It is according to this classification that the European projects are presented in the following sections.

The use of different classification schemes, namely the CORDIS indices, the Horizon 2020 topics and the five economic sectors, entailed a range of semantic interoperability issues:

- The mapping of the scientific areas of the projects to the economic sectors has been made with a certain degree of abstraction, given that there is not always a direct correlation. Some indices have a straightforward mapping to one economic sector, e.g. EDU (Education, Training) to the quaternary sector, while others correspond to two different sectors, e.g. the index ABI (Agricultural Biotechnology) combines Agriculture (which belongs to the primary sector) and Biotechnology (which belongs to the secondary sector).
- Indices used for project classification within each EU research funding framework mainly serve funding purposes; they do not aim at the exact description of the scientific domain of a project or the area of commercial activity[151]. Thus, certain indices were non-informative in what concerned the need for mapping onto thematic domains and relevant economic sectors.
- Some of the projects appear more than once in different subsections of this report due to the fact that multiple scientific fields, and consequently economic sectors, are involved[152].

Figure 23 and the corresponding Figure 24 show the total number of projects that have been funded within the **FP7** and **Horizon 2020** frameworks per economic sector, their total cost and the EU maximum contribution[153].

The **tertiary** sector, covering health care, social care and IT services amongst others, has attracted most attention and investment, with **313** projects totalling an EU funding of **852,4 M€**. The **quaternary** sector, focusing on research, development and education, follows with **273** projects which were granted **319,9 M€** from EU (total cost: **402 M€)**. Note that the investment made in the **secondary** sector, comprising all types of industry and energy, for almost the same number of projects (**253)** is double the investment made in the quaternary sector: **806,4 M€**, of which **633,8 M€** correspond to EU funding.

	Number of projects	Total Investment	EU funding
Primary sector	24	98,2 M€	75,3 M€
Secondary sector	253	806,4 M€	633,8 M€
Tertiary sector	313	1,1 B€	852,4 M€
Quaternary sector	273	402 M€	319,9 M€
Quinary sector	116	192,1 M€	149 M€
All FP7 & Horizon 2020 projects	30456	67,4 B€	51,2 B€

Figure 23. **EU Investment per economic sector**

[151] This is the case, for example, of the METACARDIS project which has been classified under the broad category of SCIENTIFIC RESEARCH, one of the RTD Horizontal Topics, not revealing its scientific area, which is Medicine, and specifically the use of medical knowledge for improved health care services.

[152] Such is the case of the **SENSEI** project, which came with three indices mapping to two different sectors: the **ELM** index (Electronics, Microelectronics) was mapped to the secondary sector while the **IPS** (Information Processing, Information Systems) and **TEL** (Telecommunications) indices mapped to the tertiary sector.

[153] Due to the fact that some projects, as mentioned, appear in more than one sectors, the respective financial investment has been taken into account in both sectors.

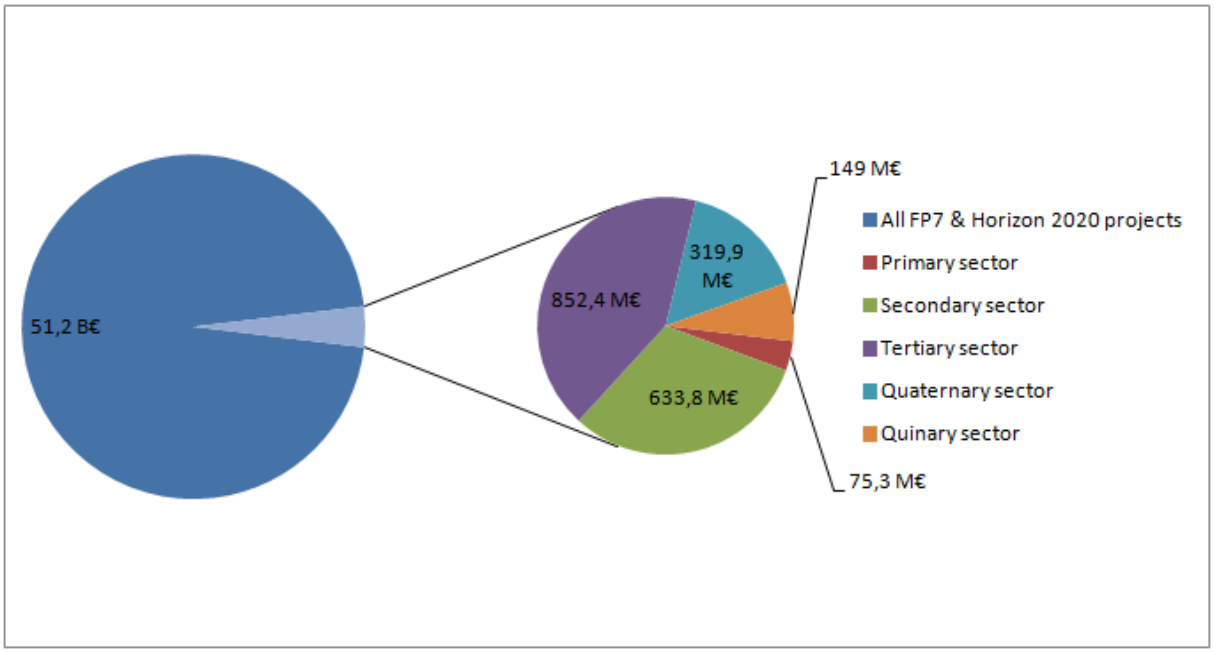


Figure 24. Funding of EU FP7 and Horizon 2020 projects

3.4.3 Scientific publications on TDM across all sectors

In order to estimate the distribution of ***Text and Data Mining*** scientific publications per economic sector we used 1,195,499 abstracts extracted from the latest dumps that have been made available at the **CORE (COnnecting REpositories)** site[154]. From the 1,195,499 abstracts in our dataset, 11504 are related to ***Text and Data Mining***, i.e. 0,96% of the total. This finding is commensurate to the findings of [Filippov, 2014] and [Tsai, 2012][155].

These abstracts were analysed using Latent Dirichlet Allocation (LDA)[156], a probabilistic Bayesian model of text generation.[157] We used the MALLET[158] implementation of LDA and learnt a model of 100 topics from the 1,195,499 abstracts. Each topic *t* (in our analysis) is represented by a small set of words; the ones that are more likely to belong to *t* according to the estimated LDA probabilities (word-topic distribution). Based on the word set, each topic was manually assigned an application area label (indicatively, Topic Id 26 → Aerospace, Topic Id 37 → IT Services) from the respective classification that is described in FutureTDM Deliverable D3.1 Research Report on TDM Landscape**.**

Topic ID	Topic Words
26	ray mass observations emission star galaxies sources formation present dust observed find similar line source density solar high spectral

[154] https://core.ac.uk/intro/data_dumps

[155] Hsu-Hao Tsai, "Global Data Mining: An Empirical Study of Current Trends, Future Forecasts and Technology Diffusions," Expert Systems with Applications, 39(9): 8172–8181, 2012.

[156] LDA assumes that each document collection is generated for a mixture of K topics, and that each document of the collection discusses these K topics to a different extent. It is worth noting that LDA is a bag-of-words model, i.e., it assumes that word order within documents is not important.

[157] Blei et al., 2003.

[158] http://mallet.cs.umass.edu/

37	scheme proposed channel performance based schemes multiple network transmission system error filter channels code complexity time di networks codes

Figure 25. Indicative sets of words per topic

From all abstracts, we kept (for our analysis) only the ones in which a ***TDM***-related term occurs (e.g. ***Text Mining***, ***Text Analysis***, ***Text Analytics***, ***Data Mining***, ***Natural Language Processing***). We used the topic-document probabilities that are returned from LDA and indicate which topics are discussed in each abstract. The sector label (e.g. Aerospace) of the LDA topic with the highest topic-document probability is assigned to each abstract. These assignments were used to estimate the distribution of ***TDM*** papers per application area as shown in Figure 26.

Application Area	No of papers	%
Research	3370	29,29%
Medicine	1952	16,97%
IT services	1713	14,89%
Education	1446	12,57%
Health care	1043	9,07%
Social care	458	3,98%
Engineering	290	2,52%
Finance	245	2,13%
Agriculture	203	1,76%
Public Administration	198	1,72%
Environment	178	1,55%
Aerospace	116	1,01%
Entertainment	104	0,90%
Energy	77	0,67%
Transportation	61	0,53%
Natural Resources	27	0,23%
Construction	15	0,13%
Microelectronics	8	0,07%
Total	11504	100,00%

Figure 26. Distribution of TDM papers per application area

3.4.4 Infrastructures

The European RI ecosystem covers five **research macro-domains**: Energy, Environment, Health & Food, Physical Sciences & Engineering, and Social & Cultural Innovation[159]. These macro-domains have been mapped to the respective economic sectors (Figure 27) and are presented in the following sections.

[159] European Strategy Forum on Research Infrastructures. (2016). *Strategic Report on Research Infrastructures, Roadmap 2016*. Available at: http://www.esfri.eu/roadmap-2016

It should be noted that the quinary sector (Funding and High-level decision) does not appear independently in the graph, given that high level decision making is applied to all macro-domains and, consequently, the corresponding RIs have been subsumed in these.

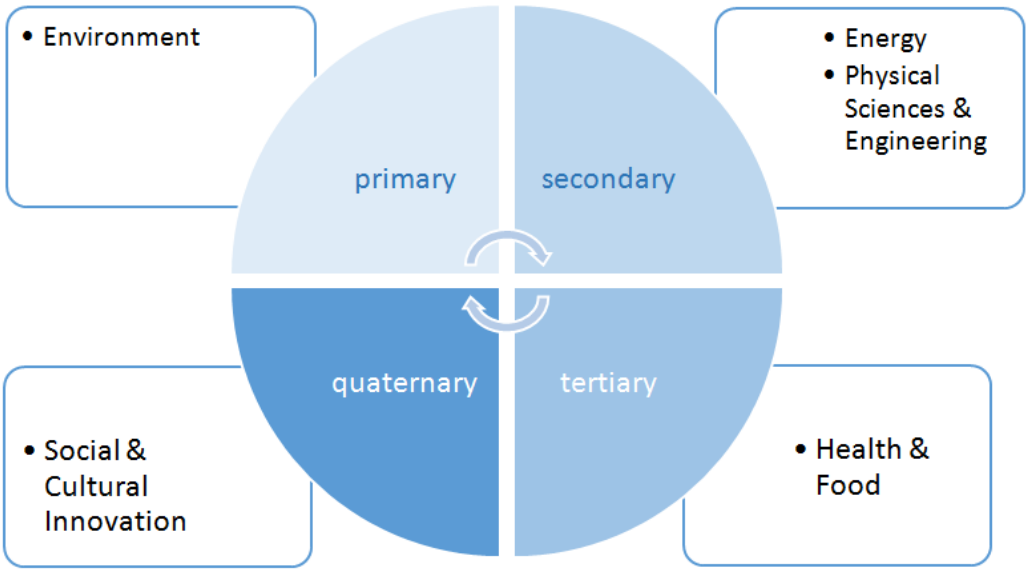


Figure 27. Mapping of RIs macro-domains to economic sectors

In addition, all the FP7 and Horizon 2020 projects which have resulted in infrastructures are mapped to these macro-domains (energy, health, social and cultural innovation, etc.) and are presented in the respective following sections per economic sector.

Projects aiming at the creation, development and maintenance of infrastructures represent the 1,3% of all funded projects[160], while they have received 4% of the total project investment by the EU, fact which indicates the strategic priority that the EU has given to RIs. Specifically, the total investment[161] made in infrastructures within the frameworks of FP7 and Horizon 2020 amounts to approximately 2,8 B€, of which the EU funding rises up to 2 B€.

[160] Infrastructures are the objective of 428 out of 30884 projects and respective calls.

[161] For the estimation of total and EU investment only FP7 and Horizon 2020 projects have been taken into consideration. The ESFRI landmarks are presented separately.

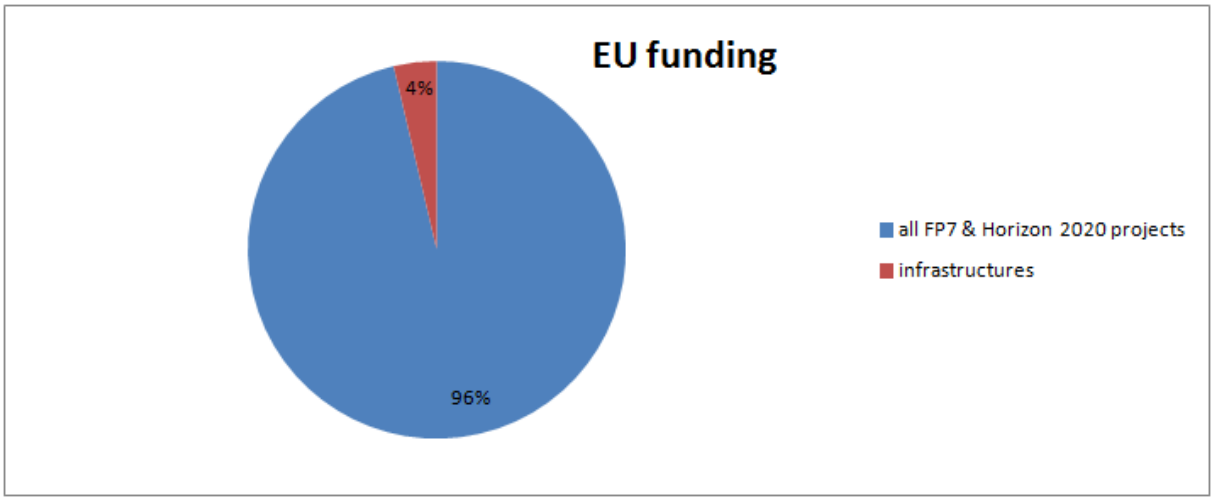


Figure 28. Proportion of EU funding of infrastructures

As already mentioned, infrastructures are a great source of data in massive quantities accompanied by tools and technologies developed for the analysis of this data. Most important, the majority of infrastructures offer open access to the data they collect. In the following sections we present for each economic sector the available infrastructures: existing, under construction, infrastructures chosen as landmarks by the European Strategy Forum on Research Infrastructure (ESFRI) and infrastructures which explicitly provide open access to their data.

