

The background of the cover is a microscopic image of neurons, showing their cell bodies and long, branching processes. A semi-transparent olive-green rectangle is centered on the page, containing the title and publisher information.

WHITE PAPER

OPEN SCIENCE IN A DIGITAL REPUBLIC - STRATEGIC GUIDE

CNRS

SCIENTIFIC AND TECHNICAL INFORMATION
DEPARTMENT

White Paper — Open Science in a Digital Republic — Strategic Guide

Study Review and Proposals for Implementing the Act

Scientific and Technical Information Department - CNRS

DOI: 10.4000/books.oep.1735
Publisher: OpenEdition Press
Place of publication: Marseille
Year of publication: 2017
Published on OpenEdition Books: 31 January 2017
Serie: Laboratoire d'idées
Electronic ISBN: 9782821878433



<http://books.openedition.org>

Electronic reference

SCIENTIFIC AND TECHNICAL INFORMATION DEPARTMENT - CNRS. *White Paper — Open Science in a Digital Republic — Strategic Guide: Study Review and Proposals for Implementing the Act*. New edition [online]. Marseille: OpenEdition Press, 2017 (generated 26 mars 2020). Available on the Internet: <<http://books.openedition.org/oep/1735>>. ISBN: 9782821878433. DOI: <https://doi.org/10.4000/books.oep.1735>.

This text was automatically generated on 26 March 2020.

© OpenEdition Press, 2017
Creative Commons - Attribution-NonCommercial-NoDerivs 3.0 Unported - CC BY-NC-ND 3.0

These Study Review and Proposals have been designed to structure and facilitate the implementation of the Act: they therefore concur with the positions adopted by the sponsors of the White Paper *Open Science in a Digital Republic*, who all clearly wish to associate this new vision for research with practices that will enable it to become a reality.

As a supplementary report to the White Paper, these Guidelines aim to assist players in what are often complex procedures: researchers, technicians, publishers and users of the results of public science now need to move forward in the same spirit of discovery that characterises the major digital projects for science. First among these is the ISTEX Investments for the Future project, which is the driving force.

TABLE OF CONTENTS

Preface

Renaud Fabre

Introduction

Sharing and free analysis of scientific texts and data

Strategic and operational guidelines ...

... following Articles 30 and 38 of the Digital Republic Act (the part known as the so-called *Petite Loi*, or minor bill, resulting from the first reading in the Senate) ...

... following the White Paper Open Science in a Digital Republic ...

... proposing an analysis of the implications of free access to scientific publications ...

... proposing an analysis of the notion of text and data mining and of the corresponding value chain

...

... offering a benchmark of the effects of TDM in countries that have legalised the practice ...

... formulating proposals to provide a framework for implementing the legal provisions relating to Open Science

Open access to scientific publications

Introduction

Open Access: A response to the risk of “misappropriation”

The consecration of a legal right to open access

Article 30 of the Digital Republic Act: Open access

Making scientific texts freely available

Contractual clauses granting exclusive transfer of copyright to have no effect

The recommendations of the European Commission on embargo periods

Prohibiting the privatisation of research data

Freedom to analyse scientific results

Introduction

What is text and data mining?

TDM and the law

TDM techniques

The economics of TDM

What issues does this raise for the work of science?

Analysing scientific publications and data

Setting out the issues and designing research projects

Optimising the governance of scientific systems

Exploiting scientific data

Support for public decision-making

How to organise TDM?

TDM structures and research centres in other countries
 The framework proposed for European projects under H2020
 Data repositories in France: A few examples

Proposals for applying the act
*Introduction**Defining standards*

A reference framework for interoperability specific to Open Science
 Certification or accreditation procedure

Creating a network of digital data curation entities

Conceptualising a typical digital data curation entity
 Forming a network of digital data curation entities

*An ethical framework for TDM via an “Ethics Charter”**Training researchers and research personnel on TDM practice*

Training for the different specialities
 The emergence of new trades and qualifications
 The initial training of researchers
 Occupational training and awareness-raising activity

*Creation of a national agency for Open Science**Summary diagram of the overall framework**Roadmap for the implementing decree for Article 38*

Statement of the principles of Open Science
 The creation of “networked science”
 A model contract for the transfer of copyright between authors and publishers
 Creating an interoperability framework and standards
 An Ethics Charter for digital science
 Creation of a national agency for Open Science
 Creation of a European agency for Open Science

Appendix
*Analytical table comparing different TDM legislation**References**Acknowledgements*

Preface

Renaud Fabre

- 1 The French Parliament has now passed the Digital Republic Act, leading to the next, important, stage of implementation. The strong consensus that has prevailed in the world of public research until now will facilitate the drafting of implementing decrees and instructions for applying the Act. All the communities of researchers, users and beneficiaries of digital science systems (CNRS Scientific Board, ADBU, the Couperin Consortium, EPRISt, CPU, CNum, etc.) came together to support the legislative provisions that have just been successfully concluded.
- 2 These Study Review and Proposals have been designed to structure and facilitate the implementation of the Act: they therefore concur with the positions adopted by the sponsors of the White Paper *Open Science in a Digital Republic*, who all clearly wish to associate this new vision for research with practices that will enable it to become a reality.
- 3 As a supplementary report to the White Paper, these Guidelines aim to assist players in what are often complex procedures: researchers, technicians, publishers and users of the results of public science now need to move forward in the same spirit of discovery that characterises the major digital projects for science. First among these is the ISTEX Investments for the Future project, which is the driving force.
- 4 These Study Review and Proposals contain comparisons with practices in other countries, and proposals or reflections that may be useful in applying the Act. The Prime Minister has stated that the Act should come into force by the end of January 2017. The short time frame is the primary reason why the CNRS's Scientific and Technical Information Department (DIST) wished to produce this document immediately after the conclusive vote by Parliament, when, on Wednesday 28 September, the Senate adopted the Bill with a strong majority, and in terms very close to those proposed by the National Assembly.
- 5 The story of the Act's adoption by Parliament shows how a wide consensus was created, from among both the Majority and the Opposition, and within the Government, where the ministers in charge of research and the digital economy joined forces at an early stage to support text and data mining (TDM) and the free circulation of scientific

publications for research purposes: the Government and Parliament thus showed their agreement with the expectations voiced during the national consultation launched by the Prime Minister in September 2015, and subsequently underlined by the presidents of universities and research organisations.

- 6 Europe has been closely following the debate in France and has chosen to move towards greater freedom of circulation and more intensive exploration of scientific results. Now that the Act is soon to come into force in France, Europe's position is both a benchmark and a new challenge: the competitiveness of French science depends very much on this new world of Open Science initiated by Carlos Moedas, the European Commissioner for Research, Science and Innovation.
- 7 For all of these reasons, the CNRS feels that there are four important lessons to be drawn from these Study Review and Proposals:
 - **Act in a comprehensive manner:** Legal, scientific and technical considerations interact in many ways and must all be taken into account, often in a manner specific to each ecosystem in the world of research. Precise systemic analyses will help avoid excessively vertical or formal approaches, taking advantage of examples from other countries as a matter of course.
 - **Act like a European player:** France now has a dynamic relationship on these issues with the European Commission, where French experience is acknowledged. We need to closely follow the ongoing projects to review European directives; these will provide road maps for the key digital infrastructures that will stimulate the competitiveness of research in the EU.
 - **Make sure that the different provisions of the Act can be used jointly in an optimal way:** The free provision of public data, the free sharing of scientific publications and text and data mining are three inseparable approaches that should be clearly visible to all researchers, users and beneficiaries of public research.
 - **Launch an experimental phase:** There is no denying the importance and the scale of the changes emerging in the digital organisation of scientific work, just as we must not underestimate the complexity of the interactions that are going to occur in the digital system for scientific and technical information (STI). It seems necessary to experiment, assess the new formulae and not declare them to be set in stone from the outset. This task will no doubt require writing progress reports at each step.
- 8 The new science of digital knowledge engineering continues to undergo huge changes. An effort to catch up is under way: it will lead to a new level of competitiveness for research, a step that has already been anticipated by the major universities that are pioneers in the field, grouped together in the League of European Research Universities (LERU).
- 9 These are the key ideas behind these Study Review and Proposals, which are designed to provide assistance with the implementation of this founding legislation for the work of science and the sharing of STI.

The signatories of the White Paper *Open Science in a Digital Republic* (reminder):

The members of the Executive Committee of the ISTEEX Investments for the Future project:

Grégory Colcanap, Coordinator of the Couperin Consortium

Renaud Fabre, Director of the CNRS DIST

Jérôme Kalfon, Director of the ABES

Jean-Marie Pierrel, Professor at Lorraine University

Laurent Schmitt, Head of the Projects and Innovation Department, INIST CNRS

The key witnesses:

Alain Beretz, President of the University of Strasbourg

Jean Chambaz, President of the UPMC

Bruno Chaudret, President of the Scientific Board of the CNRS

Bruno David, President of the French Natural History Museum

Daniel Egret, Astronomer (Paris Science et Lettres), Former President of the Observatory of Paris

Claude Kirchner, Adviser to the President of INRIA, Senior Researcher

Benoît Thieulin, President of the French Digital Council

AUTHOR

RENAUD FABRE

Director of the CNRS DIST

Introduction

Sharing and free analysis of scientific texts and data

Strategic and operational guidelines ...

- 1 The objective of these Study Review and Proposals for implementation is to present all the scientific communities, parliamentarians, scientific publishers and the public in general with practical ways of implementing the new legal provisions introduced by the Digital Republic Act in the field of digital practices for science.

... following Articles 30 and 38 of the Digital Republic Act (the part known as the so-called *Petite Loi*, or minor bill, resulting from the first reading in the Senate¹) ...

- 2 These Guidelines report the comments on and analysis of Articles 30 and 38 of the Digital Republic Act, which introduced into French law the legal basis for Open Science by creating:
 - the right for scientific publications to be made available after an embargo period (Article 30);
 - the right to explore and mine data for the purposes of public research via an exception to copyright and to the right of database producers (Article 38).

... following the White Paper Open Science in a Digital Republic ...

- 3 These Guidelines are a supplement to the White Paper *Open Science in a Digital Republic* published in March 2016 by the CNRS on behalf of the ISTEX project, which served as a guiding thread for preparatory debates leading to the passing of the Act.

