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 D5.4

Roadmap for increasing uptake of TDM

© 2017 FutureTDM | Horizon 2020 | GARRI-3-2014 | 665940

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, 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)

© 2017 FutureTDM | Horizon 2020 | GARRI-3-2014 | 665940

Deliverable

Number:	**D5.4**
Title:	**Roadmap for increasing uptake of TDM**
Lead beneficiary:	**BL**
Work package:	WP5: ELABORATE: Legal framework, policy priorities, roadmaps and practitioner guidelines
Dissemination level:	Public (PU)
Nature:	Report (RE)
Due date:	30.06.2017
Submission date:	30.06.2017
Authors:	**Kiera McNeice**, BL
Contributors:	**Ben White**, BL
Review:	**Kanella Pouli**, ARC

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

Disclaimer: The content of this publication is the sole responsibility of the authors, and does not in any way represent the view of the European Commission or its services.
This report by FutureTDM Consortium members can be reused under the CC-BY 4.0 licence (https://creativecommons.org/licenses/by/4.0/).

Table of Contents

© 2017 FutureTDM | Horizon 2020 | GARRI-3-2014 | 665940

1. INTRODUCTION

The incentives for supporting text and data mining technologies (TDM) in Europe could not be clearer. The estimated value of TDM in the European Big Data Market is expected to grow to USD 10.3 billion in 2021, while the knock-on impact of TDM on the wider European economy has the potential to rise to USD 110.1 billion in 2020.[1] It is therefore crucial that the European Commission take steps to support this rapidly growing area of technology, to maximise the benefits for Europe.

In fact, many industries and researchers are finding that big data analytics are not just an opportunity, but a necessity to deal with the sheer amount of content they work with. There is a growing realisation that data science is the new IT, and becoming crucial to all areas of the economy.

The FutureTDM project was created with the aim of understanding and reducing barriers to the uptake of TDM. Over the past 22 months, the project has engaged with stakeholders from all across Europe[2] in order to develop a comprehensive understanding of the current TDM landscape, the barriers that exist, and a policy framework through which to address those barriers.[3]

This roadmap brings together the outcomes of all the research activities of the FutureTDM project to present an EU-level path to stimulating TDM opportunities in Europe.

[1] FutureTDM D5.2: Trend analysis, future applications and economics of TDM (PDF)
[2] At FutureTDM Knowledge Cafes and other community events
[3] FutureTDM D5.1: Policy Framework (PDF)

© 2017 FutureTDM | Horizon 2020 | GARRI-3-2014 | 665940

2. ROADMAP STRUCTURE

This roadmap outlines a plan to increase the uptake of TDM technologies in Europe, focussing on three conceptual phases:

Phase I: Content Availability

Without data, there can be no data analytics. The first step towards increasing the use of TDM is to ensure that there is as much data available to practitioners as possible.

Phase II: Support Early Adopters

Many people already have a professional and / or personal interest in exploring the use of data analytics technologies and tools; the next phase of this roadmap is to ensure that existing and future TDM practitioners have as much support as possible to overcome the barriers they face.

Phase III: The Next Generation

Once large scale datasets are more freely available, and support for practitioners is established, the final phase of this roadmap focusses on providing as much of society as possible with the fundamental awareness and skills to pursue TDM technologies, if they so choose. The 'next generation' in this context refers not only to younger generations, but to anyone who might become a practitioner or supporter of TDM through acquisition of new skills and knowledge.

Each phase of the roadmap is divided into five sections:

- A single statement of the overall **Objective** of this phase;
- An outline of the current **Situation** regarding TDM, in the context of this phase;
- A breakdown of the main **Challenges** to achieving this phase's objective;
- Key **Principles** to address each challenge highlighted;
- **Activities** that the Commission and other parties should engage in to implement those principles and meet the phase's objective.

The challenges, principles and suggested activities outlined are all based on the extensive research and stakeholder engagement carried out over the course of the FutureTDM project – particularly the FutureTDM Policy Framework, which gives a more comprehensive overview of the ways in which all stakeholders in the TDM landscape can help to support the uptake of TDM technologies in Europe.

2.1 Phase I: Content Availability

Objective

Ensure more large datasets are genuinely available to as many practitioners of TDM as possible.

Situation

Without large datasets that are genuinely available for TDM – that is, legally and practically discoverable and re-usable – there can be no text or data mining. While large industries may have

© 2017 FutureTDM | Horizon 2020 | GARRI-3-2014 | 665940

significant stores of private and proprietary content to work with, academic and smaller commercial practitioners can be severely limited by the lack of large, open datasets.