3.5 Text and Data Mining in different scientific areas and fields of activity

In FutureTDM Deliverable D3.1 the total field of TDM activity has been charted in such a way that the different scientific research and commercial activities are grouped and mapped onto five economic sectors.[162]

In this section of the present report each sector will be depicted in detail, aiming to shed light on the main areas of interest for ***Text and Data Mining***: notably, the objectives of each sector, the available resources and content, the existing infrastructures and platforms and, finally, the relevant publications in the field.

The resources presented in each sector, especially repositories and large databases, are not exclusively European since they gather and present material (e.g. publications) on a global scale. This is the case of PubMed, for example, which, although US based, offers access to biomedical literature from around the world, Europe included.

3.6 Primary sector

3.6.1 Introduction to the sector

The primary sector includes all raw materials and natural resources as well as the methods used for their mining (in the prototypical sense of the word) and/or production. All human activities related to raw materials and natural resources, such as agriculture, farming, forestry, fishing, mining etc. also pertain to the primary sector.

[162] See ANNEX B: Economic sectors and Application Areas

The main objectives of ***TDM*** research in the primary sector include, among others

- the standardisation, understanding and forecasting of natural phenomena (e.g. earthquakes) and climate changes interacting with the lives of humans and other species,
- the prediction of better or worse/problematic crops or certain climate phenomena which affect them
- the correlation of pesticide use and animal/human diseases
- the better use of water resources/materials for production optimization.

3.6.2 Language/knowledge Resources, Tools and Technologies for TDM in the primary sector

The available language resources for the primary sector are mainly structured datasets in the form of ontologies, taxonomies, lexica and glossaries on agriculture and the environment. In addition, there are platforms, registries and repositories which provide data, tools and technologies. The following table contains some of the typical examples of resources for the primary sector presented according to their type.

Name	Type	Description[163]
Agrovoc[164]	controlled vocabulary	Covering all areas of interest of the Food and Agriculture Organization (FAO) of the United Nations, including food, nutrition, agriculture, fisheries, forestry, environment etc.
AGRIS[165]	database	Global public database providing access to bibliographic information on agricultural science and technology.
FAO Glossary of Biotechnology for Food and Agriculture[166]	glossary	Available through interface; portal created to store, manage and update concepts, terms and definitions Related to the various fields of FAO's activity.
Phytosanitary glossary[167]	glossary	Terms and definitions associated with the phytosanitary systems worldwide.
VOA3R[168]	platform	Re-using existing and mature metadata and semantics technology for retrieval of relevant open content and data.

[163] The information provided in the description of each resource has been verbatim copied from the relevant sites.
[164] http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus
[165] http://agris.fao.org/agris-search/index.do
[166] http://www.fao.org/biotech/biotech-glossary/en/
[167] http://www.fao.org/faoterm/collection/phytosanitary-glossary/en/
[168] http://voa3r.eu/

Name	Type	Description[163]
Chil knowledge platform[169]	portal	A knowledge web portal specialised in the agricultural, agro-industrial, environmental and rural sector.
Organic.Lingua[170]	portal	Enhance an existing Web portal with educational content on Organic Agriculture (OA) and Agroecology (AE), introducing automated multi-lingual services.
VEST Registry[171]	registry	Through the VEST Directory, AIMS becomes a reference portal to discover the vocabularies, technology and metadata sets used in the agricultural domain.
AgroPortal[172]	repository	AgroPortal is a repository which provides access and sharing for ontologies on the environment, agriculture and agronomy.
CAPSELLA[173]	repository	CAPSELLA (Collective Awareness Platforms for Environmentally-Sound Land Management Based on Data Technologies and Agrobiodiversity) aims to raise awareness about the proper use and management of agricultural land and agro-biodiversity for addressing major sustainability threats on several layers: ecological, societal, economic, food quality.
IFAD's Glossary - IFADTERM[174]	terminology	Glossary of terms produced by the International Fund for Agricultural Development (IFAD), available through the FAO Term Portal; available in Arabic, English, French and Spanish.

Figure 29. Resources used for TDM in the primary sector

3.6.3 Research Infrastructures relevant to the primary sector

The capital value of the **ESFRI infrastructures** concerning the environment rises to 1.6 B€. These infrastructures provide data collected from sensors on/below the ground, in the air/water or space which after being analysed give the capacity to understand the past and current as well as predict the future evolution of the atmospheric environment, the evolution of marine ecosystems and the impact of climate change on the lives of different species:

- **ACTRIS[175]:** Aerosols, Clouds and Trace gases Research Infrastructure
- **DANUBIUS-RI[176]:** International Centre for Advanced Studies on River-Sea Systems
- **EISCAT_3D[177]:** Next generation European incoherent scatter radar system

[169] http://www.chil.org
[170] https://dkm.fbk.eu/projects/organiclingua
[171] http://aims.fao.org/vest-registry
[172] http://agroportal.lirmm.fr/
[173] http://www.capsella.eu/
[174] http://www.fao.org/faoterm/news/detail/en/c/409955/
[175] http://www.actris.eu/
[176] http://www.danubius-ri.eu/about-us/

- **EPOS**[178]**:** European Plate Observing System
- **SIOS**[179]**:** Svalbard Integrated Arctic Earth Observing System
- **EMSO**[180]**:** European Multidisciplinary Seafloor and water-column Observatory
- **EURO-ARGO ERIC**[181]**:** European contribution to the international Argo Programme
- **IAGOS**[182]**:** In-service Aircraft for a Global Observing System
- **ICOS ERIC**[183]**:** Integrated Carbon Observation System
- **LifeWatch**[184]**:** e-infrastructure for Biodiversity and Ecosystem Research

Infrastructures concerning the environment within the framework of **FP7** and **Horizon 2020** have been granted a total of amount of 363 M€ with an EU contribution of 289,2 M€. The following have been selected among 54 projects, on the basis of the investment made which exceeds 10 M€ for each one, as indicative cases of research applications for the primary sector:

- **ENVRI PLUS**[185]**:** Environmental Research Infrastructures Providing Shared Solutions for Science and Society
- **NERA**[186]**:** Network of European Research Infrastructures for Earthquake Risk Assessment and Mitigation
- **AQUAEXCEL**[187]**:** AQUAculture infrastructures for EXCELLence in European Fish research
- **HYDRALAB IV**[188]**:** "HYDRALAB IV More than water; dealing with the complex interaction of water with environmental elements, sediment, structures and ice"

The openness of data indicated in previous sections has also been addressed by infrastructures. The following infrastructures explicitly state that they provide **open access** to their data[189]:

- **EPN2020-RI**[190]**:** EUROPLANET 2020 Research Infrastructure (*"The Europlanet 2020 Research Infrastructure (EPN2020-RI) will address key scientific and technological challenges facing modern planetary science by providing **open access** to state-of-the-art research data, models and facilities across the European Research Area"*),
- **FIXO3**[191]**:** Fixed Point Open Ocean Observatories Network (*"The FixO3 network will provide free and **open access** to in situ fixed point data of the highest quality"*),
- **EarthServer**[192]**:** European Scalable Earth Science Service Environment *("EarthServer aims at open access and ad-hoc analytics on Earth Science (ES) data, based on the OGC geo service standards Web Coverage Service (WCS) and Web Coverage Processing Service (WCPS)"*),

---

177 https://eiscat3d.se/
178 https://www.epos-ip.org/
179 http://www.sios-svalbard.org/
180 http://www.emso-eu.org/
181 http://www.euro-argo.eu/
182 http://www.iagos.org/
183 https://www.icos-ri.eu/
184 http://www.bbmri-eric.eu/
185 http://www.envriplus.eu/
186 http://www.nera-eu.org/
187 http://www.aquaexcel.eu/
188 http://hydralab.eu/
189 The information provided in the description of each resource has been verbatim copied from the relevant sites.
190 http://www.europlanet-2020-ri.eu/
191 http://www.fixo3.eu/
192 www.earthserver.eu

- **agINFRA**[193]**:** A data infrastructure to support agricultural scientific communities (*"agINFRA will try to remove existing obstacles concerning the **open access** to scientific information and data in agriculture, as well as improve the preparedness of agricultural scientific communities to face, manage and exploit the abundance of relevant data that is (or will be) available and can support agricultural research"*),
- **ODIP 2**[194]**:** Extending the Ocean Data Interoperability Platform (*"a number of prototype interoperability solutions will be developed which will be implemented by the regional data infrastructures to provide users with **open access** to good quality multidisciplinary data and associated services"*),
- **ICARE-2010**[195]**:** International Conference on Airborne Research for the Environment (*"There will be a special focus on **open access** to airborne research infrastructures, joint development of a heavy-payload and long endurance aircraft, availability of a stratospheric aircraft in Europe and the development of UAS for environmental research"*).

3.6.4 EU research projects relevant to the primary sector

A total amount of approximately **98,2 M€**, with a European Commission maximum contribution of **75,3 M€**, has been invested on **24** projects concerning ***TDM*** in the **primary** sector. Most projects focus on the collection and processing of data from **natural resources, oceanic data, climate data** or **agricultural data**. The following have been selected among the top twenty projects, based on their total investment, as representative cases for the primary sector:

- HEALS[196]: Health and Environment-wide Associations based on Large population Surveys
- MARS[197]: Managing Aquatic ecosystems and water Resources under multiple Stress
- SMARTWATER4EUROPE[198]: Demonstration of integrated smart water supply solutions at 4 sites across Europe
- ADAPTAWHEAT[199]: Genetics and physiology of wheat development to flowering: tools to breed for improved adaptation and yield potential
- HYDRONET[200]: Floating Sensorised Networked Robots for Water Monitoring
- BACI[201]: Detecting changes in essential ecosystem and biodiversity properties – towards a Biosphere Atmosphere Change Index
- SUSTAINMED[202]: Sustainable agri-food systems and rural development in the Mediterranean Partner Countries

3.6.5 Scientific publications relevant to the primary sector

The majority of the publications of TDM interest (retrieved from the CORE repository) which fall in the primary sector deal with the areas of Agriculture and Environment; other areas (e.g. Natural Resources) are comparatively less researched into while some others (e.g. Wholesale Trade) are not represented in the publications' set.

193 http://aginfra.eu/
194 http://www.odip.org/content/news_details.asp?menu=0100000_000014
195 http://www.eufar.net/events/154
196 http://www.heals-eu.eu/
197 http://www.mars-project.eu/
198 https://sw4eu.com/
199 https://www.jic.ac.uk/adaptawheat/index.htm
200 http://cordis.europa.eu/result/rcn/53857_en.html
201 http://baci-h2020.eu/index.php/
202 https://sustainmed.iamm.fr/

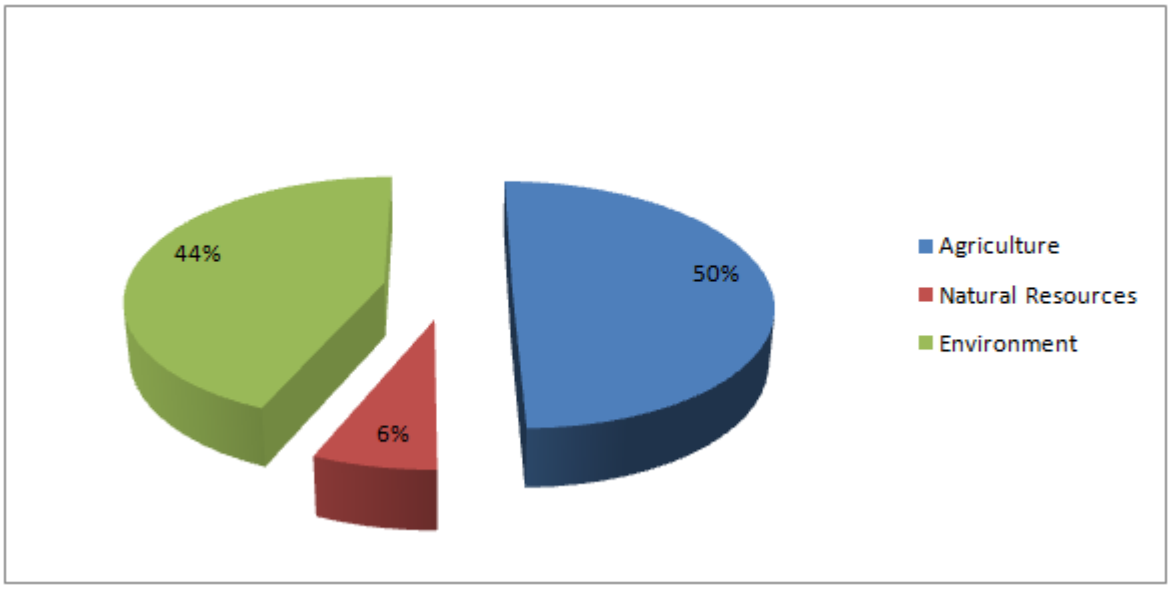


Figure 30. Publications in the primary sector

3.7 Secondary sector

3.7.1 Introduction to the sector

The secondary sector includes all the methods/techniques/tools/machinery used in order to transform the primary sector materials or resources into goods and products. All human activities related to manufacturing, processing and construction pertain to the secondary sector.