- 4 This White Paper gave an account of the practices and needs of researchers with regard to the use of scientific and technical information and digital tools. It also presented a comparative analysis of the texts relating to text and data mining (TDM) in other countries. These elements demonstrated the need for public research to introduce new rights into our legal framework.
- 5 It was the result of:
- a collective undertaking as part of ISTEX (ISTEX, the Excellence Initiative of Scientific and Technical Information, is a project for a digital multi-use platform to the highest international standards, accessible remotely by every scientific community and offering “all the currently available means of consultation and analysis in all scientific communities”²);
 - powerful testimonials from leading figures in the world of research: universities, the League of European Research Universities (LERU), the CNRS Scientific Board, the Bibliographic Agency for Higher Education (ABES), the Couperin Consortium and the University of Lorraine representing the Conference of University Presidents (CPU) as members of the Executive Committee of ISTEX, the Ethics Committee of the CNRS, and the National Digital Council;
 - legal expertise provided by the Cabinet Alain Bensoussan.
- 6 The White Paper called for the following Guidelines:³

Main orientations:

- **Create:** Create a right to Open Science guaranteeing free access and free reuse of data from public research.
- **Balance:** Redefine the economic balance of the digital science ecosystem.
- **Secure:** Adopt Article 18b (new) of the Bill for a Digital Republic as suggested by the Joint Commission [new Article 38 of the *Petite Loi*] creating an exception to copyright and the right of database producers in favour of text and data mining for the results of public research (research articles and data), in order to secure automated data-processing practices and reduce the risk of misappropriation.
- **Compete:** Enable French public research to acquire legal and technical resources that are at least equivalent to those of its European and American counterparts, and in line with the international Open Science movement.
- **Protect:** Protect legitimate interests: exploitation, secrecy, patents, copyright, privacy and personal data.

- 7 The Digital Republic Act has transposed into law the majority of the proposals contained in the White Paper, which is welcomed by all the signatories of the White Paper and of these Guidelines.
- 8 To assist with the drafting of the implementing decrees provided for, in particular in Article 38 of the Digital Republic Act, these Guidelines propose a discussion of the notion of text and data mining and the issues it raises, as well as a comparative analysis of existing structures in France and abroad.

... proposing an analysis of the implications of free access to scientific publications ...

- 9 Although Article 30 of the Digital Republic Act establishing the principle of free access to scientific publications makes no reference to an implementing decree, there is no reason why details should not be added and a framework for implementation and ethical values affirmed.

... proposing an analysis of the notion of text and data mining and of the corresponding value chain ...

- 10 The concept of text and data mining itself covers a range of different real-life applications depending on the chosen analytical viewpoint, whether legal, technical or economic. While TDM implies a researcher, a research subject and tools for automated analysis, other actors are also involved in the value chain:
- scientific publishers;
 - the authors of scientific publications;
 - researchers, laboratories and research institutes;
 - STI correspondents;
 - the publishers of digital analytical tools;
 - the publishers of submission platforms offering access to scientific data;
 - organisations that host data;
 - start-ups or other companies proposing innovative services.
- 11 All these actors, as well as the new players who will emerge on the margins of TDM, form the complex ecosystem in which scientific data live and breathe. Moreover, the world of digital data analysis is today dominated by large American firms. The development and use of digital analytical tools are opportunities for French public research as it enters the new era of digital science.

... offering a benchmark of the effects of TDM in countries that have legalised the practice ...

- 12 These Guidelines provide a benchmark and a comparative analysis of the technical, legal and economic approaches to TDM in the countries that have already legalised the practice, as well as the issues raised by TDM and the levers for fostering it observed in these countries.
- 13 Our approach was to observe foreign practices in order to propose, in the light of French needs and specificities, the ideal legal and organisational framework for implementation.

... formulating proposals to provide a framework for implementing the legal provisions relating to Open Science

14 These Guidelines formulate proposals to provide a framework for implementing the legal provisions of the Digital Republic Act:

- a legal framework, by defining the boundaries of the notion of text and data mining and its scope of application;
 - a technical framework, by creating interoperable platforms, involving the definition of standards, allowing access to all scientific and technical information, as well as tools designed for text and data mining or which can be used for this purpose. The ISTEX Platform could pioneer this process;
 - a structural framework, by creating a network of certified “digital data curators”, whose mission would be to preserve the files produced on completion of research activities and to organise their availability;
 - an ethical framework, by defining good practices for the use of TDM in scientific research;
 - this legal, ethical, organisational and structural framework could be headed by a National Agency for Open Science responsible for the governance of STI and guaranteeing its efficiency.
-

NOTES

1. Text of the *Petite Loi* for a Digital Republic: <http://www.senat.fr/petite-loi-ameli/2015-2016/744.html>
2. <http://www.istex.fr/>
3. White Paper *Open Science in a Digital Republic*, March 2016, p. 11.

Open access to scientific publications

Introduction

- 1 The White Paper *Open Science in a Digital Republic* reports and defends a twofold observation:
 - a review of the current situation on the uses of French public research suggests a pressing need to catch up, given that today the use of digital practices in science currently lags behind the major emerging and/or established practices in the leading countries of science; as the CNRS strategy “A better sharing of knowledge”¹ has shown;
 - the changes under way must move towards a “right of shared resources and protected uses”, and towards a right to Open Science guaranteeing free access and free reuse of data from publicly funded research.
 - 2 Article 30 of the Digital Republic Act formally recognises this need for researchers to have access to their colleagues’ work and for the creation of a right to access and share knowledge, thus responding to the risk of the misappropriation of knowledge and conforming to the trend followed by our European neighbours.
-

NOTES

1. <http://www.cnrs.fr/dist/strategie-ist.htm>

Open Access: A response to the risk of “misappropriation”

- 1 To be able to carry out their work correctly, researchers need to be able to freely access both the scientific data and the publications of their peers (where this is the result of research published by a private publisher).
- 2 The business models (author-pays or reader-pays) and the legal models (exclusive transfer of rights, subscription contract) of scientific publishing lead to a form of appropriation of scientific knowledge by private publishers. Although some publishers allow authors to deposit their articles in an institutional archive after an embargo period, others retain the entirety of the rights for the full duration of copyright protection (70 years from the death of the author).
- 3 These models for funding publication in the digital age have obliged public research institutions to pay for restricted access to knowledge arising from the research programmes they fund.
- 4 The study conducted by the CNRS's Scientific and Technical Information Department (DIST) entitled *Who should finance scientific publication: The "Reader" and/or the "Author"?* (January 2016) sets out precisely this need to reform the "author-pays" model (payment of article processing charges) and the "reader-pays" model (subscription) by developing a hybrid model that particularly takes into account the financial impacts and the risks of the privatisation of knowledge.
- 5 As this model is no longer viable, especially economically, scientific publication should be opened up by creating a right of access to all publications for public research. In its study, the CNRS specifies that:

"The goal is to achieve comprehensive security in all the parameters of movement towards Open Science. When 'negotiations' are conducted in relative secrecy at national level, with no overall safety net and no announcements (yet) about their content or results, this leaves publishers with the power of decision over digital STI. This situation contains the risk, as observed by the OECD, of an 'every man for himself' situation, with confusion and the fragmentation of international scientific collaborative studies, and with foreign publishers being able to pursue their own interests in the way they share out the results of public research."

The consecration of a legal right to open access

- 1 There is wide agreement on the risk of drifting towards the privatisation of knowledge among all scientific communities and in particular the higher education institutions, for which the total cost of subscriptions to publishers' platforms are increasing exponentially.
- 2 These elements have been welcomed by the sponsors of the Digital Republic Act, who have introduced the principle of a French version of open access to scientific publications.

Article 30 of the Digital Republic Act: Open access

- 3 The Digital Republic Act consecrates this right of access to scientific publications in the following terms:
 - In Chapter 3 of Title 3 of Book V of the Research Code, an Article L. 533-4 shall be inserted as follows:

“Art. L. 533-4. – I. – When a scientific text arising from a research activity financed at least 50% by grants allocated by the state, by regional or local authorities or public institutions, by grants from national funding agencies or by European Union funds is published in a periodical appearing at least once a year, its author, even after having granted exclusive rights to a publisher, has the right to make available free of charge in an open format, in digital form, subject to the agreement of any co-authors, the final version accepted for publication, as soon as the publisher itself makes the latter available free of charge in digital form, and, failing this, on expiry of a period running from the date of first publication. This period is a maximum of six months for a publication in the field of the sciences, technology and medicine, and twelve months in that of the human and social sciences.

The version made available in application of the first subparagraph may not be exploited in the framework of a commercial publishing activity.

II. – Once the data from a research activity, financed at least 50% by grants allocated by the state, by regional or local authorities or public institutions, by grants from national funding agencies or by European Union funds, are no longer protected by specific rights, or special regulations, and they have been made public

by the researcher, the research establishment or organisation, they can be freely reused.

III. – The publisher of a scientific text mentioned in I shall not limit the reuse of research data made public in the framework of its publication.

IV. – The provisions of this article are public policy and any clause to the contrary is deemed to be unwritten.”

4 The Article organises open access as follows:

- concerning publications:

- Article 30 provides the right for the author of a scientific text to make the final version of the manuscript accepted for publication freely available in an open digital format;
- this version may be made available either immediately, if the publisher places the publication online free of charge, or after an embargo period;
- embargo periods are six months for a publication in the field of the sciences, technology and medicine and twelve months in that of the human and social sciences, in compliance with EU recommendations;
- clauses on exclusive transfer of copyright laid down in publishing contracts shall not hinder the author’s right to make content available;

- on research data:

- research data may be reused freely once the research institution has made them public;
- the publisher may not retain ownership of the research data associated with a publication.

- The provisions of Article 30 are public policy and any clause to the contrary shall be deemed unwritten.

Making scientific texts freely available

5 **A requirement for researchers.** A general consensus emerged from the public consultation on the Digital Republic Act regarding a clear demand to strengthen the rights of researchers to disseminate their work freely, when the work has been financed by public funds.

6 **Legal consecration.** The new legislation has introduced into the French Research Code a right for the author of a scientific text to make **the final version of the manuscript accepted for publication** available free of charge, when this text is the result of a research activity financed at least 50% by public funds.

7 While the intention of the Act to open up access to and facilitate the sharing of scientific publications can only be welcomed, some clarification is needed, particularly concerning the notion of “final version of the manuscript accepted for publication”.

8 **Clarifications.** The problem is that the law (neither in the Intellectual Property Code nor in the Heritage Code) provides no definition of “manuscript”, “author’s version”, “publisher’s version”, “pre-publication”, “post-publication”, etc. These terms taken from publishing practice must be defined, legally qualified and given legal recognition (ownership of rights and the corresponding rights of exploitation).

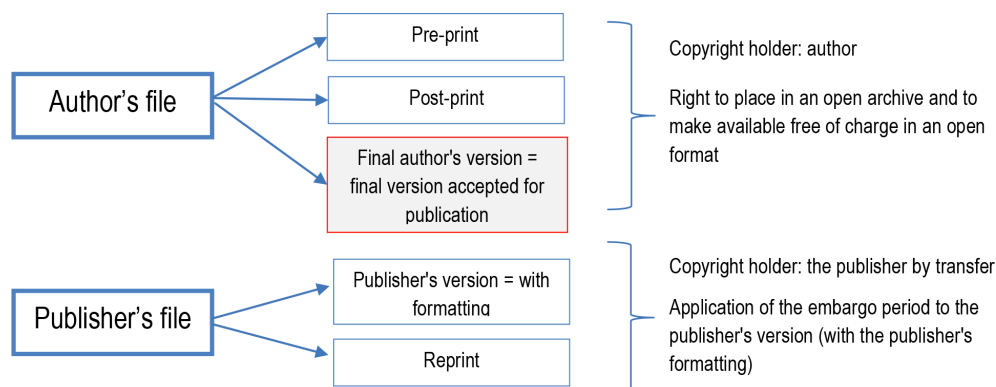
9 The scientific community and in particular the CNRS’s STI correspondents (CORIST) have considered the definition of the terms “manuscript” and “final version” in the light of current practice. When he was interviewed in the framework of the White Paper, Claude Kirchner of INRIA summarised the problem in the following manner:

“Any embargo period could concern only the ‘publisher’s version’ in its final form, in order to retain its commercial potential. Such restrictions are acceptable only if

the ‘author’s version’ can be freely distributed, and the duration of the embargo should then be set in compliance with international practices.”