The first step to increasing the uptake of TDM in Europe must be to increase the amount of data that is open for all TDM practitioners to use.

Challenges

Legal:
When compared to countries like the USA, whose 'Fair Use' exception to copyright extends to the processing and re-use of data for 'transformative' TDM activities, the lawfulness of TDM is often unclear in Europe.[4] Where copyright exceptions exist they are limited to specific member states, and the precise scope of beneficiaries is unclear.

Without legal clarity on what can or cannot be done with third parties' datasets without obtaining explicit permissions from rights holders, such datasets are in effect unusable to many TDM practitioners.

Obtaining explicit permissions from rights holders is in many cases prohibitively resource-intensive for start-ups, SMEs, and academic researchers, and in other cases simply impossible for anyone to achieve in practice. For example, for a company who wishes to scrape and analyse content from the open web, finding, contacting, and obtaining permissions from the rights holder of every website scraped would simply not be feasible.

Data Protection laws also vary across member states, with different national interpretations of EU directives, making it difficult for TDM practitioners to know how to comply.

Knowledge/Awareness:
Although the Open Data and Open Knowledge movement is a growing influence, particularly among publicly-funded researchers, much of the focus of sharing research data is on supplying data to human readers, rather than making it available for machine processing. A dataset that may be 'open' for a human to access and use is useless to a computer algorithm if it does not have machine-readable metadata including, among other things, clear licensing information.[5]

Technical:
While academic institutions are increasingly offering repositories for researchers to store and share data, at a technical level these are extremely diverse, often lacking machine-readable metadata and licensing information. Therefore, although more and more research data is being nominally made available under open licences, these data may be effectively isolated in un-interoperable 'silos' which are difficult or impossible to discover and integrate with other data to form large datasets for TDM. [6]

[4] See FutureTDM D3.3+: Baseline Report of Policies and Barriers of TDM in Europe (PDF), Part II
[5] See FutureTDM Guidelines for Data Management.
[6] A February 2017 LIBER webinar reported that in an assessment of open repositories in the context of FAIR Data Principles, 76% of data were accessible, 41% were findable, 38% were interoperable and just 18% reusable.
© 2017 FutureTDM | Horizon 2020 | GARRI-3-2014 | 665940

Heterogeneity of data sources is not an insurmountable problem, but is made much more difficult by poor quality annotations, metadata, and architecture, all of which hinder TDM practitioners from combining datasets from multiple sources.

Datasets may also exist in formats that are difficult for machines to read and re-use, such as PDFs or images, or use varying technical standards and protocols, requiring significant manual effort to homogenise and make available for large-scale analysis.

Principles

Legal:
To maximise the re-usability of datasets, the findings of the FutureTDM project make it clear that Europe needs to broaden, harmonise, and clarify exceptions to copyright and database laws, as well as harmonising and clarifying the balance between privacy and big data in data protection regimes.

Knowledge/Awareness:
Initiatives aimed at encouraging open data sharing must highlight the needs of machine reading as well as potential human readers, in particular the importance of machine-readable licensing metadata.

Technical:
Europe must continue to support the creation and use of data and metadata standards, centralised access to data, and other ways of connecting data from multiple sources.

Activities

1. *Introduce harmonised, mandatory exceptions to copyright and database rights, for all TDM activities, without limitation to specific sectors or types of TDM practitioners.*

The rights of intellectual property rights holders must be respected, and exceptions should not apply unless TDM practitioners have lawful access to rights holders' content. But in cases where TDM practitioners do not trade on the underlying creative or artistic expressions of the content they analyse and process, it is not reasonable for their activities to be restricted by copyright.

Limiting exceptions to 'non-commercial' purposes or 'research organisations' would in practice introduce significant legal uncertainty, given that university researchers often collaborate with or are partly funded by commercial partners. It would also exclude all commercial players from carrying out *any* TDM activities in cases when it is simply not possible to obtain permissions or licences from all relevant rights holders (e.g. mining the open web). This would have a significant impact on the potential economic benefits of TDM in Europe,[7] as much of the economic value created in Europe in the sphere of data analytics comes from the private sector.

In order to be genuinely effective, exceptions introduced must apply to both copyright and database rights. They must be mandatory across Europe, to minimise the possibility that different implementations and interpretations by member states will continue to fragment the TDM landscape. They must not be overridable by contract, and in order to ensure the integrity of research outcomes,

[7] Several representatives of start-ups and SMEs consulted during the course of the FutureTDM project expressed a concern that without clear legal exceptions allowing TDM for all purposes, many Europeans will choose to start TDM-related companies in countries where clear legal permissions for TDM do exist, such as the USA.