3.7.2 Language/knowledge Resources, Tools and Technologies for TDM in the secondary sector

Indicative examples of language resources for the secondary sector are **ontologies** for genes and key concepts in bioinformatics, **thesauri** for energy issues and physical sciences and **dictionaries** for chemical entities. Furthermore, there are **databases** for chemical structures and **platforms** for drugs, **repositories,** etc. relevant to the secondary sector as the following table indicates, presented by resource type:

Name	Type	Description[203]
ChemSpider[204]	database	ChemSpider is a free chemical structure database providing fast access to over 50 million structures, properties, and associated information. By integrating and linking compounds from ~500 data sources, ChemSpider enables researchers to discover the most comprehensive view of freely available chemical data from a single online search. It is owned by the Royal Society of Chemistry.
DrugBank[205]	database	The DrugBank database is a unique bioinformatics and cheminformatics

[203] The information provided in the description of each resource has been verbatim copied from the relevant sites.
[204] http://www.chemspider.com/
[205] http://www.drugbank.ca/

Name	Type	Description[203]
		resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information.
Global Reporting Initiative (GRI) Sustainability Disclosure Database[206]	database	The GRI Sustainability Disclosure Database is an extensive repository of sustainability reports that helps you search for and locate the information you need.
UniProt[207]	database	The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.
Chemical Entities of Biological Interest[208]	dictionary	Chemical Entities of Biological Interest (ChEBI) is a freely available dictionary of molecular entities focused on 'small' chemical entities.
EMBRACE Data and Methods ontology[209]	ontology	Ontology of well established, familiar concepts that are prevalent within bioinformatics, including types of data and data identifiers, data formats, operations and topics.
Gene Ontology[210]	ontology	Gene ontology (GO) is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species.
neXtProt[211]	platform	Exploring the universe of human proteins.
Open PHACTS Discovery Platform[212]	platform	The Open PHACTS Discovery Platform has been developed to reduce barriers to drug discovery in industry, academia and for small businesses. Data sources include ChEBI, ChEMBL, ChemSpider, ConceptWiki, DisGeNET, DrugBank, Gene Ontology, neXtProt, UniProt and WikiPathways. The platform is founded on semantic web and linked data principles and uses industrial strength tools such as Virtuoso to provide fast and robust access to the chemistry and biological data sources that you trust.
WikiPathways[213]	platform	WikiPathways is an open, public platform dedicated to the curation of biological pathways by and for the scientific community.

[206] http://database.globalreporting.org/
[207] http://www.uniprot.org/
[208] http://www.ebi.ac.uk/chebi/
[209] http://edamontology.org/page
[210] http://supfam.org/SUPERFAMILY/dcGO/
[211] http://www.nextprot.org/
[212] https://www.openphacts.org/
[213] http://www.wikipathways.org/index.php/WikiPathways

Name	Type	Description[203]
BiodiversityCatalogue[214]	registry	The BiodiversityCatalogue is a centralised registry of curated biodiversity Web services. It allows you to easily discover, register, annotate, monitor and use Web services.
BioPortal[215]	repository	BioPortal is the world's most comprehensive repository of biomedical ontologies.
ConceptWiki[216]	repository	ConceptWiki is a universal open access repository of editable concepts. ConceptWiki features, for each specific concept, an "Also Known As" table containing terms that can be used in different languages to identify the concept, as well as identifiers for the concept from various databases. Any additional information is available by clicking on a "More about this concept" link. The ConceptWiki focuses on the life sciences and on people working in the life sciences. The terminology and identifiers can be freely downloaded and can be used to as a thesaurus to identify references to the concepts in text and in databases.
INIS thesaurus[217]	thesaurus	The domain of knowledge covered by the INIS Thesaurus includes physics (in particular, plasma physics, atomic and molecular physics, and especially nuclear and high-energy physics), chemistry, materials science, earth sciences, radiation biology, etc.

Figure 31. Resources used for TDM in the secondary sector

3.7.3 Research Infrastructures relevant to the secondary sector

The capital value of the secondary sector ESFRI infrastructures rises to 13,2 B€. These infrastructures deal with energy issues, for example Solar Thermal and Solar Chemistry (CST) technologies, physical sciences (e.g. atomic, molecular, plasma and nuclear physics for a big variety of scientific applications, ranging from biology, chemistry and medicine to astrophysics in the laboratory) and engineering.

- **ECCSEL[218]:** European Carbon Dioxide Capture and Storage Laboratory Infrastructure
- **EU-SOLARIS[219]:** European SOLAR Research Infrastructure for Concentrated Solar Power
- **MYRRHA[220]:** Multi-purpose hYbrid Reactor for High-tech Applications
- **WindScanner[221]:** European WindScanner Facility
- **CTA[222]:** Cherenkov Telescope Array
- **EST[223]:** European Solar Telescope
- **KM3NeT 2.0[224]:** M3 Neutrino Telescope 2.0: Astroparticle & Oscillations Research with Cosmics in the Abyss

[214] https://www.biodiversitycatalogue.org
[215] http://bioportal.bioontology.org/
[216] http://www.conceptwiki.org/
[217] https://www.iaea.org/inis/products-services/INIS-Thesaurus/index.html
[218] http://www.eccsel.org/
[219] http://eusolaris.eu/
[220] http://myrrha.sckcen.be/
[221] http://www.windscanner.eu/
[222] https://portal.cta-observatory.org/
[223] http://www.est-east.eu/

- **JHR**[225]**:** Jules Horowitz Reactor
- **E-ELT**[226]**:** European Extremely Large Telescope
- **ELI**[227]**:** Extreme Light Infrastructure
- **EMFL**[228]**:** European Magnetic Field Laboratory
- **ESRF UPGRADES**[229]**:** Phase I Phase II: Extremely Brilliant Source
- **European Spallation Source ERIC**[230]**:** European Spallation Source
- **European XFEL**[231]**:** European X-Ray Free-Electron Laser Facility
- **FAIR**[232]**:** Facility for Antiproton and Ion Research
- **HL-LHC**[233]**:** High-Luminosity Large Hadron Collider
- **ILL 20/20**[234]**:** Institut Max von Laue-Paul Langevin
- **SKA**[235]**:** Square Kilometre Array
- **SPIRAL2**[236]**:** Système de Production d'Ions Radioactifs en Ligne de 2e génération

Moreover, 115 projects resulting in infrastructures have been funded within the FP7 and Horizon 2020 frameworks with a total of 847,8 M€ of which EU contributed 564,4 M€. A selection of the highest investment projects (based on their total cost) is presented below:

- **EUCARD**[237] **:** European Coordination for Accelerator Research and Development
- **AIDA**[238]**:** Advanced European Infrastructures for Detectors at Accelerators
- **EMI**[239]: European Middleware Initiative
- **NFFA-Europe**[240]: Nanoscience Foundries and Fine Analysis – Europe
- **SOPHIA**[241]: PhotoVoltaic European Research Infrastructure

The following infrastructures support and promote open access to their data:

- **OPTICON**[242]**:** Optical Infrared Co-ordination Network for Astronomy (*"The most important need for most astronomers is to have **open access** to a viable set of medium aperture telescopes, with excellent facilities, complemented by superb instrumentation on the extant large telescopes, while working towards next generation instrumentation on the future flagship, the European Extremely Large Telescope. OPTICON has made a substantial contribution to preparing the realisation of that ambition"*),

---

224 http://www.km3net.org/
225 http://www.cad.cea.fr/rjh/
226 http://www.eso.org/public/teles-instr/e-elt/
227 http://www.eli-laser.eu/
228 http://www.emfl.eu/
229 http://www.esrf.eu
230 http://www.europeanspallationsource.se
231 http://www.xfel.eu
232 http://www.fair-center.de
233 http://home.cern/
234 http://www.ill.eu
235 http://www.skatelescope.org
236 http://www.ganil-spiral2.eu
237 http://eucard2.web.cern.ch/
238 http://aida2020.web.cern.ch/
239 http://www.eu-emi.eu/
240 http://www.nffa.eu/
241 http://www.sophia-ri.eu/
242 http://www.astro-opticon.org/

- **EUFAR2**[243]**:** European Facility for Airborne Research in Environmental and Geo-sciences (*"EUFAR aims at providing researchers with **Open Access** to the airborne facilities the most suited to their needs"*),
- **SUP@VAMDC**[244]: Support at the Virtual Atomic and Molecular Data Centre (*"This programme includes the development of education-related tools linking VAMDC's scientific repositories and research data infrastructures, including establishing a free **open access** repository containing all peer-reviewed articles resulting from the VAMDC programme"*),
- **MNTEE**[245]: MNT Europe Extension (*"In order to be able to handle future **open access**, a central contact point should be created, which allows all partners in the project to access and distribute all incoming contacts"*).

3.7.4 EU research projects relevant to the secondary sector

A total amount of 806,4 M€, with a European Commission maximum contribution of 633,8 M€, has been invested on 253 projects concerning ***TDM*** in the secondary sector. The majority of projects focus on the collection and processing of biological and medical data for biotechnology applications and the subsequent provision of better health care services. Many projects concern the processing of energy data and the development of green or smart energy systems. The following have been selected among the top twenty projects, based on their total cost, as representative cases for the secondary sector:

- SYSTEMS MICROSCOPY[246]: Systems microscopy – a key enabling methodology for next-generation systems biology
- GEN2PHEN[247]: Genotype-To-Phenotype Databases: A Holistic Solution
- APO-SYS[248]: "Apoptosis systems biology applied to cancer and AIDS. An integrated approach of experimental biology, data mining, mathematical modelling, biostatistics, systems engineering and molecular medicine"
- EURECA[249]: Enabling information re-Use by linking clinical REsearch and CAre
- LinkedDesign[250]: Linked Knowledge in Manufacturing, Engineering and Design for Next-Generation Production
- Fortissimo 2[251]: Factories of the Future Resources, Technology, Infrastructure and Services for Simulation and Modelling 2.

3.7.5 Scientific publications relevant to the secondary sector

Half the publications of TDM interest (retrieved from the CORE repository) which fall in the secondary sector concern Engineering. The other areas of interest cover Aerospace and Energy followed by Construction and Microelectronics as depicted in Figure 32.

[243] http://www.eufar.net/
[244] http://www.sup-vamdc.vamdc.org/
[245] http://cordis.europa.eu/project/rcn/88391_en.html
[246] http://www.systemsmicroscopy.eu/
[247] http://gen2phen.org/
[248] http://cordis.europa.eu/result/rcn/56809_en.html
[249] http://eurecaproject.eu/
[250] http://www.linkeddesign.eu/
[251] http://i4ms.eu/projects/projects_detail.php?post_id=13

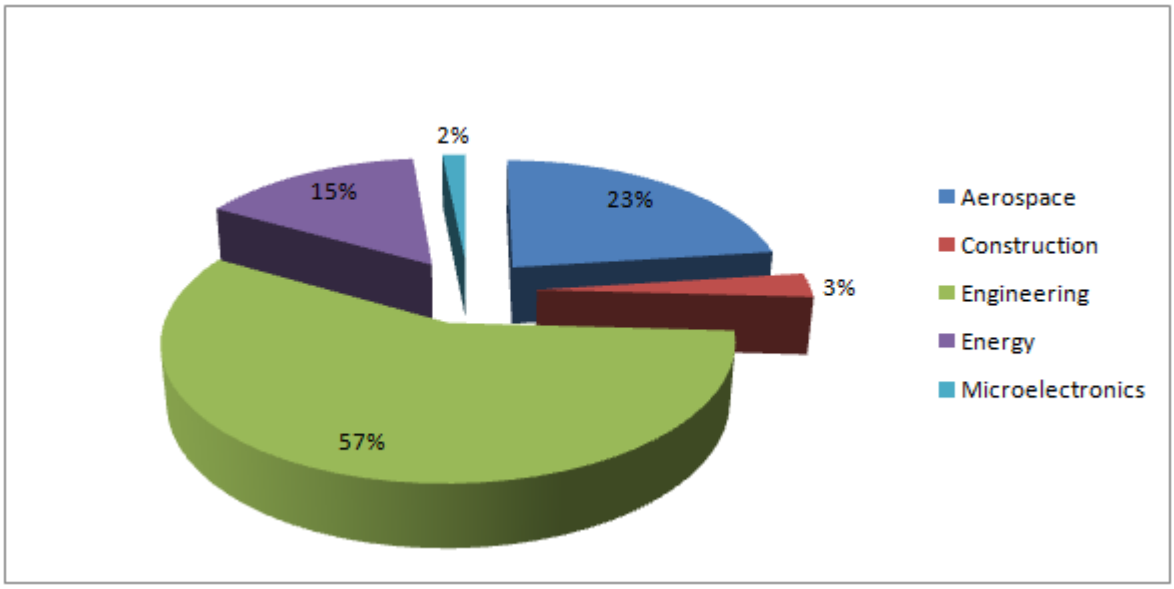


Figure 32. Publications in the secondary sector

3.8 Tertiary sector

3.8.1 Introduction to the sector

The tertiary sector includes all the services provided to individuals or organized groups such as businesses. All human activities related to retail and wholesale sales, transportation, entertainment, tourism, insurance, and many more pertain to the tertiary sector.

3.8.2 Language/knowledge Resources, Tools and Technologies for TDM in the tertiary sector

Many of the resources for the tertiary sector are built on data gathered from medical information, patients' medical records, related scientific publications and rather scarce resources for other businesses. The collections and services created aim at helping to get a better understanding of the needs and the outcomes of the health care and customer services provided and subsequently improve them, while optimizing business processes to achieve cost reductions. Figure 33 provides an insight to typical resources for the tertiary sector.

Name	Type	Description
Medical Subject Headings[252]	controlled vocabulary	Controlled vocabulary for classification relevant to the medical domain.
Clinical Record Interactive Search[253]	database	The Clinical Record Interactive Search (CRIS) system has been developed for use within the Maudsley Biomedical Research Centre and Dementia Unit (BRC/U). It provides authorised researchers with regulated, secure access to anonymised information extracted from South London and Maudsley NHS Foundation Trust (SLaM) electronic clinical records system.
Diabetes Open Directory[254]	database	The Diabetes Open Directory (DOD) is a project intended to serve the needs of scientists in the field of biomedical diabetes research by

[252] https://www.nlm.nih.gov/mesh/introduction.html
[253] http://www.maudsleybrc.nihr.ac.uk/about-us/core-facilities/clinical-record-interactive-search-cris/

Name	Type	Description
		providing many recent full-text-format articles from a range of topic-related journals. The DOD is a web, which intends to grow by collecting and publishing information regarding biomedical diabetes research. The system is in its test stage currently, but any information inserted by the users will be preserved and go into the final version.
DisGeNET[255]	database	DisGeNET is a discovery platform integrating information on gene-disease associations (GDAs) from several public data sources and the literature (Piñero et al., 2015). The current version contains (DisGeNET v4.0) contains 429,036 associations, between 17,381 genes and 15,093 diseases, disorders and clinical or abnormal human phenotypes.
South East London Community Health Study[256]	database	The study encompasses a wide range of health service users as possible. The information collected gives us a better understanding of the health needs of the community and enables service providers to plan and improve services more effectively.
Freme[257]	framework	FREME addresses the general systemic and technological challenges to validate that multilingual and semantic technologies are ready for their integration in real life business cases in innovative way. These technologies are capable to process (harvest and analyse) content, capture datasets, and add value throughout content and data value chains across sectors, countries, and languages.
Europe PMC[258]	repository	Europe PMC includes resources from PubMed and PubMed Central (PMC), projects developed at the NCBI in the USA. Europe PMC is part of a network of PMC International (PMCI) repositories that also includes PMC Canada.
PubMed[259]	repository	PubMed comprises over 25 million citations for biomedical literature from MEDLINE, life science journals, and online books. PubMed citations and abstracts include the fields of biomedicine and health, covering portions of the life sciences, behavioral sciences, chemical sciences, and bioengineering. PubMed also provides access to additional relevant web sites and links to the other NCBI molecular biology resources.
AnnoMarket[260]	web service registry	AnnoMarket aims to revolutionise the text annotation market, by delivering annomarket.com – an affordable, open marketplace for pay-as-you-go, cloud-based extraction resources and services, in multiple languages. The project is driven by a commercially-dominated consortium, from 3 EU countries and with 41% of the budget assigned to SMEs. The techniques will be generic with many business applications, e.g. large-volume multi-lingual information management, business intelligence, social media monitoring, customer relations management.