(INRIA hearing, Claude Kirchner, 15 October 2015)¹

- 10 Implementation decrees would provide an opportunity to propose definitions corresponding to their use in practice and especially in scientific publishing.
- 11 To this end, these Guidelines propose that a **reference base of uses** be created, which could contain a glossary and a definition of the terms used in practice, as well as the regime applicable to each different version of an article. The different versions of an article are referred to as follows:



- 12 The concept of “final version of the manuscript accepted for publication” seems to mean the author’s last version before publication, and therefore before formatting by the publisher. Article 30 of the Digital Republic Act should therefore be clarified by decree to specify which version is covered by the embargo period.

△ Decree: to create a reference base of uses and to specify which version of the manuscript is subject to an embargo.

Contractual clauses granting exclusive transfer of copyright to have no effect

- 13 The text of Article 30 stipulates that the right of researchers to make their scientific publications freely available applies “even after having granted exclusive rights to a publisher”.
- 14 Since publishing contracts between a researcher and a publisher most often take the form of a standard form contract, in the interest of open access the new legislation declares that any clause granting exclusive copyright should be “unwritten”.
- 15 **Proposal: Model contract.** In order to guarantee the rights of researchers regarding their published material and to take into account the risks of contractual asymmetry, a model contract could be promulgated by decree for transferring copyright for use in public research.

- 16 This contract would lay down the rules governing the relationship between the parties and protect researchers in their relationship with publishers. It would in particular ensure that there was no exclusive transfer, and guarantee the rights of researchers to:
- authorise the depositing and the reproduction in an open archive of the publication in the author's version immediately, and in the publisher's version after expiry of an embargo period;
 - allow the immediate exploration of the content of the article using digital data-processing tools;
 - prevent all forms of private retention or reservation of ownership concerning the content of the article and the corresponding data.
- 17 This contract could be promulgated by decree and thus have a regulatory status that would be imposed on the publisher for any scientific publication resulting from public research.

△ Decree: creation of a model contract concerning the transfer of rights intended for scientific publications.

The recommendations of the European Commission on embargo periods

- 18 Seeking a new balance between the positions of the different stakeholders in the digital age and the knowledge society, the Government has included the following in the Act:
- the possibility of dissemination by free access of publicly funded scientific work, upon expiry of what is known as an “embargo period”;
 - “embargo periods” of six and twelve months, at the end of which authors of publications financed by public funds must make their texts freely available. If articles are made available by the online publisher free of charge, authors will be able to exercise their right immediately.
- 19 **EC recommendations.** The embargo periods laid down by the Act are the maximum deadlines provided for by the recommendation of the European Commission (C(2012) 4890):²
- 20 The EC recommends that Member States:
- “Define clear policies for the dissemination of and open access to scientific publications resulting from publicly funded research. These policies should provide for:
 - concrete objectives and indicators to measure progress,
 - implementation plans, including the allocation of responsibilities,
 - associated financial planning”.
 - They should also ensure that there is “open access to publications resulting from publicly funded research as soon as possible, preferably immediately and in any case no later than 6 months after the date of publication, and 12 months for social sciences and humanities”.
- 21 **Measures taken in other countries.** The French embargo periods are the same or similar to those applied under law by our European neighbours:
- in Germany: embargo period of 12 months with no distinction between disciplines;
 - in Spain: depositing in an institutional archive as early as possible, without exceeding 12 months, with no distinction between disciplines.

Prohibiting the privatisation of research data

- 22 **Need for sharing.** The French Research Code defines the following missions of public research (Article L.112-1 of the Research Code), among others:
- “sharing and disseminating scientific knowledge”;
 - “open access to scientific data”.
- 23 All the scientific communities agree that it is necessary to have free and unimpeded access to scientific data, in the greater interest of research, for which the stakes are very high.
- 24 In an article entitled “Préserver les données de la recherche à l’ère du Big Data” (Preserving research data in the age of Big Data),³ the problem of the preservation and the sharing of research data is fully and accurately presented and discussed.
- “As the massive amounts of data produced by research continue to increase exponentially, the issue of data storage has become crucial, both for preserving our scientific heritage and to enable their use by the scientific community. ... As analytical instruments and software improve, almost all disciplines have witnessed an explosion in the amount of data created each year. And these data are precious since they have very often been generated by complex and costly studies, in high energy physics for example, or else they are the result of periodic observations carried out over extremely long periods, such as tracking the position of stellar objects or demographic data.”
- 25 Article 30 of the Digital Republic Act expresses this need for free and open access “to data from a research activity”, and also the need to prevent any privatisation of these data, particularly by publishing contract.
- “II. – Once the data from a research activity, financed at least 50% by grants allocated by the state, by regional or local authorities or public institutions, by grants from national funding agencies or by European Union funds, are no longer protected by specific rights, or special regulations, and they have been made public by the researcher, the research establishment or organisation, they can be freely reused.
- III. – The publisher of a scientific text mentioned in I shall not limit the reuse of research data made public in the framework of its publication.”
- 26 The text establishes the principle of the free reuse of data generated by public research. However, the boundaries of “these research data resulting from a research activity at least 50% of which was financed by public funds” are not specified, and the procedures for sharing and accessing these data are not defined.
- 27 These clarifications, which are necessary for good governance of research data and Open Science, must be included in an implementing decree. The sharing of knowledge is the vital and historical basis of the scientific approach, and indispensable for research. The digital transition has disrupted previous practice by giving access to a growing and comprehensive mass of data, instantaneously and anywhere in the world. Big Data, when applied to science, entails the development of tools and practices involving intelligent exploration by automated data analysis and observation services.
- 28 The use of these text- and data-mining tools and the advent of new transversal and multidisciplinary scientific practices give rise to multiple issues, scientific of course but also human, economic and ethical. The legislation takes these issues into account, as well as the need to introduce this right to text and data mining in French legislation,

thus enabling French research to compete with its British, American or Canadian counterparts. What is now needed is an organisational structure capable of ensuring that these principles are implemented efficiently.

NOTES

1. Claude Kirchner, 15 October 2015, White Paper, p. 226.
2. https://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf
3. Guillaume Garvanèse, “Préserver les données de la recherche à l’ère du Big Data”, *CNRS Le Journal*, 9 September 2016, <https://lejournel.cnrs.fr/articles/preserver-les-donnees-de-la-recherche-a-lere-du-big-data>

Freedom to analyse scientific results

Introduction

- 1 Text and data mining (TDM) is a set of techniques enabling researchers to explore and process vast amounts of data. It opens new fields of research and provides new ways of extending knowledge. Although its potential is still to be fully developed, TDM addresses many scientific and economic challenges. On the one hand, it helps intensify and stimulate research. On the other, it can have considerable economic value, offering savings in time and money for research expenditure. It is also a factor for the improvement of public decision-making.
- 2 The concept of TDM covers a range of real-life situations that should be clarified from a legal, technical and economic point of view. There are considerable issues related to TDM and the levers for fostering its use in scientific research as well as more generally, which are taking science into a new era, involving a necessary framework to cover multiple practices. In order to establish firm foundations for this new situation, we propose analysing comparable practices: the way TDM is organised in the United Kingdom and the United States; the framework suggested in projects funded under H2020, and, finally, the practices that already exist in France.

What is text and data mining?

- 1 Data mining is a recent concept that first appeared in 1989 under the name Knowledge Discovery in Databases (KDD), also known in French as *Extraction de Connaissances à partir des Données* (EDC).
- 2 The term “text and data mining” appeared for the first time in the field of marketing at the beginning of the 1990s. This concept, as applied in marketing, is closely linked to the notion of the “one-to-one relationship” (Michael Berry and Gordon Linoff, creators of Data Mining in Marketing), i.e. the customisation of the relationship between a company and its clients.
- 3 These Guidelines focus on the application of TDM to science, but TDM is practised in many sectors of activity such as, for example¹:
 - direct marketing: in this area, TDM techniques are used, for example, to segment customer databases and predict customers’ purchasing intentions in order to optimise the marketing pitch;
 - communication: anti-spam filtering of emails and the global Echelon system for the interception of private and public communications (SIGINT) developed by the United States, the United Kingdom, Canada, Australia and New Zealand in the framework of the UKUSA Agreement, both use TDM techniques;
 - banking and finance;
 - insurance and health;
 - medical and pharmaceuticals.
- 4 TDM has been adopted in the sciences in the last few years, with the development of open archives such as arXiv or HAL, so as to optimise the exploration of their databases, which are constantly growing in volume.
- 5 The notion of text and data mining applied to the scientific field is today widely used to designate a variety of tools and activities. We therefore propose analysing the notion of TDM from a legal, technical and economic perspective in order to answer the following questions:
 - What is TDM?
 - What are the operations involved in TDM?
 - In what areas can TDM be applied?
 - How can we measure the efficiency of TDM?

TDM and the law

The consecration of a legal right to TDM via an exception

- 6 **Twofold exception.** Article 38 of the Act (*Petite Loi*) establishes a right to text and data mining by introducing an exception to copyright and to the right of database producers, under the following terms:

The Intellectual Property Code is modified as follows:

1° After the second subparagraph of 9° of Article L. 122-5, a 10° shall be inserted as follows:

“10° Digital copies or reproductions made from a lawful source, in view of the exploration of texts and data for public research needs included in or associated with scientific results for the needs of public research, excluding any commercial purpose. A decree lays down the conditions under which the exploration of texts and data is implemented, as well as the terms for storage and communication of the files produced on conclusion of the research activities for which they were produced; these files constitute the research data”;

2° After 4° of Article L. 342-3, a 5° shall be inserted as follows:

“5° Digital copies or reproductions of the database made by a person with lawful access, in view of text and data mining included in or associated with scientific results in a research framework, excluding any commercial purpose. The storage and communication of technical copies resulting from processing, on conclusion of the research activities for which they were produced, are carried out by organisations appointed by decree. Other copies or reproductions are destroyed.”

- 7 **Lack of a definition.** The text does not define the concept of data exploration or mining. When both terms are used in the same text, a comment is called for: the use of the term “exploration” of texts and data in the first part of the text, introducing an exception to copyright, and then of “mining” (*fouille*) in the second part, creating an exception to the right of the database producer, may raise difficulties of interpretation. The implementing decree could start by specifying that the concepts of exploration and mining cover the same practices.