© 2017 FutureTDM | Horizon 2020 | GARRI-3-2014 | 665940

they must allow copies of datasets to be retained for the specific purpose of verifiability and reproducibility of research results.

2. *Clarify any limitations to exceptions to intellectual property rights.*

Rights holders must be allowed to protect the security and technical integrity of their content supply mechanisms through the use of reasonable and proportionate technical protection measures (TPMs). But lawmakers must clarify and provide guidance on what constitutes a 'reasonable and proportionate' technical protection measure, without unduly restricting the activities of machine access to content.

European universities currently spend over ca. one billion euros each year on subscriptions to electronic journal content, much of which funding comes from public sources. The needs of content owners to protect their systems must be balanced against the needs of subscribers, in practical terms, to be able to use the content they subscribe to for large-scale TDM activities.

Any other limitations to legal exceptions must likewise be explicitly defined to minimise legal uncertainty.

3. *Clarify Data Protection frameworks to aid compliance.*

The General Data Protection Regulation must be supplemented with clear guidance on terms such as 'archiving purposes' and 'statistical and scientific purposes' so that TDM practitioners may understand to what extent Data Protection regulations restrict their use of personal data. As with any other legal frameworks affecting TDM, the GDPR must seek to balance the interests of the various affected parties, bearing in mind the public interest in having a vibrant Big Data environment in Europe.

4. *Offer evaluation and certification of Data Protection practices to reduce uncertainty.*

Many potentially highly valuable applications of TDM, for example in the health sector, will inevitably require the use and processing of personal data. Offering a clear process by which organisations carrying out TDM on personal data can have their Data Protection processes evaluated and approved will help reduce legal uncertainty for those organisations, and help them to comply with Data Protection measures.

5. *Commit to supporting open sharing of data in ways that are genuinely re-usable by machines as well as human readers.*

The European Commission has committed to supporting the European Open Science Cloud, whose first report emphasised the importance of machine-readable and machine-actionable data to enable automation of data processing.[8] The remit and objectives of the Commission's FAIR Data Expert Group likewise aims to evaluate the European Commission template for FAIR Data Management Plans in the context of making DMPs more machine-actionable.[9] Machine actionability must be a core consideration for any initiatives which encourage open sharing of content.

[8] Realising the European Open Science Cloud, retrieved 30 June 2017
[9] FAIR data Expert Group Call for Contributions, retrieved 30 June 2017

© 2017 FutureTDM | Horizon 2020 | GARRI-3-2014 | 665940

6. *Ensure funders and researchers make publicly-funded content genuinely open and re-usable for machines as well as human readers.*

Many public funding bodies for research across the EU are adopting policies that mandate researchers make their research data open; such mandates should specify the ways in which data can be made re-usable for machines as well as humans. For example, mandates should specify the use of standard open licences wherever possible;[10] oblige researchers and publishers receiving public money to include licensing information in machine-readable metadata; encourage content creators to use open, machine-readable standards for their data; and supply guidance for researchers on how to do this.

7. *Support the development of centralised, integrated platforms providing access to data sources.*

The best way to ensure datasets are discoverable, accessible and interoperable is to aggregate as many as possible in centralised, standardised, and integrated content repositories, which make clear the rights associated with the materials ingested. Such repositories[11] should be supported at national and international levels, use open standards for content and metadata wherever possible, and ideally provide access to data via open, user-friendly APIs.

Such repositories also provide a place for researchers and other content creators to share their content and data in cases where they may not have a dedicated institutional repository.

8. *Encourage all content repositories to use open metadata standards and expose their metadata.*

In some cases, organisations may still prefer to host their own content repositories rather than depositing content and data into a central repository. In such cases, providing access to open, machine-readable metadata that conforms to consistent standards at least ensures that the content of those repositories is still discoverable by machines and automated processes, and therefore visible to the TDM world.

2.2 Phase II: Support Early Adopters

Objective

Help and support those who are already interested in using and developing TDM tools and applications.

Situation

Many entrepreneurs, researchers, and indeed large companies are already pursuing TDM technologies of their own accord. The work of these practitioners is already beginning to demonstrate the value that TDM can create, which will be a key step in encouraging greater uptake of TDM technologies.

Assuming they have access to adequate datasets to work with, the Commission must do what it can to connect and support these early adopters of TDM, and reduce the barriers they face.