Figure 33. Resources used for TDM in the tertiary sector

3.8.3 Research Infrastructures relevant to the tertiary sector

The capital value of the ESFRI infrastructures for the tertiary sector is approximately[261] 1,5 B€. These infrastructures fall under the broad category of Health & Food and are presented here

[254] http://www.diabetes-od.org/directory/content/index_eng.html
[255] http://www.disgenet.org/web/DisGeNET/menu
[256] http://www.kcl.ac.uk/innovation/groups/selcoh/index.aspx
[257] http://www.freme-project.eu/
[258] https://europepmc.org/About;jsessionid=bdupZmoHkSZXBKMRPfD5.0
[259] http://www.ncbi.nlm.nih.gov/pubmed
[260] https://annomarket.eu/

mainly because they study the interrelation of parameters, such as agriculture or climate change coming from the primary and secondary sector, on human health and the health care services.

- **AnaEE**[262]**:** Infrastructure for Analysis and Experimentation on Ecosystems
- **EMBRC**[263]**:** European Marine Biological Resource Centre
- **EMPHASIS**[264]**:** European Infrastructure for multi-scale Plant Phenomics and Simulation for food security in a changing climate
- **ERINHA**[265]**:** European research infrastructure on highly pathogenic agents
- **EU-OPENSCREEN**[266]**:** European Infrastructure of Open Screening Platforms for Chemical Biology
- **Euro-BioImaging**[267]**:** European Research Infrastructure for Imaging Technologies in Biological and Biomedical Sciences
- **ISBE**[268]**:** Infrastructure for Systems Biology Europe
- **MIRRI**[269]**:** Microbial Resource Research Infrastructure
- **BBMRI ERIC**[270]**:** Biobanking and BioMolecular resources Research Infrastructure
- **EATRIS ERIC**[271]**:** European Advanced Translational Research Infrastructure in Medicine
- **ECRIN ERIC**[272]**:** European Clinical Research Infrastructure Network
- **ELIXIR**[273]**:** A distributed infrastructure for life-science information
- **INFRAFRONTIER**[274]**:** European Research Infrastructure for the generation, phenotyping, archiving and distribution of mouse disease models
- **INSTRUCT**[275]**:** Integrated Structural Biology Infrastructure

From the FP7 and Horizon 2020 frameworks 98 projects have resulted in infrastructures for the tertiary sector. The investment made from the EU is 492,5 M€ with a total investment of 599,5 M€. Some of the most expensive projects (with a total cost for each one of more than 10M€) are the following:

- **CORBEL**[276]: Coordinated Research Infrastructures Building Enduring Life-science services
- **BIOMEDBRIDGES**[277]: Building data bridges between biological and medical infrastructures in Europe
- **EVAg**[278]: European Virus Archive goes global
- **TRANSVAC**[279]: European Network of Vaccine Development and Research

---

[261] The estimation of the capital value of all infrastructures was not possible due to the fact that there is no information for ⅓ of them.
[262] http://www.anaee.com/
[263] http://www.embrc.eu/
[264] http://www.plant-phenotyping.org/
[265] http://www.erinha.eu/
[266] http://www.eu-openscreen.eu/
[267] http://www.eurobioimaging.eu/
[268] http://project.isbe.eu/
[269] http://www.mirri.org/home.html
[270] http://www.bbmri-eric.eu/
[271] http://www.eatris.eu/
[272] http://www.ecrin.org/
[273] http://www.elixir-europe.org
[274] http://www.infrafrontier.eu
[275] http://www.structuralbiology.eu
[276] http://www.corbel-project.eu/home.html
[277] http://www.biomedbridges.eu/
[278] http://www.european-virus-archive.com/

- **EURIPRED**[280]: European Research Infrastructures for Poverty Related Diseases

The following infrastructures support and promote open access to their data:

- **Global BioImaging**[281]: Global BioImaging Project - International imaging infrastructure services for the life science community ("*Euro-BioImaging together with the AMMRF, NIF and India-BioImaging will exchange best practice in imaging facility management and operation, quality management, **open access** policies;"*)
- **pro-iBiosphere**[282]: Coordination and policy development in preparation for a European Open Biodiversity Knowledge Management System, addressing Acquisition, Curation, Synthesis, Interoperability and Dissemination ("*A system that facilitates **open access** to taxonomic data is essential because it will allow a sustainable provision of high quality data to partners and users, including e-science infrastructure projects as well as global initiatives on biodiversity informatics"*)

3.8.4 EU research projects relevant to the tertiary sector

A total amount of **1,1** B€, with a European Commission maximum contribution of **852,4 M€**, has been invested on **313** projects concerning ***TDM*** in the tertiary sector. Most projects focus on the collection and processing of data from medical records and biology databases**.** The next category in frequency is that of IT services followed by the business application area. The following projects have been selected among the top twenty, based on their total cost which for each one ranges from 10 M€ to 23 M€, as representative cases for the tertiary sector:

- SENSEI[283]: Integrating the Physical with the Digital World of the Network of the Future
- SYSTEMS MICROSCOPY[284]: Systems microscopy – a key enabling methodology for next-generation systems biology
- SIIP[285]: Speaker Identification Integrated Project
- APO-SYS[286]: "Apoptosis systems biology applied to cancer and AIDS. An integrated approach of experimental biology, data mining, mathematical modelling, biostatistics, systems engineering and molecular medicine"
- KAP[287]: Knowledge, Awareness and Prediction of Man, Machine, Material and Method in Manufacturing
- mPlane[288]: mPlane – an Intelligent Measurement Plane for Future Network and Application Management
- BRAVEHEALTH[289]: Patient Centric Approach for an Integrated, Adaptive, Context Aware Remote Diagnosis and Management of Cardiovascular Diseases

[279] http://www.transvac.org/
[280] http://www.euripred.eu/
[281] http://www.eurobioimaging.eu/content-page/global-bioimaging-project
[282] http://www.pro-ibiosphere.eu/
[283] https://www.utwente.nl/ctit/research/research_projects/concluded/international/fp7/fp7-ip/sensei/
[284] http://www.systemsmicroscopy.eu/
[285] http://www.siip.eu/SIIP_Project
[286] http://www.systemsmedicineireland.ie/research/colorectal-cancer/apo-sys/
[287] http://cordis.europa.eu/project/rcn/95347_en.html
[288] http://www.ict-mplane.eu/
[289] https://www.utwente.nl/ewi/telemedicine/Projects/bravehealth/

3.8.5 Scientific publications relevant to the tertiary sector

The publications of TDM interest (retrieved from the CORE repository) for the tertiary sector cover a variety of fields as depicted in Figure 34. The majority of publications (60%) pertain to the fields of Medicine, Health Care and Social Care, while 30% are related to generic IT services. The remaining publications are related to Finance, Public Administration, Entertainment and Transportation.

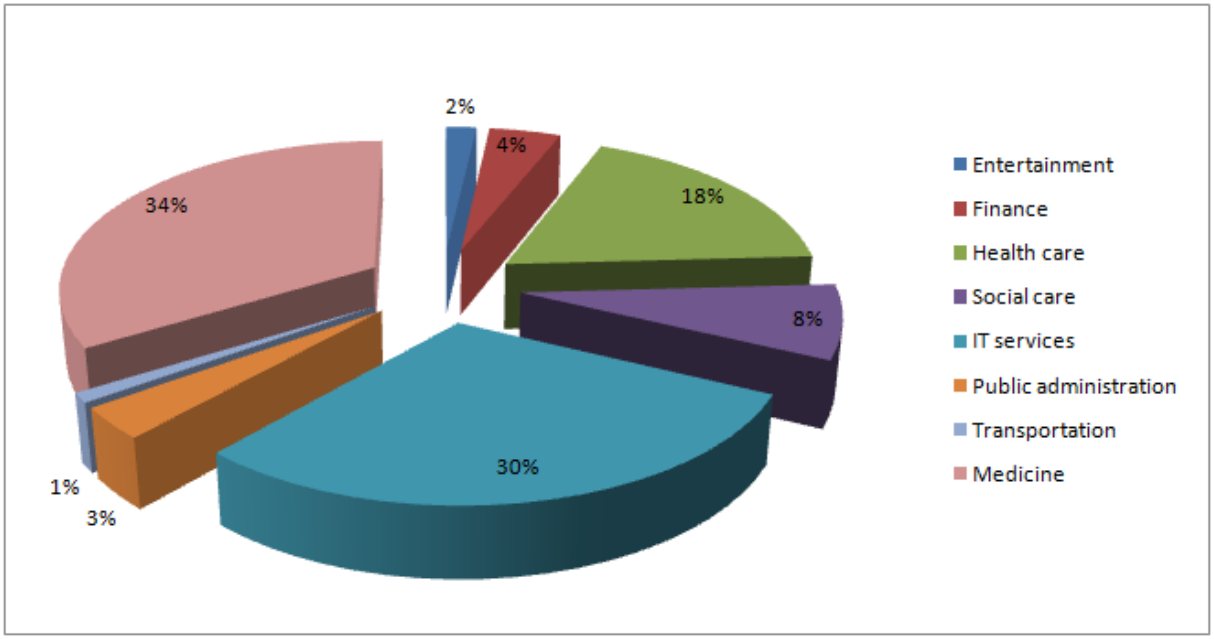


Figure 34. Publications in the tertiary sector

3.9 Quaternary sector

3.9.1 Introduction to the sector

The quaternary sector includes all services dealing with knowledge and information such as research, development and education among others. Research and development are terms applying to all scientific fields; therefore the following sections include projects, infrastructures, language resources and publications which could also be identified as belonging to any of the first three economic sectors but are presented here because the focus lies on the innovation, introduction, and improvement of products and processes of the various disciplines and application areas.

3.9.2 Language/knowledge Resources, Tools and Technologies for TDM in the quaternary sector

The resources provided for the quaternary sector are mostly in the form of **structured knowledge representations**, e.g. ontologies, classification schemes, thesauri, knowledge bases, etc. Some additional existing infrastructures, platforms and repositories serving the needs of research and education are also presented in the following table[290]. In some cases they offer

[290] For this table, we have relied on, reused and expanded data from OpenMinTed Deliverable D5.1.

language resources as well as web services especially with regard to Humanities and Social Sciences.

Name	Type	Description
IULA-UPF CLARIN Competence Center[291]	catalogue	Competence Centre IULA-UPF-CC CLARIN manages, disseminates and facilitates this catalogue, which provides access to reference information on the use of language technology projects and studies in different disciplines, especially with regard to Humanities and Social Sciences.
LDC catalogue[292]	catalogue	The Linguistic Data Consortium (LDC) is an open consortium of universities, libraries, corporations and government research laboratories. It was formed in 1992 to address the critical data shortage then facing language technology research and development.
CLARIN national repositories[293]	centre registry	Infrastructures dedicated to various languages, developed from the respective countries members of the European CLARIN infrastructure, a pan-European network of organisations for the collection, documentation, curation and distribution of Language Resources, Technologies and Language processing web services.
Library of Congress Subject Headings[294]	classification system	The classification system used for classifying material in the Library of Congress, which has spread to other libraries and organizations as well.
Universal Decimal Classification[295]	classification system	Classification system widely used in libraries, bibliographic, documentation and information services.
Freebase[296]	collaborative knowledge base	Freebase is an open, Creative Commons licensed graph database with more than 23 million entities.
COAR Resource Type Vocabulary[297]	controlled vocabulary	Controlled Vocabulary for types of digital resources, such as publications, research data, audio and video objects, etc.
LRE Map[298]	database	The LRE Map is a new mechanism intended to monitor the use and creation of language resources by collecting information on both existing and newly-created resources during the submission process. It is a collective enterprise of the LREC community, as a first step towards the creation of a very broad, community-built, Open Resource Infrastructure.
CASRAI dictionary[299]	dictionary	Dictionary on research output types
Big Data Europe[300] (Empowering Communities with Data Technologies)	infrastructure	Big Data Europe aims to enable European companies to build innovative multilingual products and services based on semantically interoperable, large-scale, multi-lingual data assets and knowledge, available under a variety of licenses and business models.