△ The implementing decree could start by specifying that the concepts of exploration and mining cover the same practices. The decree should follow the European Union Directives in containing an article entitled “Definition”.

- 8 **Scope of the notion.** Although the actual concepts of text and data mining and exploration are not defined, the text lays down limits and a framework for this practice:

Criteria	Article 38
Legal basis	An exception to copyright and the right of the database creator: the right to copy and reproduce material digitally for the purpose of TDM
Scope of TDM	Mining of text and data included in or associated with scientific texts
Who benefits from the exception?	/

Limits	TDM limited to the needs of scientific research/a research framework Non-commercial purposes Lawful source/lawful access to texts and data subject to TDM
---------------	---

Introducing a TDM exception in the proposed Directive on Copyright in the Digital Single Market

- 9 **Preliminary reports.** The White Paper *Open Science in a Digital Republic* referred to many reports, some of them commissioned by the European Commission, which advocated the revision of Directive 2001/29/EC “On the harmonisation of certain aspects of copyright and related rights in the information society” and the introduction of a right to TDM:
- the Sirinelli Report for the French Higher Council for Literary and Artistic Property (CSPLA) *Rapport de la mission sur la révision de la directive 2001/29/CE sur l’harmonisation de certains aspects du droit d’auteur et des droits voisins dans la société de l’information* (Report of the mission on the revision of Directive 2001/29/EC on the harmonisation of certain aspects of copyright and related rights in the information society) of December 2014 states that “demands for the creation of new exceptions to copyright relate in particular to the activities known as text and data mining (TDM)”;²
 - the study by Wolf & Partners in March 2014, entitled *Study on the Legal Framework of Text and Data Mining*,³ for the European Commission;
 - a group of experts of the European Commission also published, in April 2014, a report entitled *Standardisation in the area of innovation and technological development, notably in the field of text and data mining*;⁴
 - the Reda Report: this report, adopted by the European Parliament on 9 July 2015, “stresses the need to properly assess the enablement of automated analytical techniques for text and data (e.g. ‘text and data mining’ or ‘content mining’) for research purposes”;
 - the European Commission press release of 9 December 2015 presenting the measures to improve access to online content and the Commission’s vision of an overhauled copyright. In this context, “the Commission intends to work on key EU exceptions to copyright” and, in particular, “will revise EU rules to make it easier for researchers to use ‘text and data mining’ technologies to analyse large sets of data”.
- 10 **Proposed Directive.** The proposed Directive on Copyright in the Digital Single Market (COM(2016) 593 final) was published by the European Commission on 14 September 2016.
- 11 Through this proposed Directive on Copyright in the Digital Single Market (COM(2016) 593 final),⁵ the Commission “proposes modern EU copyright rules for European culture to flourish and circulate ... The proposals will also bring tools for innovation to education, research and cultural heritage institutions”.⁶ The objective of this Directive is to adapt the provisions relating to copyright to the increasing use of digital technologies, in particular in the field of scientific research, noting the multifaceted application of the provisions of the InfoSoc Directive and in particular the exceptions.⁷
- 12 **Defining TDM.** Article 2 of the Directive offers a definition of the notion of text and data mining:

“text and data mining means any automated analytical technique aiming to analyse text and data in digital form in order to generate information such as patterns, trends and correlations”.

- 13 **Exception.** Article 3 introduces an exception to copyright and an exception to the right of the database creator in favour of text and data mining in the following terms:

Article 3 – Text and data mining

1. Member States shall provide for an exception to the rights provided for in Article 2 of Directive 2001/29/EC, Articles 5(a) and 7(1) of Directive 96/9/EC and Article 11(1) of this Directive for reproductions and extractions made by research organisations in order to carry out text and data mining of works or other subject-matter to which they have lawful access for the purposes of scientific research.

2. Any contractual provision contrary to the exception provided for in paragraph 1 shall be unenforceable.

3. Rightholders shall be allowed to apply measures to ensure the security and integrity of the networks and databases where the works or other subject-matter are hosted. Such measures shall not go beyond what is necessary to achieve that objective.

4. Member States shall encourage rightholders and research organisations to define commonly-agreed best practices concerning the application of the measures referred to in paragraph 3.

- 14 This exception can be analysed according to the following criteria (criteria used for an analysis of the French text):

Criteria	Proposed Directive
Legal basis	An exception to copyright and the right of the database creator: the right to make reproductions or extracts for the purpose of TDM
Scope of TDM	TDM on works or other objects
Who benefits from the exception?	Research organisations (the notion is broadly defined in Article 2 of the proposed Directive ⁸)
Limits	TDM limited to the needs of scientific research Non-commercial purposes Legal access to the source material of TDM

- 15 The Commission justifies the general principles behind this text in the following manner:

Text and data mining:

- Option 1 consisted in self-regulation initiatives from the industry.
- Other options consisted in the introduction of a mandatory exception covering text and data mining.
 - In Option 2, the exception covered only uses pursuing a non-commercial scientific research purpose.
 - Option 3 allowed uses for commercial scientific research purpose but limited the benefit of the exception to some beneficiaries.
 - Option 4 went further as it did not restrict beneficiaries.
- Option 3 was deemed to be the most proportionate one.

- 16 The Commission specifies that the purpose of this text is to provide a legal clarification and a framework of fair competition so that European researchers can use innovative techniques for data analysis, thus enabling them to more quickly find innovative solutions in response to major challenges such as global epidemics and climate change, and promoting cross-border and interdisciplinary collaborations. This exception supports European competitiveness by promoting Open Science.⁹
- 17 Carlos Moedas, European Commissioner for Research, Science and Innovation, has justified the need for this exception in the following manner:
- “Science needs a copyright law that reflects the reality of the modern age. We must remove barriers that prevent scientists from digging deeper into the existing knowledge base. This proposed copyright exception will give researchers the freedom to pursue their work without fear of legal repercussions, and so allow our greatest minds to discover new solutions to major societal problems.”
- 18 France and Europe are developing a legislative arsenal authorising the use of automated analysis techniques, but it is also interesting to look at the provisions adopted by other countries.

Shifting limits to the notion of TDM in the legislation of different countries

- 19 The table below presents an analytical reading of the notion of TDM in British, American and Japanese law, which have each introduced a legal right to TDM.

Country	Source	Text	Characteristics of the TDM
---------	--------	------	----------------------------

United Kingdom	<p>Act of Parliament</p> <p>Article 29 A introduced in 2014¹⁰ in the Copyright, Designs and Patents Act (1988)</p>	<p>29A Copies for text and data analysis for non-commercial research</p> <p>(1) The making of a copy of a work by a person who has lawful access to the work does not infringe copyright in the work provided that—</p> <p>(a) the copy is made in order that a person who has lawful access to the work may carry out a computational analysis of anything recorded in the work for the sole purpose of research for a non-commercial purpose, and</p> <p>(b) the copy is accompanied by a sufficient acknowledgement (unless this would be impossible for reasons of practicality or otherwise).</p> <p>(2) Where a copy of a work has been made under this section, copyright in the work is infringed if—</p> <p>(a) the copy is transferred to any other person, except where the transfer is authorised by the copyright owner, or</p> <p>(b) the copy is used for any purpose other than that mentioned in subsection (1)(a), except where the use is authorised by the copyright owner.</p> <p>(3) If a copy made under this section is subsequently dealt with—</p> <p>(a) it is to be treated as an infringing copy for the purposes of that dealing, and</p> <p>(b) if that dealing infringes copyright, it is to be treated as an infringing copy for all subsequent purposes.</p> <p>(4) In subsection (3) “dealt with” means sold or let for hire, or offered or exposed for sale or hire.</p> <p>(5) To the extent that a term of a contract purports to prevent or restrict the making of a copy which, by virtue of this section, would not infringe copyright, that term is unenforceable.</p>	<p><u>Legal basis:</u> Exception to copyright for the purpose of “computational analysis”</p> <p><u>Scope:</u> Works and all related data</p> <p><u>Beneficiary:</u> /</p> <p><u>Limits:</u></p> <p>Lawful access</p> <p>Non-commercial purpose</p> <p>For the sole purpose of research</p> <p>Attribution of authorship</p> <p>No copy may be transferred to any other person/no copy may be transferred or licensed by contract</p>
----------------	---	--	---

United States	<p>Act</p> <p>Federal Agency Data Mining Report Act of 2007: Article 42 U.S. Code § 2000ee-3</p>	<p>The term “data mining” means a program involving pattern-based queries, searches, or other analyses of 1 or more electronic databases, where—</p> <p>A. a department or agency of the Federal Government, or a non-Federal entity acting on behalf of the Federal Government, is conducting the queries, searches, or other analyses to discover or locate a predictive pattern or anomaly indicative of terrorist or criminal activity on the part of any individual or individuals;</p> <p>B. the queries, searches, or other analyses are not subject-based and do not use personal identifiers of a specific individual, or inputs associated with a specific individual or group of individuals, to retrieve information from the database or databases; and</p> <p>C. the purpose of the queries, searches, or other analyses is not solely—</p> <p>i. the detection of fraud, waste, or abuse in a Government agency or program; or</p> <p>ii. the security of a Government computer system.</p> <p>11</p>	<p><u>Legal basis:</u> Positive right to conduct software-based searches</p> <p><u>Scope:</u> Models, searches, or other types of analysis of one or more electronic databases</p> <p><u>Beneficiary:</u> Federal agencies</p> <p><u>Limits:</u> Predictive analysis concerning terrorist or criminal activity</p> <p>No use of personal identifiers</p>
United States	<p>Court ruling</p> <p>Authors Guild v. Hathi Trust, 755 F.3d 87 (2d Cir. 2014).</p>	<p>The court held that the HDL’s first use—creation of a full-text searchable database—was fair. It found that use “quintessentially transformative” because “the result of a word search is different in purpose, character, expression, meaning, and message from the page (and the book) from which it is drawn.” The court further held that the copies were reasonably necessary to facilitate the HDL’s services to the public and to mitigate the risk of disaster or data loss. In addition, it held that the full -text search posed no harm to any existing or potential traditional market for the copyrighted works. The court also held that the second use—access for the print -disabled—was fair. It concluded that providing such access was a valid purpose under the first statutory factor, even though it was not transformative. The court held that it was reasonable for the defendants to retain both text and image copies because the text copies were required for text searching and text -to-speech capabilities, and the image copies provide an additional method by which many disabled patrons can access the works. Finally, the court held that the fourth factor favored fair use given the insignificance of the present-day market for books accessible to the handicapped.¹²</p>	<p><u>Legal basis:</u> The reproduction and use of digital books for their conversation, for text searches or for assisting the visually impaired is not an infringement of copyright, as it corresponds to “fair use”</p> <p><u>Scope:</u> Works/databases</p> <p><u>Beneficiary:</u> Parties to the dispute</p> <p><u>Limits:</u> Four criteria for fair use:</p> <ul style="list-style-type: none"> - non-commercial purpose - the nature of work protected by copyright - the portion of the work used - the use must have no financial effect