[10] Note that in the case of Creative Commons licences, version 4.0 should be preferred as earlier versions do not address database rights.
[11] The European Open Science Cloud and related GO FAIR initiative are leading examples of the move to develop better integrated infrastructure across Europe.

© 2017 FutureTDM | Horizon 2020 | GARRI-3-2014 | 665940

Challenges

Community:
At the moment the uptake of TDM is fragmented across different sectors, businesses and fields, often limited to areas where early adopters have pushed for the use of TDM in their immediate personal areas of influence. Best practices are not shared across domains or sectors, hindering the advancement of TDM use.

Many others who are interested in adopting or using TDM technologies are unsure whom they can consult for advice or support. Without clear centralised sources of information, documentation, and expertise, the spread of awareness and use of TDM will remain limited to ad hoc personal networks.

Economic Support:
Although funding exists for research into development of novel TDM techniques, many researchers report difficulty in accessing funding for the 'less exciting' aspects of the TDM value chain: Scaling of processes, storage and preservation of datasets, and security of access to data.

Skills:
Many TDM practitioners have little to no understanding of relevant legal and licensing issues, or how to ensure that their activities are lawful with respect to intellectual property and Data Protection.[12] Skills gaps also exist between experts in TDM technology, and experts in subject domains to which TDM may be applied; and between TDM skills used and taught in academia, and those required in an industrial setting.

Principles

Community:
Support and encourage the creation of multi-stakeholder communication channels and platforms, and centralised sources of information about TDM.

Economic Support:
Funders must provide for all stages of the TDM value chain, including activities which are not strictly novel research.

Skills:
Encourage TDM practitioners to 'satellite skills' and awareness of issues such as relevant laws and regulations; facilitate collaborations to bridge skills gaps between TDM and subject experts, and between academia and industry.

Activities

1. *Establish and support working groups, expert groups, and communication channels for TDM practitioners.*

Communication channels connecting TDM practitioners should be highly visible and easily discoverable to anyone interested in using or applying TDM technologies. This will help foster exchanges of ideas

[12] This also has implications for the availability of data; if researchers are not educated in the fundamental legal and licensing issues that relate to data, they are less likely to realise the importance of applying open, standard licences when sharing their own content.

© 2017 FutureTDM | Horizon 2020 | GARRI-3-2014 | 665940

and best practices around existing tools and their applications, bridging gaps between economic sectors and subject domains, and accelerating the benefits that can be derived from the use of TDM. It will also make it easier for new and potential TDM practitioners to discover support networks and communities to whom they can turn for support, reducing barriers to first-time TDM practitioners.

Existing and future projects around the use of TDM should be encouraged to participate in networks like Joinup[13] to share and benefit from best practices.

2. *Increase visibility of TDM resources and experts*

To supplement the support networks discussed above, information about TDM resources and experts should be as easy to discover as possible. By ensuring everyone who is interested in TDM has access to relevant resources, training and experts, this again reduces barriers to adoption and use of TDM technologies.

In the best case this would be achieved through centralised national and international knowledge and resource portals. The FutureTDM and OpenMinTeD projects have made a start on aggregating knowledge and resources, but there is still much more that can be done to aggregate and disseminate resources and support for TDM practitioners.

3. *Support public-private collaborations around TDM technologies.*

Collaborations between academia and industry will help to align the skills that TDM practitioners learn and use in an academic research setting, and those most relevant to commercial and industrial applications. These opportunities for knowledge exchange will in turn make it easier to foster further public-private collaborations, accelerating the impact of TDM education and research.

4. *Encourage research libraries to facilitate TDM practitioners' access to satellite skills.*

Libraries often already undertake many research supporting activities, for example around data management, content sharing, and copyright compliance. These activities already involve connecting researchers to technical and legal experts where necessary. Libraries are therefore ideally placed to expand this role to also support TDM research activities.

5. *Encourage universities to develop strategic policies to support TDM.*

Many of those involved in the research and application of TDM technologies are members of research universities. As multi-disciplinary institutions, universities are ideally placed to mediate sharing of knowledge and best practices across different subject domains. Universities should therefore be encouraged to actively develop strategic approaches to supporting TDM, in coordination with their libraries, by connecting existing and future TDM practitioners with each other and with relevant resources and experts in the university community.

6. *Broaden funders' remit to support TDM infrastructure and other related needs.*

Supporting novel research into TDM applications and tools is of course important, but proving that research outcomes can be translated into genuine, real-world applications requires funding for a host

[13] https://joinup.ec.europa.eu/

© 2017 FutureTDM | Horizon 2020 | GARRI-3-2014 | 665940

of related costs. Among other things these include integrating datasets from different sources; cleaning and homogenising datasets for analysis; building infrastructure to store, preserve and analyse data at scale; and developing and evaluating procedures to ensure compliance with Data Protection and other regulations.