[291] http://lod.iula.upf.edu/index-en.html
[292] https://www.ldc.upenn.edu/
[293] https://centres.clarin.eu/
[294] http://id.loc.gov/authorities/subjects.html
[295] http://www.udcc.org/index.php/site/page?view=about
[296] http://wiki.freebase.com/
[297] https://www.coar-repositories.org/activities/repository-interoperability/ig-controlled-vocabularies-for-repository-assets/deliverables/
[298] http://www.resourcebook.eu/
[299] http://dictionary.casrai.org/Output_Types
[300] http://www.big-data-europe.eu/

Name	Type	Description
ORTOLANG[301]	infrastructure	ORTOLANG is an EQUIPEX project accepted in February 2012 in the framework of investissements d'avenir. Its aim is to construct a network infrastructure including a repository of language data (corpora, lexicons, dictionaries etc.) and readily available, well-documented tools for its processing.
Wikipedia[302]	Internet encyclopaedia	Wikipedia is an Internet encyclopedia, supported and hosted by the non-profit Wikimedia Foundation. It is a free-of-cost encyclopedia with its articles being free-content; those who use Wikipedia can edit almost any article accessible.
Wikidata[303]	knowledge base	Wikidata is a free and open knowledge base that can be read and edited by both humans and machines.
DBpedia[304]	knowledge base and ontology	DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web.
YaGO[305]	large semantic knowledge base	YAGO is a huge semantic knowledge base, derived from Wikipedia WordNet and GeoNames. Currently, YAGO has knowledge of more than 10 million entities (like persons, organizations, cities, etc.) and contains more than 120 million facts about these entities.
UBY[306]	large-scale lexical semantic resource combining several existing resources	UBY is a large-scale lexical-semantic resource for natural language processing (NLP) based on the ISO standard Lexical Markup Framework (LMF). UBY combines a wide range of information from expert-constructed and collaboratively constructed resources for English and German
WordNet[307]	lexical database	WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.
Dewey Decimal Classification[308]	library classification scheme	Library classification scheme, also widely used for the classification of documents.
META-SHARE[309]	network of repositories	META-SHARE is an open and secure network of repositories for sharing and exchanging language data, tools and related web services.
OLiA[310]	ontologies for linguistic annotation	The Ontologies of Linguistic Annotation (OLiA) are a repository of linguistic data categories used for corpus annotation, Natural Language Processing (NLP) tools, machine-readable dictionaries and other linguistic resources.
PROV Ontology[311]	ontology	Ontology for declaring provenance information; implemented in OWL 2.0

[301] https://www.ortolang.fr
[302] https://en.wikipedia.org
[303] https://www.wikidata.org
[304] http://wiki.dbpedia.org/
[305] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/
[306] https://www.ukp.tu-darmstadt.de/data/lexical-resources/uby/
[307] https://wordnet.princeton.edu/
[308] https://www.oclc.org/dewey.en.html
[309] http://www.meta-share.org/
[310] http://nachhalt.sfb632.uni-potsdam.de/owl/#olia
[311] http://www.w3.org/TR/prov-o/

Name	Type	Description
Semantic Web for Research Communities[312]	ontology	Ontology for modelling entities of research communities such as persons, organizations, publications (bibliographic metadata) and their relationships.
GOLD[313]	ontology for linguistic description	The purpose of the GOLD Community is to bring together scholars interested in best-practice encoding of linguistic data. This standard encompasses linguistic concepts, definitions of these concepts and relationships between them in a freely available ontology.
OpenCyc[314]	open source subset of the CYC knowledge base	The OpenCyc Platform is a gateway to the full power of Cyc, the world's largest and most complete general knowledge base and commonsense reasoning engine.
ALVEO[315]	platform/infrastructure	Alveo provides on-line infrastructure for accessing human communication data sets (speech, texts, music, video, etc.) and for using specialised tools for searching, analysing and annotating that data.
Language Grid[316]	platform/infrastructure	Language Grid is an online multilingual service platform which enables easy registration and sharing of language services such as online dictionaries, bilingual corpora, and machine translators.
LAPPS Grid[317]	platform/infrastructure	The Language Application (LAPPS) Grid project is a collaborative effort among US partners Vassar College, Brandeis University, Carnegie-Mellon University, and the Linguistic Data Consortium at the University of Pennsylvania, and is funded by the US National Science Foundation. The project is establishing a framework that enables language service discovery, composition, and reuse and supports state-of-the-art evaluation of natural language Processing (NLP) components.
QT21[318]	platform/infrastructure	QT21 is repository devoted to the sustainable sharing, processing and dissemination of language resources.
CLARIN - Virtual Language Observatory (VLO)[319]	registry	The Virtual Language Observatory (VLO) can be characterized as a search engine or a metadata-based portal for language resources.
Connecting Repositories[320]	repository	The mission of CORE (COnnecting REpositories) is to aggregate all open access research outputs from repositories and journals worldwide and make them available to the public.
datahub[321]	repository	The Datahub is a free, powerful data management platform from the Open Knowledge Foundation, based on the CKAN data management system. CKAN is a tool for managing and publishing collections of data.
ELRA Catalogue of Language Resources[322]	repository	The ELRA Catalogue of Language Resources offers a repository of Language Resources (LRs)

[312] http://ontoware.org/swrc/
[313] http://www.linguistics-ontology.org/gold.html
[314] http://www.opencyc.org/
[315] http://alveo.edu.au/
[316] http://langrid.org/en/index.html
[317] http://www.lappsgrid.org/
[318] http://qt21.metashare.ilsp.gr/
[319] https://vlo.clarin.eu
[320] http://core.ac.uk/
[321] https://datahub.io/

Name	Type	Description
SPARC[323]	repository	SPARC (the Scholarly Publishing and Academic Resources Coalition) works to enable the open sharing of research outputs and educational materials in order to democratize access to knowledge, accelerate discovery, and increase the return on our investment in research and education.
ZENODO[324]	repository	Zenodo builds and operates a simple and innovative service that enables researchers, scientists, EU projects and institutions to share, preserve and showcase multidisciplinary research results (data and publications) that are not part of the existing institutional or subject-based repositories of the research communities.
Bielefeld Academic Search Engine[325]	repository/search engine	BASE is one of the world's most voluminous search engines especially for academic open access web resources. BASE is operated by Bielefeld University Library.
EuroVoc[326]	thesaurus	Multilingual, multidisciplinary thesaurus covering the activities of the EU (in particular the European Parliament); it is extensively used for the classification of EU publications and, in general, public sector administrative documents.
Thesaurus for the Social Sciences (TheSoz)[327]	thesaurus	The Thesaurus for the Social Sciences (Thesaurus Sozialwissenschaften) contains about 12,000 entries, of which more than 8,000 are descriptors (authorised keywords) and about 4,000 non-descriptors. Topics in all of the social science disciplines are included.

Figure 35. Resources used for TDM in the quaternary sector

3.9.3 Research Infrastructures relevant to the quaternary sector

The capital value of the ESFRI infrastructures for the quaternary sector is approximately[328] 614,3 M€. The quaternary sector infrastructures concern social and cultural innovation, research and development. Their main interest and purpose is to promote research in various scientific fields which will in addition improve the quality of life for citizens and societies in general.

- **E-RIHS[329]:** European Research Infrastructure for Heritage Science
- **CESSDA[330]:** Consortium of European Social Science Data Archives
- **CLARIN ERIC[331]:** Common Language Resources and Technology Infrastructure
- **DARIAH ERIC[332]:** Digital Research Infrastructure for the Arts and Humanities
- **ESS ERIC[333]:** European Social Survey
- **SHARE ERIC[334]:** Survey of Health, Ageing and Retirement in Europe

[322] http://catalog.elra.info/index.php
[323] http://sparcopen.org/
[324] http://www.zenodo.org/
[325] http://base-search.net
[326] http://eurovoc.europa.eu/
[327] http://www.gesis.org/en/services/research/tools-zur-recherche/social-science-thesaurus/
[328] The estimation of the capital value of all infrastructures was not possible due to the fact that there is no information for half of them.
[329] www.e-rihs.eu
[330] http://www.cessda.net
[331] http://www.clarin.eu
[332] http://www.dariah.eu
[333] http://www.europeansocialsurvey.org
[334] http://www.share-project.org/

- **PRACE**[335]**:** Partnership for Advanced Computing in Europe

The FP7 and Horizon 2020 projects which have resulted in the creation/development of **infrastructures** have been granted a total amount of 1 B€ with a contribution of EU rising up to 693 M€. From the 161 projects pertaining to the quaternary sector we present some of the highest funded with a total investment for each one exceeding 10 M€ and rising up to 177 M€:

- **GN3**[336]: Multi-Gigabit European Research and Education Network and Associated Services (GN3)
- **PARTHENOS**[337]: Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies
- **EUROFLEETS2**[338]: New operational steps towards an alliance of European research fleets
- **CHARISMA**[339]: Cultural heritage Advanced Research Infrastructures: Synergy for a Multidisciplinary Approach to Conservation/Restoration
- **TIARA**[340]: Test Infrastructure and Accelerator Research Area
- **EUDAT2020**[341]**:** EUropean DATa 2020

Of particular interest are the following infrastructures, either for **providing open access** to their data or for being ***TDM* oriented**:

- **OpenAIRE2020**[342]**:** Open Access Infrastructure for Research in Europe 2020 ("*OpenAIRE2020 represents a pivotal phase in the long-term effort to implement and strengthen the impact of the **Open Access** (OA) policies of the European Commission (EC), building on the achievements of the OpenAIRE projects*[343]"),
- **EHRI**[344]: European Holocaust Research Infrastructure ("*Although EHRI is primarily geared towards the needs of scholarly communities, the online availability and **open access** to reliable and properly contextualized Holocaust material is relevant and important for the larger public well beyond scholarly communities, as its research topic is deeply rooted in the development of European societies*"),
- **EGI-Engage**[345]: Engaging the EGI Community towards an Open Science Commons ("*The mission of EGI-Engage is to accelerate the implementation of the Open Science Commons vision, where researchers from all disciplines have easy and **open access** to the innovative digital services, data, knowledge and expertise they need for their work*"),
- **CALIPSO**[346]: Coordinated Access to Lightsources to Promote Standards and Optimization (*The consortium is characterised by common objectives, harmonised decisions, transnational **open access** based on excellence and joint development of new instruments*),

[335] http://www.prace-ri.eu/
[336] http://www.geant.org/Projects/GEANT_Project_GN4/Pages/Home.aspx
[337] http://www.parthenos-project.eu/
[338] http://www.eurofleets.eu/np4/302.html
[339] http://www.charismaproject.eu/
[340] http://www.eu-tiara.eu/
[341] https://www.eudat.eu/
[342] https://www.openaire.eu/
[343] The OpenAire projects include the **OpenAIRE** (Open Access Infrastructure for Research in Europe), **OpenAIREplus** (2nd-Generation Open Access Infrastructure for Research in Europe) and the **OpenAIRE2020** (Open Access Infrastructure for Research in Europe 2020).
[344] http://www.ehri-project.eu/
[345] https://www.egi.eu/about/egi-engage/
[346] http://www.calipso.wayforlight.eu/

- **OpenMinTeD**[347]: Open Mining INfrastructure for TExt and Data ("***Text and Data Mining*** *is emerging as a powerful tool for harnessing the power of structured and unstructured content and data, by analysing them at multiple levels and in several dimensions to discover hidden and new knowledge"*),
- **Sci-GaIA**[348]: Energising Scientific Endeavour through Science Gateways and e-Infrastructures in Africa ("*Sci-GaIA plans to work with new and emerging CoPs to develop these exciting technologies, to strengthen e-Infrastructure service provision, especially in terms of* ***open access*** *linked data, and to deliver training and dissemination workshops. This will give a sustainable foundation on which African e-Infrastructures can be developed and be linked to scientific networks across Africa"*).
- **RDA Europe**[349]: Research Data Alliance - Europe 3 ("*The Research Data Alliance (RDA) is rapidly building the social and technical bridges that enable* ***open sharing*** *and re-use of data on a global level"*).

3.9.4 EU research projects relevant to the quaternary sector

A total amount of **402 M€,** with a European Commission maximum contribution of **319,9 M€,** has been invested on **273** projects concerning TDM in the **quaternary** sector. This sector includes all these projects that are classified by the EC as scientific research under RTD horizontal topics without further distinguishing, in most cases, the particular scientific area, the development of which the particular action will affect. As a result strong infrastructural projects like EUDAT2020 (see section 3.9.3) and projects requiring substantial financial investment in the area of Medicine and Health Sciences, e.g. the METACARDIS project, are grouped together. The following have been selected among the top twenty projects, based on their total cost, as representative cases for the tertiary sector:

- METACARDIS[350]: Metagenomics in Cardiometabolic Diseases
- RD-CONNECT[351]: "RD-CONNECT: An integrated platform connecting registries, biobanks and clinical bioinformatics for rare disease research"
- IRIMA[352]: Industrial Research and Innovation Monitoring and Analysis
- SCY[353]: Science Created by You (SCY)
- NEXT-TELL[354]: Next Generation Teaching, Education and Learning for Life

3.9.5 Scientific publications relevant to the quaternary sector

The publications concerning ***TDM*** for the quaternary sector fall into two big categories: the first being **research**, which is a generic term including many scientific fields, and the second one being **education**. The research publications presented here are those for which no specific scientific field was denoted from the set of words of topic *t* (see section 3.4.3). Otherwise, the classification of the topic was done in accordance to the application areas of the primary, secondary or tertiary sector.