United States	<p>Court ruling</p> <p>Authors Guild v. Google, INC.¹³</p> <p>16 October 2015</p>	<p><i>In sum, we conclude that (1) Google's unauthorized digitizing of copyright-protected works, creation of a search functionality, and display of snippets from those works are non-infringing fair uses. The purpose of the copying is highly transformative, the public display of text is limited, and the revelations do not provide a significant market substitute for the protected aspects of the originals. Google's commercial nature and profit motivation do not justify denial of fair use.</i></p> <p><i>(2) Google's provision of digitized copies to the libraries that supplied the books, on the understanding that the libraries will use the copies in a manner consistent with the copyright law, also does not constitute infringement. Nor, on this record, is Google a contributory infringer.</i></p>	<p><u>Legal basis:</u> Google's legal right to make certain passages from books available in digital format on the basis of fair use</p> <p><u>Scope:</u> Works</p> <p><u>Beneficiary:</u> Parties to the dispute</p> <p><u>Limits:</u> Four criteria for fair use:</p> <ul style="list-style-type: none"> - non-commercial purpose - the nature of work protected by copyright - the portion of the work used - the use must have no financial effect <p><u>Scope:</u></p> <p>With this ruling, the United States has given its researchers a significant advantage by granting them the possibility of digitising very large lawfully accessible datasets, of sharing the corpora and developing search functions and data-processing algorithms¹⁴</p>
---------------	---	---	---

Japan	<p>Act</p> <p>Japan Copyright Act – Article 47 septies introduced in 2009</p>	<p><i>For the purpose of information analysis (“information analysis” means to extract information, concerned with languages, sounds, images or other elements constituting such information, from many works or other much information, and to make a comparison, a classification or other statistical analysis of such information; the same shall apply hereinafter in this Article) by using a computer, it shall be permissible to make recording on a memory, or to make adaptation (including a recording of a derivative work created by such adaptation), of a work, to the extent deemed necessary. However, an exception is made of database works which are made for the use by a person who makes an information analysis.¹⁵</i></p>	<p><u>Legal basis:</u> The text provides for an “information analysis” exception for purposes of comparison, classification, and statistical analysis</p> <p><u>Scope:</u> Information of any kind</p> <p><u>Beneficiary:</u> /</p> <p><u>Limit:</u> Information analysis is not restricted to public research, or to non-commercial purposes</p>
-------	--	---	---

A comparative analysis of legal provisions

- 20 A comparative analysis of French, British, American and Japanese legal provisions, as well as of those in the proposed Directive, inspires the following remarks in the light of the four analytical criteria used (the legal basis, scope, beneficiaries and limits of TDM):

16

Criteria	Comparative analysis
Legal basis	<p>An exception to copyright and the right of the database creator is favoured in each of the countries that have legislated to authorise TDM.</p> <p>The United Kingdom prefers the expression “computational analysis” to “text and data mining”.</p>
Scope	<p>Different national legal codes offer different degrees of scope for TDM:</p> <ul style="list-style-type: none"> - Japanese law is the broadest, stipulating that TDM can be practised on any kind of information; - English law and the proposed Directive reserve it for works and all associated data; - French law refers to “text and data included in or associated with scientific texts”.
Beneficiary	<p>Neither English nor French law specifies any particular beneficiary of these provisions. However, the beneficiaries concerned are indirectly the public research organisations; TDM is limited to purposes of public research.</p> <p>The proposed Directive expressly limits the use of TDM to research organisations.</p>
Limits	<p>All the legal systems mainly specify three limits. TDM should be:</p> <ul style="list-style-type: none"> - limited to the needs of scientific research; - used for non-commercial purposes; - applied to texts and data with lawful access.

- 21 This comparative analysis shows that there are certain similarities between the different legislative frameworks governing TDM in the countries analysed.
- 22 In order that the French legislation should not fall short of the provisions in other countries and the proposed Directive, we recommend that the implementing decree clarify the concept of “texts and data included in or associated with scientific texts” to make it as broad as possible.

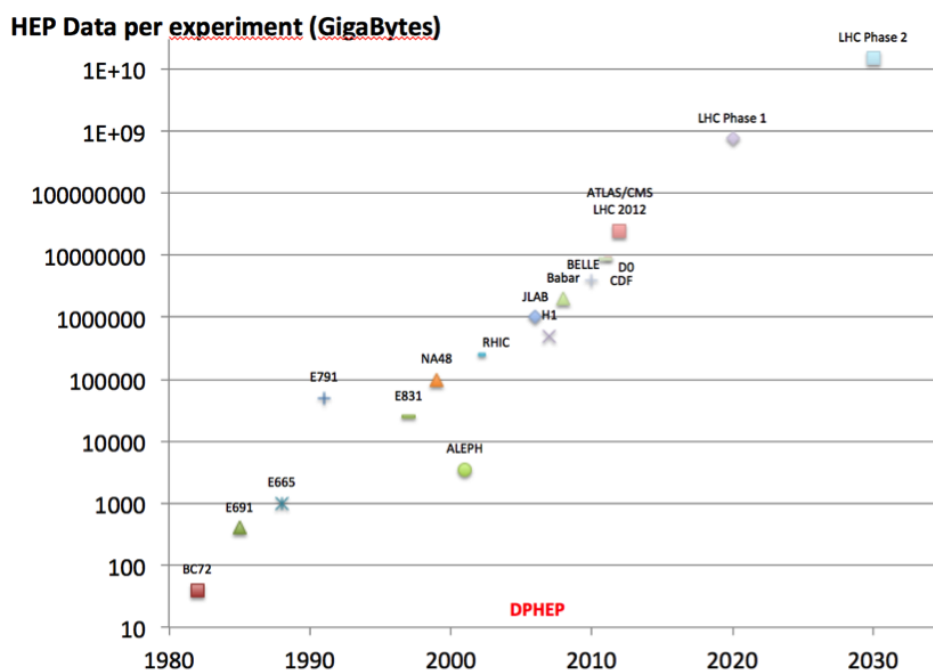
△ Definition of the notion of TDM and the scope of its application. Clarification of the concept of “texts and data included in or associated with scientific texts” to make it as broad as possible.

TDM techniques

- 23 From a technical point of view, TDM can be analysed in relation to:
- the necessary technological objects involved;
 - the necessary technical operations;
 - its general features.

The technological objects necessary for TDM

- 24 To mobilise text- and data-mining practices, researchers need:
- input data;
 - processing tools;
 - output data, when processing tools are applied to the input data.
- 25 **The raw material: Input data.** Technically speaking, TDM could be described as a process for the extraction of knowledge from selected texts and data, based on mechanisms for the identification of hitherto unknown structures that are scientifically valid and exploitable. Each scientific field (astrophysics, molecular biology, sociology of organisations, marine geology, linguistics, etc.) has developed its own arsenal of resources and techniques for data collection: sensors, probes, satellites, sequencers, cameras, digitisation, simulation, chemical analysis, and so on, which lead to the accumulation of huge datasets (Big Data).
- 26 Overall scientific production is also experiencing spectacular growth due to the increase in scientific activity around the world, exacerbated by pressure to publish. Global annual production of scientific publishing has progressed significantly; for example, between 1996 and 2012 it rose from 1,134,000 to 2,250,000 articles per year on Elsevier’s Scopus database.¹⁷ However, “despite its rapid growth, publication is not increasing in proportion to the production of research data”: more than 90% of data remain stored on local hard disks and are consequently not shared.¹⁸



<http://informatique.in2p3.fr/li/spip.php?article327>

- 27 The process of scientific analysis requires the use of research data in the broadest sense, concerning both scientific publications and the raw data, and including tables, images, statistical data, sounds, other bodies of text, and in general all the information needed by the researcher.
- 28 High Energy Physics and the LHC (Large Hadron Collider) provide an extreme example of this new scientific practice centred on data: a technical device built by engineers – the collider – is associated with instruments designed by researchers – the detectors – to study the particles emitted during the billions of collisions produced. Something like 15 petabytes of data are produced every year. Such a volume of data requires both colossal technical processing resources and a data management infrastructure capable of providing maximum availability to research teams.
- 29 But, before they can be made available, the data must be prepared. For example, it is easy to understand that the major earth and ocean observation programmes, which are usually conducted cooperatively as international operations, require processes for preparing the data before any data mining can be undertaken. These activities, grouped under the term “digital data curation”, include the selection, verification, standardisation, annotation, reformatting, enrichment and structuring of the data collected; the ultimate goal is to have qualified data that can then be used for scientifically valid data mining.
- 30 The leading digitisation programmes for old books thus include processes to create information-rich metadata, facilitating the extraction of knowledge.
- 31 The data repositories resulting from these curation processes are then made available to researchers through thematic access portals such as the International Virtual Observatory (astronomy), the World Observatory of Biodiversity, or the portal of the World Climate Data and Monitoring Programme.

- 32 Processing capabilities obviously play an essential role in the development of TDM practices. The capabilities of computer centres are constantly improving, and they are organised to perform grid computing at the international scale, and to respond to the challenges of processing the massive flows of data generated by research into the climate, the environment, health and astronomy. At the end of the processing chain, there is a need for tools capable of representing and/or displaying complex data, which require high levels of computing power.
- 33 **Analytical software.** In addition, the research practices of the different scientific communities are undergoing change, moving towards more collaborative research involving multiple actors, and are based on the latest technologies for processing and exploiting large bodies of data produced in a shared environment.¹⁹
- 34 TDM technologies aim to reveal relationships between the data items analysed, detect links of cause and effect, establish models, and validate their reproducibility. To do this, depending on the types of data and the proposed objectives, TDM is used alongside techniques from descriptive statistics, data analysis (exploratory statistics) or informatics (artificial intelligence).
- 35 Text mining uses the same techniques as data mining but it is first necessary to process the textual data with language technologies to make them compatible with the methods used in data mining. The use of data structures adapted to the properties of the texts and semantic algorithms is specific to text mining.
- 36 The data-analysis methods used (factorial, discriminant, principal components, multiple components, correspondences, etc.) bring out the different internal dimensions of the datasets, revealing the parameters that show how the data is organised. The classification methods (clustering, unsupervised learning) can identify groups of elements. The purpose of regression methods and supervised learning (artificial intelligence) is to predict the evolution of certain behaviours as a function of other variables.
- 37 TDM software contains features comprising all or part of this processing chain, ranging from access through to the preparation of data, including the application of selected computing algorithms (learning), the exploitation of output, the exploitation of models, and the visualisation of results.
- 38 The software specific to text-mining processes natural language by grammatical labelling, syntax rules, ontologies, learning from labelled corpora, etc. Based on the corpora of documents thus structured, different analytical algorithms can be used: automatic classification, analysis of trends, rules of association, etc.
- 39 Many tools and techniques for the automatic analysis of texts and data have been developed by French research laboratories and are already used by researchers.
- 40 **At the French Alternative Energies and Atomic Energy Commission (CEA),** TDM is widely used in particle physics, nuclear physics and astrophysics.²⁰ Using smart software to analyse interoperable archives is fundamental to European projects such as the European Virtual Observatory, <http://www.euro-vo.org/>, or EUROPLANET, <http://www.europlanet-eu.org/>.
- 41 In other areas, TDM is used by CEA teams for automatic language processing, in particular for the compilation, visualisation and analysis of networks of citations (<http://clair.eecs.umich.edu/aan/index.php>) or to improve the performance of

specialised search engines (<http://aclasb.dfki.de/> or <http://saffron.insight-centre.org/acl/>).