Although these are not technically novel research activities, they must be recognised as crucial parts of research and development of TDM technologies, and provided for in funding for TDM research.

2.3 Phase III: The Next Generation

Objective: Foster a 'data-savvy' society through awareness and education in fundamental skills, to broaden the pool of future TDM practitioners and promoters.

Situation

Although interest in TDM and data analytics is growing, there are still a great many people who have little concept of what TDM technologies are, let alone how they might benefit a given sector. The general attitude among stakeholders consulted in the FutureTDM project is that over the coming years, data science will become 'the new IT', and that it will become critical for all members of society to have some awareness of the uses and impact of data.

By supporting awareness and education in data literacy and related skills, the Commission will pave the way to a 'data-savvy' society that will foster future waves of TDM practitioners.

Challenges

Awareness:
In many areas, there is still a lack of awareness of what TDM and data analytics actually are, how they can be used, and what sorts of benefits can be derived from these technologies. Obviously if people are unaware of the potential applications of TDM, they will not think to pursue the use of TDM technologies.

Education:
Understanding where TDM technologies may be useful requires a basic level of data literacy – that is, an understanding of what data is, why it is valuable, and how it can be used. At the moment, there is little attention paid to data literacy in education until and unless students choose to pursue computer science or related fields in secondary education or beyond.

This means that students who choose not to explicitly study computer science or related fields are unlikely to be taught the data literacy skills that are becoming more and more fundamental to all sectors of modern society. This in turn puts those from non-technical fields at a disadvantage when it comes to understanding the potential value and applications of TDM technologies.

Economic Incentives:
There are significant economic costs associated with introducing TDM into business and education alike; this is likely to disrupt existing practices and workflows, and require new training and acquisition of staff. Without clear evidence of the economic benefits of using TDM technologies, there is little incentive for either the education system or industry to integrate TDM into their curricula or decision-making processes.

© 2017 FutureTDM | Horizon 2020 | GARRI-3-2014 | 665940

Principles

Awareness:
Work still needs to be done to raise awareness of the existence and potential applications of TDM technologies.

Education:
Introduce basic data literacy into educational curricula as early as possible, to foster a 'data-savvy' society in which citizens in all sectors understand the potential value and uses of data.

Economic Incentives:
Clear examples of the benefits of TDM are needed to incentivise people to invest in training and use of TDM technologies.

Activities

1. *Identify and promote 'unusual' applications of TDM as widely as possible.*

To increase awareness of the vast breadth of potential TDM applications, and combat perceptions that TDM is valuable only to technical and quantitative fields, the Commission and projects like FutureTDM should promote examples of TDM being applied in less common fields and sectors. These will help people across all economic sectors to gain a better understanding of what TDM technologies are, the various ways in which they can be used, and ultimately to recognise ways in which TDM may be applied in their own areas.

2. *Integrate fundamentals of data literacy and 'computational thinking' into primary educational curricula.*

FutureTDM's report on the future applications and economics of TDM clearly demonstrates the future economic value of TDM technologies.[14] Investing in education to foster a 'data-savvy' society will in turn help nurture the skills that underpin the use of big data, data analytics, and TDM, and ensure that Europe can maximise the potential economic benefits of these areas.

3. *Quantify and demonstrate specific examples of the value of TDM.*

While FutureTDM's report on the economics of TDM demonstrates the value of investing in TDM to society as a whole, specific examples are needed to demonstrate how businesses can benefit from integrating TDM into their workflows and decision-making. Case studies that show concrete savings in time, money, or other resources should be identified and disseminated to encourage industries to adopt and invest in TDM technologies.

[14] FutureTDM D5.2: Trend analysis, future applications and economics of TDM (PDF)

© 2017 FutureTDM | Horizon 2020 | GARRI-3-2014 | 665940

3. CONCLUSION

The suggested actions in this roadmap culminate from the entirety of the work done by the FutureTDM project from September 2015 to June 2017, and represent the project's views on the best path the European Commission can take towards supporting greater uptake of TDM technologies in Europe. These conclusions are supported by the project's extensive expert reports, which can be found in full on the FutureTDM website.[15]

[15] FutureTDM Knowledge Library

© 2017 FutureTDM | Horizon 2020 | GARRI-3-2014 | 665940