[347] http://openminted.eu/
[348] http://www.sci-gaia.eu/
[349] https://rd-alliance.org/
[350] http://www.metacardis.net/
[351] http://rd-connect.eu/
[352] http://iri.jrc.es/
[353] http://scycom.collide.info/
[354] http://next-tell.eu/

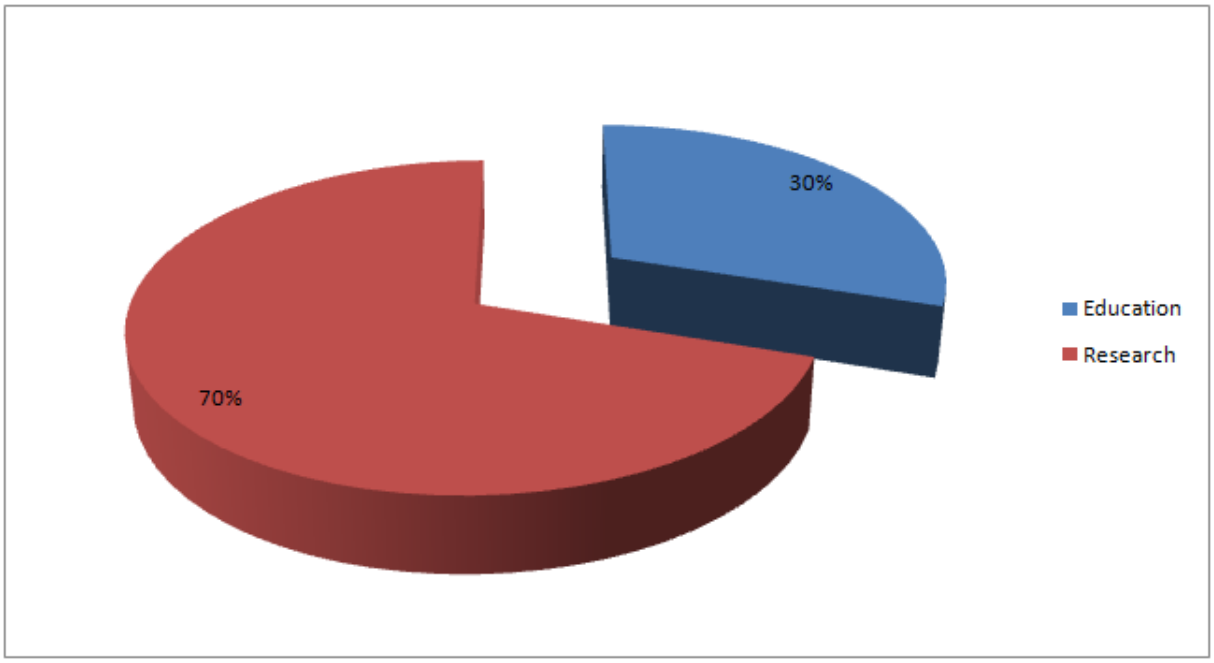


Figure 36. Publications in the quaternary sector

3.10 Quinary sector

3.10.1 Introduction to the sector

The quinary sector is the highest economic sector encompassing decision making for policy guidelines in Industry, Government, Science and Technology, having a profound impact on economy in general. It is the special characteristics of this sector and the interrelation with all the other economic sectors, which make it challenging to isolate resources and material pertaining exclusively to the quinary sector itself. Decision makers need all the available data from various types of infrastructures as well as all the ***TDM*** software tools and technologies in order to analyse, deploy them, and finally make decisions affecting different application areas within all previous sectors. Taking these peculiarities into consideration, in this report we present only the EU projects which have been funded under certain RTD Horizontal Topics, specifically those which aim at the formulation or identification of future developments in RTD and long-term strategic options.

3.10.2 EU research projects relevant to the quinary sector

A total amount of **192,1 M€**, with a European Commission maximum contribution of **149 M€**, has been invested on **116** projects concerning ***TDM*** in the **quinary** sector. Most projects focus on the evaluation or application of laws and regulations, policies, policy strategies or plans of action for research and development in science and technology. The following projects have been selected, with an emphasis on **decision making** as part of their **title** or **objective**, as the most representative for the quinary sector:

- KHRESMOI[355]: Knowledge Helper for Medical and Other Information users
- PSIP[356]: Patient Safety Through Intelligent Procedures In Medication

[355] http://www.khresmoi.eu/

- LOD2[357]: LOD2 - Creating Knowledge out of Interlinked Data
- ARENA[358]: Architecture for the Recognition of thrEats to mobile assets using Networks of multiple Affordable sensors
- FIRST[359]: Large scale information extraction and integration infrastructure for supporting financial decision making
- Dicode[360]: Mastering Data-Intensive Collaboration and Decision Making
- ADVANCE[361]: Advanced predictive-analysis-based decision-support engine for logistics

[356] http://cordis.europa.eu/project/rcn/85437_en.html
[357] http://lod2.okfn.org/
[358] https://www.tno.nl/en/focus-area/defence-safety-security/national-security-crisis-management/arena/
[359] http://project-first.eu/content/large-scale-information-extraction-and-integration-infrastructure-supporting-financial-decis
[360] http://dicode-project.eu/
[361] http://www.advance-logistics.eu/

4 Challenges of Text and Data Mining in Europe

In the previous sections the landscape of TDM in Europe has been painted, shedding light to various aspects, notably on the tremendous data growth, the wide variety of ***Text and Data Mining*** tasks, the data as well as the technologies available for TDM (internationally and at the European level), the commercial activity around TDM and, finally, the EU research activities concerned with TDM, as regards publications and R&D projects/infrastructures per economic sector.

This huge, rapidly evolving, multi-faceted and multi-player domain faces many challenges, which need to be first recognized as such and, hopefully, successfully tackled. These challenges are of technical and legal nature and also touch upon policy issues.

4.1 Technical challenges

As presented in the sections above, the TDM tools, services and infrastructures evolve in parallel with the growing data.

Despite the rapid growth, the domain does not seem to have reached a maturity level of convergence; it is characterized by **fragmentation** of sources, which are variably deposited and organized in a high number of repositories and aggregators, employing varying techniques, methods and implementation strategies.

The ensuing required **interoperability** of tools and datasets at syntactic and semantic level is still a moving target.

The issue of **persistent identification** of datasets and tools/services is still a research issue, despite some recently emerging good practices, like handles and DOIs. Although storage capacity has tremendously increased, the issue of **persistent storage** of data has not been successfully tackled. Closely connected to identification and discovery of data and tools is the extensive use of appropriate **metadata**; curation and maintenance of data and metadata is of utmost importance. Interoperability of metadata schemas is also open. Finally, the importance of language technology resources (data and tools) for all EU languages and the development and maintenance of related infrastructures offering such services is considered indispensable.

4.2 Legal and regulatory issues[362]

The legal challenges are tightly interrelated with the issue of data and tools accessibility. Despite the obvious fact that data and tools have more value when shared, used, re-used and re-purposed, for scientists, entrepreneurs and broad public, openness is far from being established. It is true that openness of data requests a big change in attitudes, methodologies, and scientific, commercial and governmental practices.

The challenge of securing the benefits of data sharing **without compromising the rights of all parties involved** (legal, ethical/personal, financial) needs an urgent answer. In parallel, **personal, private or sensitive data** need to be highly respected; any such data (e.g. medical records, trial

[362] For a detailed discussion on legal issues, please refer to FutureTDM Deliverable D3.3.

participations in new treatment protocols, etc.) should be made available after processing through careful anonymization processes.

The very nature of ***TDM*** presupposes access to big amounts of data; however, this stumbles on the size-limit for retrieval of scientific publications imposed by publishing houses striving to protect their assets. ***Text and Data Mining*** does not copy data for reproducing or redistributing it, but to process and analyse it. Building models for ***TDM*** (e.g. language models) requires access to big amounts of data; this process entails that the original data loses its integrity and (in most cases) cannot be reproduced; therefore, ***TDM*** should not be viewed as acting competitively to the original data distributors' interests.

Even if not directly negative towards TDM, many data creators and providers display **legal agnosticism**: by not making explicit the terms and conditions of use of the data, they hold TDM researchers and commercial developers hostage. Last but not least, the **multitude of licenses** for data sharing discourages providers and users alike. Despite the existence of fairly standardized licences (Creative Commons[363], AGPL[364], ApacheLicence_2.0[365], BSD[366], GFDL[367], GPL[368] and LGPL[369]), the proliferation of ad hoc licensing documents, coupled with the variety of proprietary licenses, renders the complexity of legally sharing data enormous.

4.3 Policy issues

It is clear that the EU member states do not have a uniform approach as regards data. There are **differences in the legal framework** across countries for the treatment of data (either public data produced by government organisations or privately owned).

Despite the acknowledgement of the high value of data, at both public and private sector levels, the **absence of data management plan** dominates. Sometimes even the very notion is neglected.

A number of very useful applications has already been built out of the comparatively few public data that the governments make available (e.g. geographical and transport network data). Although the importance of openness of data is acknowledged, the issue of **open data is not as yet central** in **governmental policies**, judging from the relevant legislation, its implementation tools but also form the amount of funding invested on TDM technologies.

The fact that TDM and related subjects are absent from current curricula in most countries results in skills shortage and far from promotes the TDM domain.

[363] http://creativecommons.org/licenses
[364] http://www.gnu.org/licenses/agpl-3.0.html
[365] http://www.apache.org/licenses/LICENSE-2.0.html
[366] https://opensource.org/licenses/BSD-2-Clause
[367] http://www.gnu.org/copyleft/fdl.html
[368] https://opensource.org/licenses/gpl-license.php
[369] http://www.gnu.org/licenses/lgpl.html

5 Limitations of this Research

The limitations of this research are, in a nutshell, typical problems of ***Text and Data Mining*** research. Specifically:

- Data is dynamic and flows in on a daily basis. It is clear (and it has been stated many times in the present study) that this overview claims to be an indicative snapshot of the current state of affairs in the TDM domain. It is time constrained: more data and tools will appear tomorrow which are not present in this overview[370].
- In many cases data is contributed by players in the field(s) in a crowdsourcing manner, resulting in some cases in incomplete data or data in need of curation.
- The overall landscape is fragmented, as is also the landscape for individual components of this research, which, for reasons of completeness, is based on multiple sources for data and content, tools and technologies, projects and publications. Multiplicity of sources entails semantic interoperability problems with respect to the classification schemes, even by the same source. Indicatively:
    - The EU FP7 project classification scheme was different than the one used in the Horizon 2020 framework.
    - The various repositories investigated (CORE, Science Direct, Web of Science) utilise different schemes for domain classification.
    - The analysis performed by different researchers and presented in this study (or reviewed as literature) was grounded on different bases:
        - Different sources of data, i.e. Web of Science, Science Direct, OpenAIRE/CORE
        - Different bodies of data within sources, i.e. analysis on titles only, abstracts only, some on both, some including full text where available
        - Different queries: in measuring the frequency of terms relevant to TDM, some researchers focused on only *Text Mining/Data Mining/Text and Data Mining*, while others on a few more terms. In this report, the terms researched into were *Text Mining/Data Mining/Text and Data Mining* plus a set of other conceptually subsumed and related terms (*Semantic Analysis, Sentiment Analysis, Opinion Mining, Information Extraction,* etc.)

[370] Actually, at the time of proofreading this report (08h00 CET 31 MAY 2016), the Lisbon Council launched **Text and Data Mining for Research and Innovation: What Europe Must Do Next**, "an interactive policy brief which looks at the challenge and opportunity of text and data mining in a European context".,

6 Conclusions

In the present report we aimed at describing the status quo of ***TDM*** in Europe along four dimensions:

- the overall situation as regards the area of ***Text and Data Mining*** (research issues, stakeholders, recent developments);
- the existing idiosyncrasies of the European situation;
- the existing resources, technologies, infrastructures and related scientific production (in the form of publications);
- the results and the impact on everyday life, as depicted by the mapping of the above into the five economic sectors.

Text and Data Mining is a powerful new method both for academic research and commercial offerings. Its value and most notably the importance of the mineable data and themselves are widely recognised in the EU, as attested by the Open Access policies, the emphasis on data depositing, aggregation and sharing as well as the prioritisation of services over such data by all stakeholders along the value chain. Accessibility, sharing, reuse and generation of new insights out of big volumes of data are, however, hampered by a rather vague and incomprehensive legal framework[371]. Researchers and companies in the EU are well aware of and strongly engaged in TDM activities, while an increasing shortage of TDM skilful scientists is observed.

Infrastructure and technology wise, the landscape is fragmented. There is an obvious proliferation of data depositing and documentation artifacts (repositories, metadata models, aggregators) leading to interoperability problems. Interoperability is also a requirement within the TDM technological proper, as researchers are increasingly interested in trying and combining components and modules in new technical architectures, pipelines and systems. Multilinguality, one of Europe's central and cultural assets, is an additional dimension that needs to be taken into account in a coordinated way, especially in view of the Digital Single Market vision.

A substantial investment is already being made by the EU in R&D and infrastructural projects in a wide range of areas. Coordination and uptake of R&D results is a requirement, especially for new interdisciplinary engagements where crossing the boundaries of and building bridges to other communities is considered *sine qua non*.

Commercial activity around TDM is picking up in Europe, not unexpectedly concentrated in the big EU countries (UK, Germany, France, Spain, Italy) and geared towards vertical applications and various types of analytics services.

[371] An ananlysis opf the somewhat obscure legal framework in the EU and Member States appears in FutureTDM Deliverable 3.3 Baseline report of policies and barriers of TDM in Europe.