- 42 **At INRA**, the Bibliome research team (MaIAGE-INRA Unit) has developed many TDM applications for the texts of scientific articles in areas of interest to INRA. Some of these projects also use references (PubMed), patents (EspaceNet) and professional journals (e.g. *Perspectives Agricoles*).
- 43 **The Lisis UMR** (“joint research unit”, in this case run by INRA, ENPC and UPEM) (<http://www.inra-ifris.org/>), associated with the **IFRIS**, is developing a digital platform entitled CorTeXT for the analysis of textual corpora (<http://www.cortext.net/>) in the framework of research in the human and social sciences.
- 44 **The CIRAD** is also concerned, through the IATE UMR (CIRAD, INRA, SupAgro, University of Montpellier II), and develops innovative methods and tools for the processing of data and knowledge. The objective is to propose decision-support methods and tools for use in the overall management of biomass transformation processes. These methods and tools must be capable of collecting, representing and managing different types of data and knowledge, including imperfect data (for example, unreliable or imprecise data, and so on), expert opinions, and process engineering models. In addition, the tools proposed must be capable of taking multiple criteria into account, as well as the preferences and arguments of the actors in the agricultural sector.
- 45 **At IRSTEA**, scientists from the TETIS UMR use innovative TDM solutions in the framework of research conducted in collaboration between IRSTEA, INRA and AgroParisTech: exploring scientific articles to identify new research themes by the enrichment of semantic resources (termino-ontological resources for each speciality, thesauruses, etc.) or improved tools for epidemiological surveillance in animals. In addition, TDM can identify relationships within heterogeneous datasets in very voluminous corpora that include both scientific and non-scientific texts (datasets and databases, images, etc.), which has led to the discovery of new and complementary knowledge (<https://tetis.teledetection.fr/index.php/fr/>).
- 46 **The GESTE UMR** (IRSTEA, ENGEES), in Strasbourg (<http://geste.engees.eu/>), currently has scientists working on the problem of emerging pollutants in water and changes in the behaviour of individuals and artisans to reduce emissions of these pollutants. This is a major issue in terms of health and the environment, still insufficiently understood in terms of either the risks or the solutions. The use of TDM will enable risk-mapping, in both its scientific and societal aspects, clarification of chronology, and the identification of sub-themes as well as key institutions and actors.²¹
- 47 **INRIA** has also developed GROBID (GeneRation Of Bibliographic Data),²² an automatic learning (or machine learning) tool available in open source. This application can extract, analyse and restructure scientific publications in raw formats (e.g. PDF) into a Text Encoding Initiative (TEI) format.
- 48 **Other tools.** Software such as Alceste²³ (developed by the company IMAGE and the CNRS) and Calliope²⁴ (developed by Astefo) are specialised in lexical analysis. The Gargantext project provides an example of such an analysis. It can be used to analyse texts according to a paradigm other than complete words:

“It smees that teh oredr of the leettrs in a word has no intorpance. The frist and the last leettr must be in the rghit plcae. The rest can be in a total disrodre and can stlil be read with no proelbm. This menas that we don’t raed each leettr in itslef,

but the word as a whole. A change in the frame of reference and we transpose the result to the text itself: the order of the words is unimportant compared to the context, which counts much more: counter-texted with Gargantext.”²⁵

- 49 Weka is one of the most widely used TDM applications in the world. It was originally developed by the University of Waikato in New Zealand, and enables data to be viewed and analysed rapidly. Finally, many commercial software applications have been produced by companies, such as RightsDirect,²⁶ KNIME²⁷ or RapidMiner.²⁸
- 50 Scientific data and articles, which are the primary source of information for analysis by TDM, may be imported from platforms for the sharing of scientific and technical information (such as digital libraries, databases, open archives, search engines, etc.). Some platforms are attempting to facilitate free access to scientific production such as, in France, **HAL**, **Gallica**, **NAKALA**, **ISIDORE**, **OpenEdition** and **Persée**.²⁹
- 51 **ISTEX**. Lastly, value-added services are being developed on the digital platform for access, sharing and enrichment of scientific and technical information (STI) of the ISTEX project (Excellence Initiative of Scientific and Technical Information – ANR-10-IDEX-004-02). These services are designed to enable users to analyse large bodies of data automatically, in particular via the following services:
- data enrichment;
 - machine-induced datasets;
 - nano-publications;
 - collaboration patterns;
 - influence patterns;
 - semantic analysis;
 - impact analysis;
 - automatic document generation.
- 52 Many other platforms for sharing scientific and technical information are listed by the CNRS and described in order to promote initiatives for pooling knowledge and free access to scientific production.³⁰
- 53 **Results.** Once the TDM process has been applied to the datasets, a result is generated automatically. This result is presented differently depending on the analytical technique used, which depends in turn on the analytical approach chosen by the researcher (“user-generated content”).
- 54 The result is new knowledge. This can be of several different types and exploited in several different ways:
- recommendation systems seek to predict behaviour by analysing and filtering the information contained in the data;
 - in science, the links identified between data items are used to recommend new research hypotheses.
- 55 The result produced becomes part of the scientific approach as new analytical data.

The technical operations behind TDM

- 56 It is also important to understand the technical processes underlying every processing operation in TDM.
- 57 **Atilf.** In a note dated 14 May 2014, Jean-Marie Pierrel, representing the University of Lorraine (acting on behalf of the Conference of University Presidents), who is

responsible for the development of value-added services as part of ISTEEX and Director of Atilf (Analysis and Computer Processing of the French Language), explained the technical operations carried out by the laboratories in the framework of the development of TDM. In his note, Mr Pierrel makes a distinction between two phases:

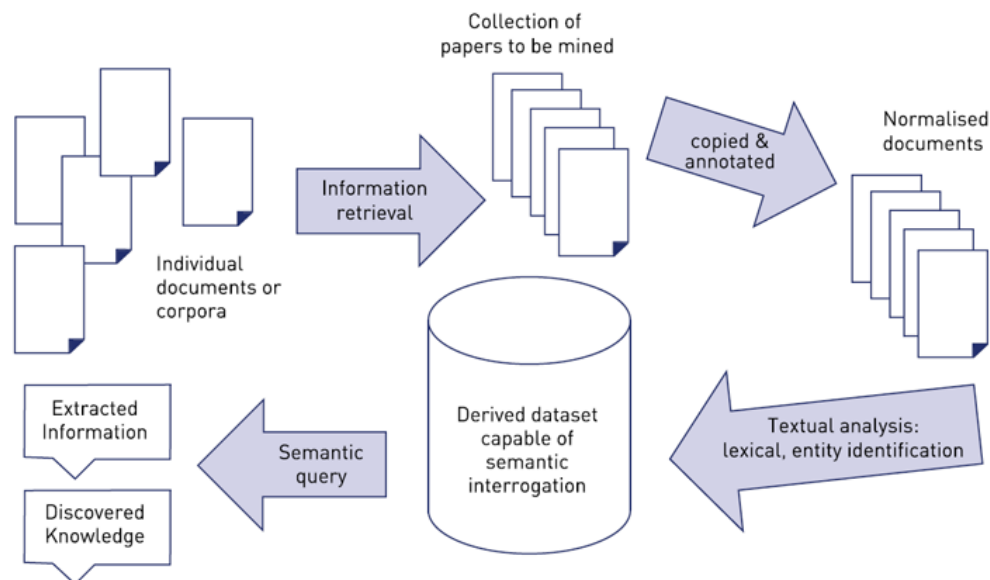
1. Work carried out on computers:

- work is copied from sub-corpora onto computers;
- linguistic annotation of the data;
- detection of terms and construction of a terminological reference framework:
 - the detection of named entities and construction of a reference framework of named entities;
 - annotation and lemmatisation of texts, and construction and management of glossaries;
 - tagging of bibliographical databases.
- use of these annotations in the procedures for selecting articles following a user request;
- mapping on the basis of the elements annotated.

2. The distribution of the results of searches using TDM

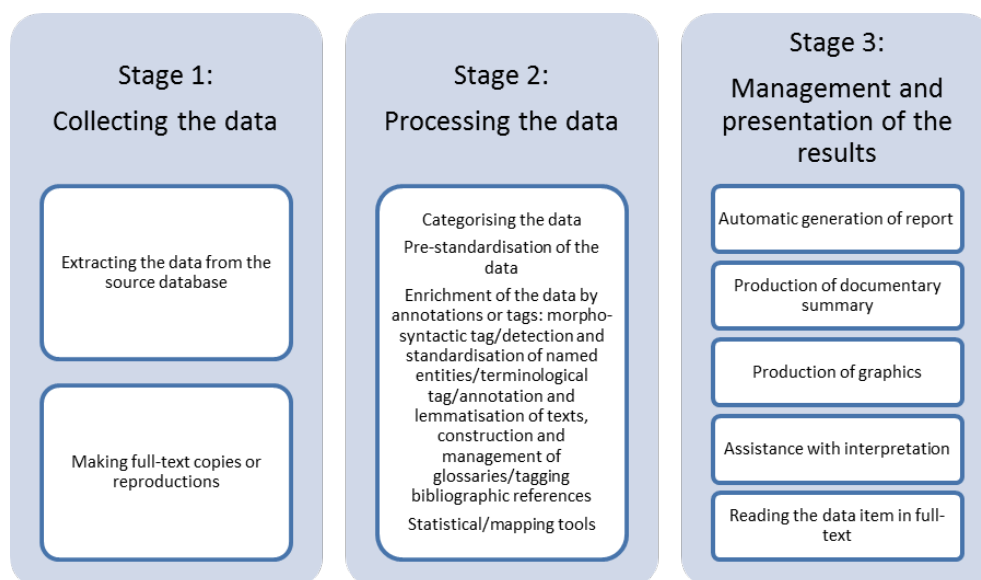
- in scientific publications;
- as the dissemination of the results of TDM.

58 **Illustration of TDM.** The following diagram illustrates the steps in the processing of information in the framework of TDM applied to direct marketing, whose principles can be transposed to the scientific field:³¹



32

59 **Summary diagram of TDM operations.** Considering all these elements, the processes for dealing with information by TDM are complex and can be summarised in three stages. The technical operations carried out in each of the steps are detailed below:



- 60 Each of these steps needs to be clearly defined and secure, standards and formats must be defined, and a procedure for carrying out the TDM operations must be drafted and possibly standardised.

“Macro” functions within TDM

- 61 Taking a functional macroscopic view of TDM provides another way of grasping the concept as regards its objectives.
- 62 The development of TDM techniques opens new perspectives for the analysis of large volumes of texts and data (Big Data). TDM uses an inductive approach to analyse all this information grouped in corpora. Text mining encompasses a set of practices and a variety of methods that are difficult to treat in a unified way.³³ Yet it is possible to paint a comprehensible picture in broad strokes.³⁴
- 63 The vast increase in the number of scientific articles in digital format has made a wide variety of information available: tables, images, statistical data, sounds, corpora of text, and so on. The latter may be written in very different styles (language, expressions, etc.). All these styles and genres of texts to be analysed thus require that researchers using TDM first create coherent corpora of data. The purpose of TDM is to achieve a specific objective. Data are not mined to give a general insight into the contents of a corpus, but as a way of exploring them to answer a specific question.
- 64 TDM therefore analyses a set of data according to a criterion of novelty (what emerges when all the texts in a corpus are cross-referenced, for example?) or a criterion of similarity (what theme recurs in all the texts of the corpus?). It is then for the researchers to find the meaning behind all the cross-references among the data. After selecting and transforming the data (according to their coding, in particular), the researcher uses TDM tools to arrive at an interpretation and assessment. Each of the stages of TDM is therefore essential, from pre-processing and formatting of data through to the conclusions.
- 65 Let us take an example to better understand these techniques. INRA is currently conducting a study with the help of text mining in order to identify new species of fish suitable for captive breeding, i.e. species that could be adapted for aquaculture.

Considerable amounts of data relating to fish (breeding, feeding, living environment) are analysed using text-mining technologies. Many of these data come from the ISTEEX corpus. The goal is to identify the standard features of aquaculture fish in order to better understand the phenomenon of domestication, as well as to identify those species of fish most similar to aquaculture species (fundamental research). The species identified could then be used in domestication experiments (applied research). The use of the results could open new opportunities for fish-farming.

The economics of TDM

- 66 TDM can generate both financial and social benefits:
- lower costs and higher productivity;
 - more innovations in products and services.
- 67 In addition, it is also important to analyse the impact of TDM on the world of publishing, in order to defuse any potential conflict while clarifying and securing the practices.

Reduced research costs and more numerous discoveries

- 68 The economic potential of TDM lies primarily in the reduction of research costs and the possibility of more numerous discoveries.
- 69 By reducing processing costs and saving time, TDM makes it possible to produce many more new research articles, thus enriching databases, etc. in a “virtuous circle”.
- 70 The Jisc,³⁵ a British organisation in charge of digital services for research, has estimated that the use and development of TDM for research would increase the productivity of public research without additional cost. TDM would provide the equivalent of 4.7 million extra hours of research work per year, for the whole of the United Kingdom.
- 71 In the same vein, TDM enables the development of interdisciplinary research activities. Corpora could be constructed covering a range of subjects, to give access to the latest progress in biochemistry for a biological study, or to reveal historical changes in sociological analyses.

More innovation

- 72 A large proportion of this new knowledge can be converted into business innovation. For example, in the field of healthcare, an analysis using PubMed³⁶ came up with new hypotheses for research. A group of researchers extracted data from a corpus of medical articles. They looked for links between “drug-disease” and “drug-phenotype”.³⁷ While browsing through all these articles, they identified genes that could be responsible for diseases because of the toxicity of certain medicinal drugs. This enrichment of the data could enable researchers and doctors to rapidly diagnose new links between a “drug” and a “disease”. From an economic point of view, this discovery could enable great progress in medical research, with limited labour costs.
- 73 The economic benefits resulting from the use of TDM therefore arise essentially from the role it can play as a source of incremental innovation. In this respect, the Jisc³⁸

emphasises that TDM “unlocks”³⁹ knowledge, which must inevitably provide both economic and social benefits.

- 74 Furthermore, TDM is a scientific and digital technique that requires considerable human capital if it is to be applied correctly. Text and data mining can therefore provide opportunities for specialised technical employment, primarily for engineers and IT specialists. In addition, research using TDM requires infrastructures and services whose creation and production will also contribute to economic growth.

Economic issues of TDM and the world of publishing

- 75 The economic issues raised by TDM go hand in hand with those facing the world of publishing. TDM is having a profound effect on the traditional notion of copyright. While there will inevitably be questions, there should be no doubt about the advantages of this practice compared to the current publishing model.
- 76 Some publishers have already protested that their profession will quite simply disappear in the light of an analysis of Article 17 of the Digital Republic Bill (Article 30 of the *Petite Loi*), which establishes a right of access to scientific publications after expiry of an embargo period. This argument was not retained in the results of the Impact Study of Article 17, published by the French National Assembly on 9 December 2015.⁴⁰
- 77 This warning has also proved unfounded in the United Kingdom, where an exception to copyright was introduced in favour of text and data mining in 2014, as well as in the United States, where the practice has also been given legal status (the case *Authors Guild v. Google* – 14 November 2013).
- 78 Furthermore, the supposed threat from TDM to the profits of the world of publishing does not seem to exceed the possible benefits to society of intensive exploitation of data. The European Commission therefore agrees about the long-term benefits of TDM and has inserted an exception in favour of TDM in the proposed Directive on Copyright in the Digital Single Market.

NOTES

1. <http://www.rithme.eu/?m=resources&p=dmdomains&lang=en>.
2. CSPLA report, p. 8.
3. http://ec.europa.eu/internal_market/copyright/docs/studies/1403_study2_en.pdf
4. http://ec.europa.eu/research/innovation-union/pdf/TDM-report_from_the_expert_group-042014.pdf
5. Proposal for a Directive of the European Parliament and of the Council on Copyright in the Digital Single Market, 14.09.16, COM(2016) 593 final, <http://ec.europa.eu/transparency/regdoc/rep/1/2016/EN/1-2016-593-EN-F1-1.PDF>
6. Press release of 14 September 2016.

7. Introducing a TDM exception to copyright under English law, based on Article 5, 3a) of Directive 2001/29/EC of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society (the InfoSoc Directive). An analysis of the texts of the Directive in their English and French versions makes it clear that the scope of the exception proposed is not the same depending on the language used.
8. Draft Directive, Article 2: “‘Research organisation’ means a university, a research institute or any other organisation the primary goal of which is to conduct scientific research or to conduct scientific research and provide educational services:
 - (a) on a non-for-profit basis or by reinvesting all the profits in its scientific research; or
 - (b) pursuant to a public interest mission recognised by a Member State; in such a way that the access to the results generated by the scientific research cannot be enjoyed on a preferential basis by an undertaking exercising a decisive influence upon such organisation.”
9. <http://ec.europa.eu/research/index.cfm?pg=newsalert&year=2016&na=na-140916>
10. http://www.legislation.gov.uk/ukxi/2014/1372/pdfs/ukxi_20141372_en.pdf
11. <https://www.law.cornell.edu/uscode/text/42/2000ee-3>
12. <https://copyright.gov/fair-use/summaries/authorsguild-hathitrust-2dcir2014.pdf>
13. <https://www.authorsguild.org/wp-content/uploads/2015/10/CA2-Fair-Use-Ruling.pdf>
14. “Comment l’affaire Google Books se termine en victoire pour le Text Mining”, 21 October 2015, <https://scinfolex.com/2015/10/21/comment-laffaire-google-books-se-terme-en-victoire-pour-le-text-mining/>
15. <http://www.cric.or.jp/english/clj/cl2.html>
16. The comparative table can be found in Annex 1.
17. CNRS STI Strategic Orientation Plan, *Les publications scientifiques : une augmentation continue et forte* (Scientific publishing: Growing constantly and rapidly), p. 15.
18. CNRS STI Strategic Orientation Plan, *Données et publications : une course poursuite* (Data and publication: A downhill race), p. 15.
19. CNRS STI Strategic Orientation Plan, p. 16.
20. See the work of the Strasbourg Astronomical Data Centre at <http://cdsweb.u-strasbg.fr/about>
21. These examples are drawn from a note by the Association of STI managers of research organisations (EPRIST) on text and data mining, *Le TDM comme outil innovant de recherche scientifique* (TDM as an innovative tool for scientific research), <http://www.cnrs.fr/dist/z-outils/documents/EPRIST%20text%20et%20data%20miningV3.pdf>
22. https://www.researchgate.net/publication/221176095_GROBID_Combining_Automatic_Bibliographic_Data_Recognition_and_Term_Extraction_for_Scholarship_Publications
23. <http://www.image-zafar.com/Logiciel.html>
24. <https://www.calliope-textmining.com/>
25. <http://gargantext.org/>
26. <http://www.rightsdirect.com/>
27. <https://www.knime.org/>
28. <https://rapidminer.com/>
29. List (with descriptions) of STI sharing platforms: <http://www.cnrs.fr/dist/acces-ist.html>
30. <http://www.cnrs.fr/dist/acces-ist.html>
31. <http://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining>
32. ©Jisc CC BY-NC-ND
33. Yannick Toussaint, *Extraction de connaissances à partir de textes structurés* (Extracting knowledge from structured texts), *Document Numérique*, 8(3), 2004, pp. 11–34.
34. Fidelia Ibekwe-Sanjuan, *Fouille de textes : méthodes, outils et applications* (Text mining: Methods, tools and applications), coll. *Systèmes d’information et organisations documentaires*, Hermès, 2007, 352 p.

35. <https://www.jisc.ac.uk/guides/text-and-data-mining-copyright-exception>
36. <http://www.ncbi.nlm.nih.gov/pubmed>
37. "A collaborative project between CTD and Pfizer: Manual curation of 88,000 scientific articles text-mined for drug-disease and drug-phenotype interactions", <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3842776/>
38. <https://www.jisc.ac.uk/guides/text-and-data-mining-copyright-exception> page 38
39. Jisc, p. 38.
40. Digital Republic Bill, NOR: EINI1524250L/Bleue ETUDE D'IMPACT, 9 December 2015 : "Focus 1: Economic impact on institutional scientific publishing in France[:]
The impact of this measure on the economic equilibrium of French institutional scientific publishing, which essentially consists of human and social science publishers, must be assessed relatively, insofar as the majority of their current turnover comprises subsidies from establishments or laboratories. The journals, moreover, represent only 18% of their publishing production on average, while between 40% and 60% of the overall sales figure associated with these journals is achieved from the publications for the year, which would remain subject to embargo at maturity of the proposed measure, ensuring that these players would be only marginally affected."
<http://www.assemblee-nationale.fr/14/projets/pl3318-ei.asp>

What issues does this raise for the work of science?