7 References

- Cunningham, S.J. and Holmes, G., 1999. Developing innovative applications in agriculture using data mining. In The proceedings of the Southeast Asia regional computer confederation conference (pp. 25-29).
- Eskevich, M., van den Bosch, A., Caspers, M., Guibault, L., Bertone, A., Reilly, S., Munteanu, C., Leitner, P., Piperidis, St., 2016. FutureTDM Deliverable D3.1 Research Report on TDM Landscape, Available at: http://project.futuretdm.eu/wp-content/uploads/2016/05/D3.1-Research-Report-on-TDM-Landscape-in-Europe-FutureTDM.pdf
- European Strategy Forum on Research Infrastructures, 2016. Strategic Report on Research Infrastructures, Roadmap 2016. Available at: http://www.esfri.eu/roadmap-2016
- Ferrucci, D., and Lally, A. 2004. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Natural Language Engineering, 10(3-4), 327–348. http://doi.org/10.1017/s1351324904003523
- Filippov, S. 2014. Mapping Text and Data Mining in Academic and Research Communities in Europe. Brussels: Lisbon Council. Available at: http://www.lisboncouncil.net/component/publication/publication/109-mapping-text-and-data-mining-in-academic-and-research-communities-in-europe.html
- Gavrilidou, M., Labropoulou, P., Desipri, E., Giannopoulou, I., Hamon, O., & Arranz, V. 2012. The META-SHARE Metadata Schema: Principles, Features, Implementation and Conversion from other Schemas. In Workshop “Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR”. Available at: LREC2012 (pp. 5–12). http://www.lrec-conf.org/proceedings/lrec2012/workshops/11.LREC2012%20Metadata%20Proceedings.pdf
- Hager, K.M. & Gu, W. 2014. Understanding the non-canonical pathways involved in p53-mediated tumor suppression. Carcinogenesis, vol. 35(4), pp.740–746.
- Han, J., Kamber, M., Pei, J. 2001. Data mining: concepts and techniques. Morgan Kaufmann. p.5.
- Hatler, M., Gurganious, D. and Chi, C. 2012. Industrial wireless sensor networks: A market dynamics report. ON World. San Diego, CA, USA.
- Keim, D., Kohlhammer, J., Ellis, G. & Mansmann, Fl. (Eds.). 2010. Mastering the Information Age: Solving Problems with Visual Analytics, Eurographics Association.
- Knoth, P. and Zdrahal, Z. 2012. CORE: Three Access Levels to Underpin Open Access, D-Lib Magazine, 18, 11/12, Corporation for National Research Initiatives, 10.1045/november2012-knoth
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, Ch. and Byers, H.A. 2011. Big Data: The next frontier for innovation, competition and productivity. McKinsey Global Institute.
- Monachini, M., Quochi, V., Calzolari, N., Bel, N., Budin, G., Caselli, T., Choukri, K., Francopoulo, G., Hinrichs, E., Krawuer, S., Lemnitzer, L., Mariani, J., Odijk, J., Piperidis, S., Przepiorkowski, A., Romary, L., Schmidt, H., Uszkoreit, H., Wittenburg, P. 2011. The

Standards' Landscape towards an Interoperability Framework. Available at: http://www.flarenet.eu/sites/default/files/FLaReNet_Standards_Landscape.pdf

- OpenMinTeD Deliverable 5.1: Interoperability Landscaping Report, 28 December 2015.
- Piperidis, S. 2012. The META-SHARE language resources sharing infrastructure: Principles, challenges, solutions. In LREC 2012. In Proceedings of the 8th International Conference on Language Resources and Evaluation. Available at: http://www.lrec-conf.org/proceedings/lrec2012/summaries/1086.html
- Rehm, G., Uszkoreit, H. (editors) 2013. *META-NET Strategic Research Agenda for Multilingual Europe 2020.* Heidelberg, New York etc.: Springer, 2013.
- Simpson, M. S., & Demner-Fushman, D. 2012. Biomedical text mining: A survey of recent progress. In Mining Text Data (pp. 465–517). Springer.
- Spangler, Sc. et al. 2014. Automated hypothesis generation based on mining scientific literature. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14). ACM, New York, NY, USA, 1877-1886. DOI = http://dx.doi.org/10.1145/2623330.2623667
- Triaille, J.P., de Meeûs d'Argenteuil, J. & de Francquen, A. 2014. Study on the legal framework of Text and Data mining (TDM), European Commission. Available at: http://ec.europa.eu/internal_market/copyright/docs/studies/1403_study2_en.pdf
- Tsai, Hsu-Hao. 2012. Global Data Mining: An Empirical Study of Current Trends, Future Forecasts and Technology Diffusions. Expert Systems with Applications, 39(9), pp. 8172–8181.
- Turner, V., Gantz, J.F., Reinsel, D., and Minton, St. 2014. The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. Available at: https://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf
- Xuezhong, Zhou, Yonghong, Peng & Baoyan, Liu. 2010. Text mining for traditional Chinese medical knowledge discovery: a survey. Journal of Biomedical Informatics.

8 ANNEX A: TABLES OF EU FUNDED RESEARCH INFRASTRUCTURES

HORIZON 2020 RESEARCH INFRASTRUCTURES

This table lists the RIs of the Horizon 2020 Programme, ordered by total cost per sector. The RIs are accompanied by

- administrative information, namely title and investment (total cost and EU contribution)
- information on their domain of application (as signified by keywords extracted from their descriptions),
- mapping to economic sector and
- relation to TDM (TDM related term extracted by their objectives description).

Acronym	Title	Domain keywords	Cost	EU Contribution	TDM keyword	sector
EarthServer-2	Agile Analytics on Big Data Cubes	Agile Analytics; Big Earth Data Cubes; sensor, image, simulation, statistics data; open standards; rasdaman Array Database technology; automatic data and query distribution; tape archive integration; 3D/4D visualization; Big Data standards; dissemination & exploitation	2839743,75	2839743	large data, big data	1
AHEAD	Integrated Activities for the High Energy Astrophysics Domain	High Energy Astrophysics; space observatories for high-energy astrophysics; space-based sensors and electronics; data analysis; technologies; background modeling; cross calibration; feasibility studies of space-based instrumentation; X-ray and gamma-ray missions; public outreach	5085247	4982477	data analysis	2
COEGSS	Center of Excellence for Global Systems Science	Global Systems Science; global coordination problems; ICT; High Performance Data Analysis; customized synthetic populations for GSS applications; HPC; real-time assessment of global risks and opportunities	4469662,5	4469662,5	data analysis	2

Acronym	Title	Domain keywords	Cost	EU Contribution	TDM keyword	sector
West-Life	World-wide E-infrastructure for structural biology	structural biology; macromolecules; prokaryotic organisms; macromolecular machinery of higher organisms; human health; experimental facilities; Protein Data Bank; public repository; pipelines for data analysis and structure determination; integrated management of structural biology data; metadata; application level service	3981125	3981125	data analysis	2
INDIGO-DataCloud	INtegrating Distributed data Infrastructures for Global ExplOitation	middleware; service models; user tools; Big Data; scientific computing; Cloud computing, storage and network; PaaS; SaaS; open source solutions; Grid based infrastructures; HPC clusters; GEANT-compliant federated and distributed AA policies	11138114	11138114	big data	3
PhenoMeNal	PhenoMeNal: A comprehensive and standardised e-infrastructure for analysing medical metabolic phenotype data	metabolome data; genotype; phenome; exposome; personalised evidence-based medicine; metabolic information; molecular phenotyping; personal genome; biomedical data management; information-mining; metabolomics applications	8018723,75	7697733,75	data analysis	3
Global BioImaging	Global BioImaging Project - International imaging infrastructure services for the life science community	bioImaging; biological and medical imaging technologies; biological and medical sciences; imaging facility management and operation; quality management; training activities; image data analysis software; open access	1780585	1780585	data analysis	3
EUDAT2020	EUDAT2020	data management; data preservation, access and sharing; trusted environment; Collaborative Data Infrastructure; data access and deposit; archiving; identification, discoverability, computability; long-tail and big data; research data	19052882	18865385	big data	4
EGI-Engage	Engaging the EGI Community towards an Open Science Commons	Open Science Commons vision; innovative digital services, data, knowledge and expertise; e-Infrastructure Commons; Open Data Commons; Knowledge Commons; improved cloud or data services; network of Competence Centres; requirements; community-specific applications; interoperability across e-Infrastructures; user-centric development model; ad hoc access policies	8662501,5	8000000	data analysis	4
OpenMinTeD	Open Mining INfrastructure for TExt and Data	digital research data; Text and data mining; content; scientific publications; interoperability; training; different scientific areas; generic scholarly communication; life sciences; food and agriculture; social sciences and humanities; legal experts	6068072,5	5375535,5	text and data mining, text mining	4

Acronym	Title	Domain keywords	Cost	EU Contribution	TDM keyword	sector
SoBigData	SoBigData Research Infrastructure	Social Mining & Big Data Ecosystem; ethic-sensitive scientific discoveries; social data mining; mathematics; ICT; human, social and economic sciences; training; networking; open science; openly available datasets	5917500	5000000	data mining, big data	4
NoMaD	The Novel Materials Discovery Laboratory	computational materials science; HPC; data sharing; physical, materials, and quantum-chemical sciences; big-data analytics; services; petascale-exascale computations	4910624,48	4910624,48	data analytics	4
EDISON	Education for Data Intensive Science to Open New science frontiers	Data Science; Data Intensive/Big Data technologies; sustainability/business model; education and training; certification; example datasets; virtual labs	2500000	2500000	large data, big data, data science	4
Sci-GaIA	Energising Scientific Endeavour through Science Gateways and e-Infrastructures in Africa	guides and educational documents; training and support; Science Gateways; e-Infrastructures services; open access linked data; training and dissemination workshops	1339125	1339125	linked data	4

ESFRI RESEARCH INFRASTRUCTURES

This table lists the RIs of the FP7 Programme, ordered by capital value per sector. The RIs are accompanied by

- administrative information, namely type of project, title, investment (cost) and URL,
- information on their domain of application (manually assigned),
- mapping to economic sector, and
- information on their domain of application (as signified by keywords extracted from their descriptions).

Type	ACRONYM	Title	Domain	Keywords	CAPITAL VALUE (M€)	URL	sector
ESFRI project	SIOS	Svalbard Integrated Arctic Earth Observing System	environment	regional observational system; Earth System Science; Global Change; remote sensing resources; pan-Arctic observational structure; regional modelling	N/A	http://www.sios-svalbard.org/	1

Type	ACRONYM	Title	Domain	Keywords	CAPITAL VALUE (M€)	URL	sector
ESFRI project	EPOS	European Plate Observing System	environment	solid Earth System; data, models and facilities; new concepts and tools for geo-hazards and geo-resources applications to environment and to human welfare	500	https://www.epos-ip.org/	1
ESFRI project	ACTRIS	Aerosols, Clouds and Trace gases Research Infrastructure	environment	aerosols, clouds, gases; precision data, services and procedures; short-lived atmospheric species; evolution of the atmospheric environment	450	http://www.actris.eu/	1
ESFRI project	DANUBIUS-RI	International Centre for Advanced Studies on River-Sea Systems	environment	large river-sea systems; environmental, social and economic sciences; knowledge exchange, harmonised data; interdisciplinary research, education and training	300	http://www.danubius-ri.eu/about-us/	1
ESFRI project	EISCAT_3D	Next generation European incoherent scatter radar system	environment	radar system; Earth's atmosphere coupled to space; incoherent scatter radars	128	https://eiscat3d.se/	1
ESFRI landmark	EMSO	European Multidisciplinary Seafloor and water-column Observatory	environment	ocean observation systems; real-time monitoring of enviromental processes; natural hazards, climate change; marine ecosystems; Arctic; Atlantic; Mediterranean; Black Sea	108	http://www.emso-eu.org/	1
ESFRI landmark	LifeWatch	e-infrastructure for Biodiversity and Ecosystem Research	environment	biodiversity research; knowledge-based strategic solutions to environmental preservation; Virtual Research Environments	66	http://www.bbmri-eric.eu/	1
ESFRI landmark	ICOS ERIC	Integrated Carbon Observation System	environment	carbon cycle; greenhouse gas budget; perturbations; carbon portal	48	https://www.icos-ri.eu/	1
ESFRI landmark	IAGOS	In-service Aircraft for a Global Observing System	environment	atmospheric composition; aerosol and cloud particles; commercial aircraft	25	http://www.iagos.org/	1
ESFRI landmark	EURO-ARGO ERIC	European contribution to the international Argo Programme	environment	in-situ ocean observations; global array of profiling floats; temperature; salinity; climate change research and monitoring	10	http://www.euro-argo.eu/	1

Type	ACRONYM	Title	Domain	Keywords	CAPITAL VALUE (M€)	URL	sector
ESFRI project	EST	European Solar Telescope	physical sciences and engineering	high-resolution solar physics; website; solar telescope	N/A	http://www.est-east.eu/	2
ESFRI project	WindScanner	European WindScanner Facility	energy	wind and turbulence fields; wind energy	45-60	http://www.windscanner.eu/	2
ESFRI landmark	ESRF UPGRADES	Phase I Phase II: Extremely Brilliant Source	physical sciences & engineering	hybrid multi-bend achromat lattice design; source brilliance and coherence; beamlines; detector projects; data management and analysis strategy; serial crystallography; structural biology and material science	150 (+)	http://www.esrf.eu	2
ESFRI landmark	European Spallation Source ERIC	European Spallation Source	physical sciences & engineering	neutrons; neutron peak brightness; interdisciplinary research in physical and life sciences; on-site Accelerator installations; Accelerator beam on the Target	1843	http://www.europeanspallationsource.se	2
ESFRI project	MYRRHA	Multi-purpose hYbrid Reactor for High-tech Applications	energy	innovative nuclear research reactor; Accelerator Driven System; radioactive waste; neutron-irradiated silicon; renewable energy; radioisotopes for medical applications; fast spectrum reactor; fusion technology	1500	http://myrrha.sckcen.be/	2
ESFRI landmark	European XFEL	European X-Ray Free-Electron Laser Facility	physical sciences & engineering	high repetition rate ultra-short X-ray flashes; brilliance; synchrotron X-ray radiation sources; atomic details of viruses; molecular composition of cells; three-dimensional images of the nanoworld; chemical reactions; processes “under extreme conditions”	1490	http://www.xfel.eu	2
ESFRI landmark	HL-LHC	High-Luminosity Large Hadron Collider	physical sciences & engineering	highest-energy particle collider in the world; Higgs boson; new physics at the energy frontier; High-Luminosity LHC; increased data rates; particle physics	1370	http://home.cern/	2

Type	ACRONYM	Title	Domain	Keywords	CAPITAL VALUE (M€)	URL	sector
ESFRI landmark	FAIR	Facility for Antiproton and Ion Research	physical sciences & engineering	accelerator complex; beams of antiprotons and ions; dynamics of matter under extreme conditions; evolution of the Universe; nucleosynthesis in stars; star explosions; injector chain; nuclear structure; nuclear astrophysics; physics of hadrons; fundamental physics with antiproton beams; physics of compressed nuclear matter; plasma physics; atomic physics; materials research; biomedical applications	1262	http://www.fair-center.de	2
ESFRI project	ECCSEL	European Carbon Dioxide Capture and Storage Laboratory Infrastructure	energy	Carbon Capture and Storage technologies; CO2 emissions; global climate change	1000	http://www.eccsel.org/	2
ESFRI landmark	JHR	Jules Horowitz Reactor	energy	material and fuel behaviour in extreme nuclear environment; irradiation loops; power reactor technologies; nuclear fuel; materials used in reactors; radioisotopes for use in medicine	1000	http://www.cad.cea.fr/rjh/	2
ESFRI landmark	E-ELT	European Extremely Large Telescope	physical sciences & engineering	Extremely Large Telescope; planets around other stars; first objects in the Universe; super-massive black holes; dark matter and dark energy; optical/near-infrared telescope	1000	http://www.eso.org/public/teles-instr/e-elt/	2
ESFRI landmark	ELI	Extreme Light Infrastructure	physical sciences & engineering	extreme light-matter interactions; attosecond resolution of coherent radiation; laser-accelerated particles; atomic, molecular, plasma and nuclear physics; biology; chemistry; medicine; astrophysics	850	http://www.eli-laser.eu/	2
ESFRI landmark	SKA	Square Kilometre Array	physical sciences & engineering	radio telescope; first structures in the Universe; physics; gravity; cosmic magnetism; origins of life	650	http://www.skatelescope.org	2

Type	ACRONYM	Title	Domain	Keywords	CAPITAL VALUE (M€)	URL	sector
ESFRI project	CTA	Cherenkov Telescope Array	physical sciences and engineering	high-energy gamma-ray astronomy; cosmic non-thermal processes; understanding of astrophysical and cosmological processes	400	https://portal.cta-observatory.org/	2
ESFRI landmark	ILL 20/20	Institut Max von Laue-Paul Langevin	physical sciences & engineering	neutron science and technology; condensed matter physics; chemistry; biology; nuclear physics; materials science; intense reactor source; signal to noise performance	171	http://www.ill.eu	2
ESFRI landmark	EMFL	European Magnetic Field Laboratory	physical sciences & engineering	magnetic fields; scientific research infrastructures; pulsed fields	170	http://www.emfl.eu/	2
ESFRI project	KM3NeT 2.0	KM3 Neutrino Telescope 2.0: Astroparticle & Oscillations Research with Cosmics in the Abyss	physical sciences and engineering	KM3 Neutrino Telescope; astrophysical objects ; highenergy neutrino emission; cosmic-ray interactions; deep-sea installations; earth and sea sciences; marine biology, oceanography and environmental sciences	137	http://www.km3net.org/	2
ESFRI project	EU-SOLARIS	European SOLAR Research Infrastructure for Concentrated Solar Power	energy	Concentrating Solar Thermal and Solar Chemistry technologies; solar energy	120	http://eusolaris.eu/	2
ESFRI landmark	SPIRAL2	Système de Production d'Ions Radioactifs en Ligne de 2e génération	physical sciences & engineering	Radioactive Ion Beam physics; hadron and isotope therapy; physics of the atom and its nucleus; condensed matter; astrophysics; properties of nuclei; contemporary nuclear physics; astrophysics; interdisciplinary research; r and rp-process nuclei; shell closure; very heavy elements; material sciences; radiobiology; energy; environment; social sciences; health; engineering; space; ICT; radiobiology	110	http://www.ganil-spiral2.eu	2

Type	ACRONYM	Title	Domain	Keywords	CAPITAL VALUE (M€)	URL	sector
ESFRI project	ERINHA	European research infrastructure on highly pathogenic agents	health and food	highly pathogenic micro-organisms; biosafety; biosecurity procedures; standards for management of biological resources; group 4 pathogens; training	N/A	http://www.erinha.eu/	3
ESFRI project	EU-OPENSCREEN	European Infrastructure of Open Screening Platforms for Chemical Biology	health and food	novel small chemical compounds; organisms; cells; cellular components; novel therapeutic targets; cellular physiology; chemical biology; open-access database	N/A	http://www.eu-openscreen.eu/	3
ESFRI project	Euro-BioImaging	European Research Infrastructure for Imaging Technologies in Biological and Biomedical Sciences	health and food	biological and medical imaging; life sciences; image data support; data management; training activities	N/A	http://www.eurobioimaging.eu/	3
ESFRI project	ISBE	Infrastructure for Systems Biology Europe	health and food	systems biology; facilities; data; models; tools; training; standardisation of biological data, tools, models, operating procedures	N/A	http://project.isbe.eu/	3
ESFRI project	MIRRI	Microbial Resource Research Infrastructure	health and food	biotechnology; mBRC activity; authentic microbial resources and associated data; legally compliant framework; bioresource holdings	N/A	http://www.mirri.org/home.html	3
ESFRI landmark	BBMRI ERIC	Biobanking and BioMolecular resources Research Infrastructure	health and food	bio-medical research; human health/disease-relevant biological resources; associated data; ethically and legally compliant; common standards; biobanking	170-220	http://www.bbmri-eric.eu/	3
ESFRI landmark	EATRIS ERIC	European Advanced Translational Research Infrastructure in Medicine	health and food	translational medicine; diagnosis; biomedical innovations; clinical needs; large and diverse clinical patient cohorts	500	http://www.eatris.eu/	3

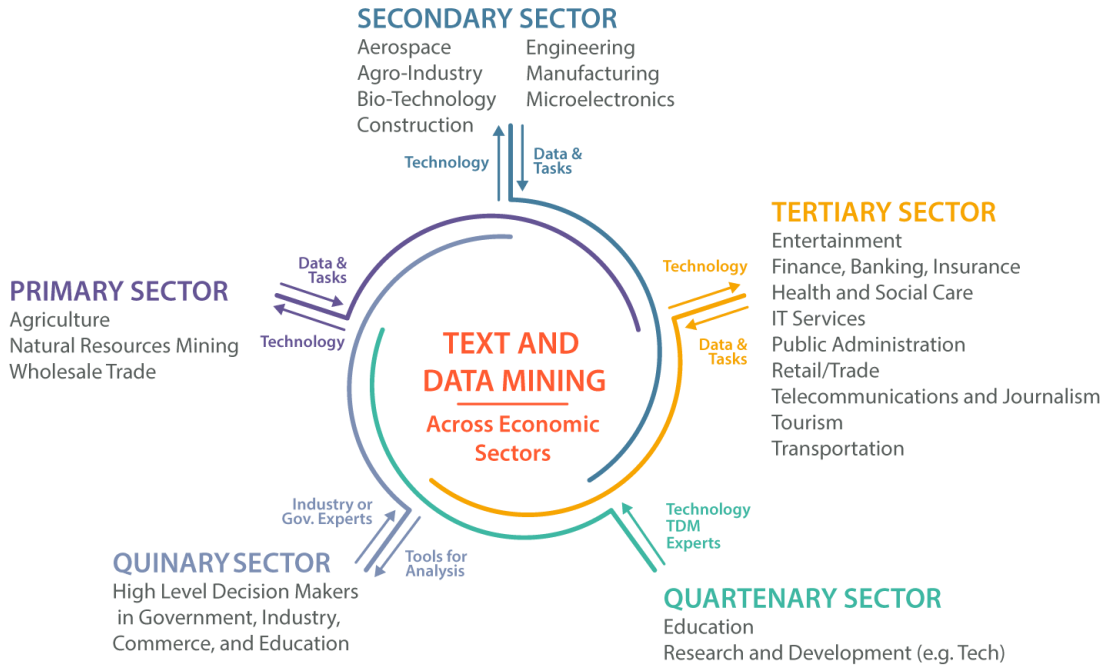
Type	ACRONYM	Title	Domain	Keywords	CAPITAL VALUE (M€)	URL	sector
ESFRI landmark	INSTRUCT	Integrated Structural Biology Infrastructure	health and food	novel small chemical compounds; organisms; cells; cellular components; novel therapeutic targets; cellular physiology; chemical biology; open-access database	285	http://www.structuralbiology.eu	3
ESFRI landmark	INFRAFRONTIER	European Research Infrastructure for the generation, phenotyping, archiving and distribution of mouse disease models	health and food	mouse disease models; biomedical research; organismic effects of genetic alterations; gene function; mouse models; research tools and associated data; phenotype analyses; medical research; cancer, metabolic and cardiovascular diseases; lung diseases; infectious diseases; rare diseases; global threats	180	http://www.infrafrontier.eu	3
ESFRI project	AnaEE	Infrastructure for Analysis and Experimentation on Ecosystems	health and food	ecosystems; sustainable future; climate change; land use; agricultural systems; climate mitigation strategies; food security; bio-economy	135,5	http://www.anaee.com/	3
ESFRI project	EMPHASIS	European Infrastructure for multi-scale Plant Phenomics and Simulation for food security in a changing climate	health and food	plant phenotyping; web-based platform; workshops; symposia	135	http://www.plant-phenotyping.org/	3
ESFRI project	EMBRC	European Marine Biological Resource Centre	health and food	marine science; marine biology; ecology; blue biotechnologies	126	http://www.embrc.eu/	3
ESFRI landmark	ELIXIR	A distributed infrastructure for life-science information	health and food	life-science; bioinformatics resources; data-related needs; marine research; plants; agriculture; health research; medical sciences	125	http://www.elixir-europe.org	3
ESFRI landmark	ECRIN ERIC	European Clinical Research Infrastructure Network	health and food	transparent clinical trials; interoperability of clinical research environment; diversity; decision in medical practice; sound scientific evidence; high-quality clinical research	1,5	http://www.ecrin.org/	3
ESFRI project	E-RIHS	European Research Infrastructure for Heritage Science	social and cultural innovation	Heritage Science; cross-disciplinary research; global heritage; access to infrastructures, methodologies, data and	N/A	www.e-rihs.eu	4

Type	ACRONYM	Title	Domain	Keywords	CAPITAL VALUE (M€)	URL	sector
				tools; training; public engagement; repositories for standardized data storage, analysis and interpretation			
ESFRI landmark	CESSDA	Consortium of European Social Science Data Archives	social and cultural innovation	social sciences; data archives; social, economic and political research	N/A	http://www.cessda.net	4
ESFRI landmark	CLARIN ERIC	Common Language Resources and Technology Infrastructure	social and cultural innovation	Language Resources and Technology; humanities and social sciences; digital language data; advanced tools and services; language data repositories; interoperability of data and tools	N/A	http://www.clarin.eu	4
ESFRI landmark	ESS ERIC	European Social Survey	social and cultural innovation	data on social attitudes and behaviours; social stability and change; citizen involvement and democracy; family and working life; personal and social wellbeing; ageism; trust in institutions; attitudes to climate change and energy security; welfare state; immigration; health inequalities	N/A	http://www.europeansocialsurvey.org	4
ESFRI landmark	PRACE	Partnership for Advanced Computing in Europe	e-RI	supercomputing; computing; data resources; services; large-scale scientific and engineering applications; data management resources and services; High Performance Computing; energy efficiency of computing systems; environmental impact	500	http://www.prace-ri.eu/	4
ESFRI landmark	SHARE ERIC	Survey of Health, Ageing and Retirement in Europe	social and cultural innovation	Health, Ageing and Retirement; multidisciplinary database of microdata; health; socio-economic status; social and family networks; demographic ageing; social policy	110	http://www.share-project.org/	4

Type	ACRONYM	Title	Domain	Keywords	CAPITAL VALUE (M€)	URL	sector
ESFRI landmark	DARIAH ERIC	Digital Research Infrastructure for the Arts and Humanities	social and cultural innovation	digital Arts and Humanities; digitally enabled research and teaching; network of people, expertise, knowledge; digital resources; best practices; methodological and technical standards	4,3	http://www.dariah.eu	4

9 ANNEX B: ECONOMIC SECTORS AND APPLICATION AREAS

Annex B presents the sectors defined in D3.1, with their respective application areas (see D.3.1, Figure 5. General economic structure and connection between TDM and all economic sectors, p.22).



"Since the beginning of humanity, the development of novel technologies has defined the society structure, and its employment distribution. Agriculture and natural resources extraction fits into the primary sector, as this is an initial type of human joint efforts to produce food for survival. Invention of machines and elaborated tools led to the development of the secondary sector which evolves these days into complex machinery engineering, constructions, and space exploration. The third sector comprises all the services that society members deliver to each other, varying from retail to transportation and from medical healthcare support to entertainment. Until recently this three components structure represented most of the societies in the world. However, the growth of data driven business, services and research, brought a discussion on the changes in the economic structure that should reflect a novel type of knowledge-based economy, where knowledge and data become a separate valuable commodity (OCDE, 1996)[372]. Thus, recently the research that supports knowledge sharing and growth, as well as education of the professionals that can carry out these activities, are put aside as a quaternary sector of economy. Moreover, the high level decision makers in governments, large industry companies and education, who have potential to shape the future of the entire sectors with their vision and following decisions, are placed into a separate, quinary sector."

[372] https://www.oecd.org/sti/sci-tech/1913021.pdf

10 ANNEX C: FREQUENCY OF TDM RELATED TERMS IN FP7 AND HORIZON 2020 PROJECTS

Projects retrieved from FP7 and Horizon 2020 objectives using a list of TDM-related search queries.

Keywords in OBJECTIVE	FP7 results	keywords in OBJECTIVE	Horizon 2020 results
data analysis	**240**	big data	**66**
machine learning	**156**	data analysis	**61**
data mining	**88**	machine learning	**51**
large data	**65**	data analytics	**21**
language processing	**39**	**data mining**	**14**
big data	**38**	large data	**14**
information retrieval	**37**	linked data	**6**
information extraction	**30**	business intelligence	**5**
linked data	**30**	data science	**5**
text mining	**17**	language processing	**5**
data science	**19**	information retrieval	**4**
data analytics	**18**	predictive analytics	**4**
information access	**17**	information access	**3**
summarization	**12**	sentiment analysis	**2**
business intelligence	**11**	**text and data mining**	**2**
knowledge discovery	**11**	information extraction	**1**
multimedia processing	**8**	knowledge discovery	**1**
conversation analysis	**7**	summarization	**1**
opinion mining	**7**	TDM	**1**
sentiment analysis	**5**	abbreviation detection	**0**
semantic mining	**4**	competitive intelligence	**0**
text analytics	**3**	concept extraction	**0**
content mining	**3**	content classification	**0**
text understanding	**3**	content mining	**0**
trend detection	**2**	conversation analysis	**0**
graph mining	**2**	datafication	**0**
predictive analytics	**2**	graph mining	**0**
sense disambiguation	**2**	linguistic identification	**0**
content classification	**2**	modality detection	**0**
textual entailment	**1**	multimedia processing	**0**
datafication	**0**	negation detection	**0**
concept extraction	**0**	opinion mining	**0**
modality detection	**0**	semantic mining	**0**
negation detection	**0**	sense disambiguation	**0**
competitive intelligence	**0**	text analytics	**0**
linguistic identification	**0**	**text mining**	**0**
abbreviation detection	**0**	text understanding	**0**
text and data mining	**0**	textual entailment	**0**
TDM	**0**	trend detection	**0**