- 1 TDM brings science and scientific research into a new age. TDM and Open Science have multiple advantages:
 - TDM clearly enables scientists to broaden their approach via the automated analysis of scientific literature and all kinds of data;
 - it also makes it easier to organise and structure research projects in France and to rationalise the decision-making processes for setting priorities and allocating budgets;
 - it creates new avenues for the exploitation of research output;
 - TDM can also contribute to public decision-making.

Analysing scientific publications and data

- 2 TDM enthusiasts are especially interested in scientific literature, articles from journals, monographs, theses and reports. All this information expressed in natural language, which is rather unstructured when compared with quantitative data, offers vast and varied reserves. Each scientific article is the result of a logical research process in the framework of a given project with a declared goal, and has a place in a scientific context specified and reflected by the article's references. When TDM is used to analyse large corpora of scientific literature covering a variety of periods, methodologies and equipment situations, either for a given issue or limited to a specific field of study, it can develop products and services of great benefit to researchers:
 - diachronic longitudinal studies on the evolution of scientific practices, the emergence of new concepts, the merging or separation of scientific fields, new and evolving relationships between different disciplines;
 - the quantity of information produced by researchers (2.5 million articles a year) shows the importance of using knowledge extraction technologies to help researchers access the contents of their colleagues' articles. Sequential reading is likely to be replaced, at least partly, by browsing through knowledge sets resulting from the initial corpora. Text- and data-mining activities thus have direct consequences for the way researchers access scientific knowledge;

- the improvement in the relevance of publishers' access portals via semantic enrichment and annotations and also via "recommendation" features are all vital for a better sharing of the knowledge contained in the articles;
 - automatic categorisation, the classification of published work, links between publications beyond simple citations: these all use natural language for their analyses and for building derived knowledge;
 - revamped bibliometrics and the development of new metrics and ways of measuring the impact of publications, leading to progress in the way researchers, structures, programmes and institutions are assessed;
 - changes in scientometrics: retrospective analysis of fields of research, measurements of research performance, cross-fertilisation between disciplines;
 - seeking out and managing pools of experts;
 - the development of the predictive approach by automated generation of hypotheses from mining scientific literature.
- 3 TDM therefore fosters the intensification of research. It stimulates fundamental research while also driving applied research. First of all, TDM makes quantitative analyses possible that are unprecedented in the history of science. Never before has it been possible to access, and potentially analyse, so many corpora of texts and data. The researcher's task of scientific interpretation does not change in itself, but the range of research is expanding. Intensification also requires a transformation of the way in which researchers perceive science. Entire paradigms may need to change, as the *Gargantext* example shows.
- 4 TDM makes it possible to find new correlations or emerging trends. As a tool for derived research, text mining can reveal interactions that are often new or previously impossible at transdisciplinary level.
- 5 Combining TDM with the dissemination of research data and the acceleration of knowledge acquisition will facilitate the reproduction of experiments and make it easier to assemble and assess evidence. In some cases, it should be easier to refute findings, leading to greater scientific relevance. Verifying research is an offshoot of the development of TDM.

Setting out the issues and designing research projects

- 6 Science progresses as the result of a series of methodological revolutions concomitant with an enrichment of the methodological arsenal available to researchers. The advent of digital science, and the pervasiveness of instruments generating massive amounts of data, has led to the birth of what Hey, Tolle and Tansley have called "data-intensive scientific discovery" (*The Fourth Paradigm*).
- 7 Scientists now entrust observation and measurement tasks either to single instruments shared with colleagues, such as particle accelerators or on-board telescopes, or to networks of many small instruments such as oceanographic buoys, and weather or seismic stations. A science of data interpretation then develops downstream, of which TDM is the main pillar. But this development of algorithms for data analysis has to be backed up by less visible but equally important functions: infrastructures for calculation and for the management and sharing of data.

- 8 The scientific approach has traditionally been based on deductive reasoning, starting from theories serving as premises to be tested. Hypotheses are based on theoretical knowledge, while the observations collected (data) are used to confirm these hypotheses (or refute them if they challenge the theory). A large part of Western education is based on this approach, as soon as children reach the “age of reason”. But this pre-eminence of theory over experimentation has been challenged over the last few years, notably with the arrival of Big Data and the development of TDM.
- 9 This way of reasoning has been turned on its head: correlations revealed between aspects of the observations concentrated in the datasets suggest hypotheses from which to establish theoretical models of behaviour. Of course, this inductive reasoning does not guarantee the conclusion and can lead to multiple successive iterations. Abductive reasoning, involving logical inference *from* observations *to* theory, focuses on the simplest and most probable hypotheses and has been described as “inference to the best explanation”¹. Inference engines using artificial intelligence are often based on this pragmatic abductive approach, which has proved so fruitful for developing TDM software.
- 10 The use of TDM and the organisation of knowledge sharing, together with the changes in scientific methods, will enable research communities to build new research projects and discover new topics.
- 11 This way of constructing a project in advance by the use of TDM offers enhanced control and productivity compared with the way decisions about research projects are currently made. Research projects structured like this will enable better allocation of resources.

Optimising the governance of scientific systems

- 12 TDM will thus contribute to a better “governance of scientific systems” by streamlining “the decision-making process”, and by “priority setting, the allocation of funds and the management of human resources in a way that effectively responds to the concerns of the various stakeholders involved in the system”.²
- 13 Indeed, TDM analysis of the results of research funding programmes (publications, patents, etc.) more accurately identifies the most fields that have proved most fruitful, as well as showing links between disciplines. Tools for viewing or mapping cooperation across different disciplines or geographical areas, either nationally or internationally, can bring out this information, which is so essential for guiding scientific research.
- 14 The use of this information by those with decisions to make about scientific policy helps organisations define their scientific strategies, both in the field of basic research and in its financial exploitation.

Exploiting scientific data

- 15 TDM and the massive analysis of data will not only make it possible to exploit the new knowledge resulting from automated processing and the subsequent development of innovations and discoveries, it will also enable the exploitation of the masses of unused

data saved on the hard drives of researchers and not shared, despite their scientific value.

Exploiting “lost science”

- 16 A “water tower effect”³ is possible for the entire world of research, with the processing of forgotten data.
- 17 At the moment, large quantities of data collected during experiments are “lost”. It has been estimated that about 10% of these data are published, while 90% remain on computer hard drives. In some disciplines, valid and important results remain unpublished and much of the data is underused or lost (this is particularly the case of data from negative outcomes that are simply forgotten). When it comes to the output of large-scale instruments, the raw data collected are so massive that they are processed directly online without being stored, such as for example those provided by satellite observations.⁴
- 18 For Cristinel Diaconu, the loss of data seems to be sadly common:
- “When we want to access the data, either they can no longer be found, or we can find them, but we don’t know what to do with them because we don’t understand what they mean. Worse still, data have sometimes been destroyed by the researchers, who thought they were useless after the end of a project. You don’t realise it at the time, but ten years later, you might be working on a project with echoes of the previous project, and the potential for discovery is lost because you no longer have the funding to repeat these manipulations.”
- 19 Yet these archived data can be a real treasure trove. Cristinel Diaconu has calculated the financial value of these data for his own field: “My team and I realised that the additional cost of preserving the data is in the order of one thousandth of the total budget. Yet the publication of new articles resulting from the exploitation of archives in the following five years provides a profit of 10%. This is research that costs practically nothing! If there is no strategy for preserving data, we miss potential discoveries and research at very low cost. If they have been properly preserved, data cost almost nothing.”

Free reuse of the results of TDM

- 20 Article 38 of the Act for a Digital Republic stipulates that “files produced at the conclusion of the research activities for which they were produced ... constitute research data.”
- 21 The legal framework governing “research data” is laid down in:
- Article 30 of this same Act: “Once the data from a research activity financed at least 50% by grants allocated by the state, by regional or local authorities or public institutions, by grants from national funding agencies or by European Union funds, are no longer protected by specific rights, or special regulations, and they have been made public by the researcher, the research establishments or organisation, they can be freely reused.”
 - Act No. 2015-1779 of 28 December 2015 on free access to and the terms of reuse of public sector information (known as the Valtier Act), which brings teaching establishments and institutions into the age of Open Data and imposes the obligation to make their data

available “in electronic form” complying with an “open and easily reusable standard that can be exploited by an automated processing system”.

- 22 These research data therefore constitute a source of knowledge that must be made available to the scientific communities for the purposes of acquiring new knowledge and enabling new research work.
- 23 However, Open Science must not hamper the financial aspects of research.
- 24 The provision of scientific data on open science platforms must not impede:
 - the exploitation of data, including by patents, and respect for secrecy and specific provisions, such as Restricted Regime Areas;⁵
 - respect for contractual rules of confidentiality.
- 25 The way research data is made available must also be organised to take into account the different practices of the various scientific communities.

Developing these innovations

- 26 TDM techniques rely on innovative tools (software, supercomputing, technologies for the massive collection and processing of data), and French public research organisations and French companies contribute to their development.
- 27 TDM is currently one of the major sectors with potential for innovation within the digital economy, and these technologies are central. This can be seen in the awards acknowledging work on TDM. Xavier Tannier, from the CNRS’s LIMSI laboratory at Orsay (Laboratoire d’information pour la mécanique et les sciences de l’ingénieur) and Iona Manolescu of INRIA have received a “Google Award”⁶ for their algorithm for text mining in print media. In the case of medical informatics, Pierre Zweigenbaum⁷ was inducted into the American College of Medical Informatics in 2014.
- 28 TDM has potential benefits for the French economy: several start-ups in France have been created as a result of the need for research projects to develop TDM tools jointly with private partners (at the CEA in particular).
- 29 In his article “Mining external R&D”,⁸ Alan Porter insists on the desirability of such innovation for businesses. Companies with research and development centres can certainly expect to benefit from applied research on a broader scale, using larger corpora. The advantage is even greater for companies that do not have an R&D centre. In this respect, text mining is a factor for growth. This is the result of the positive externalities of its development, based on three points: product innovation, productivity, and increased consumer satisfaction. The extent of this economic impact is correlated with the legal freedom to carry out TDM: the broader the possibilities of TDM, the greater the economic impact.

Support for public decision-making

- 30 TDM can help improve decision-making. In the public sector, it would facilitate the development of evidence-based policies (EBP). EBP is a policy approach based on the empirical analysis of situations. This practice, which consists in analysing large factual databases for decision-making purposes, was originally developed in the medical field (evidence-based medicine) at the beginning of the 1990s.⁹ This approach then spread to

other areas of public decision-making, such as the protection of the environment and security.

- 31 In the United Kingdom, these empirical judgements started to take an important role in public life under the government of Tony Blair. A User Guide was even drawn up by the Overseas Development Institute (ODI) in order to disseminate these practices.¹⁰ The concept is divided into several branches: evidence-based decision, evidence-informed decision, and evidence-aware decision. TDM can optimise economic, social and environmental policies, as it is not based only on opinions and theoretical models, but also on factual analysis.
- 32 TDM can be used for decision support in other fields (not necessarily public), for example in geography (economic geography, social geography, geomarketing, etc.). The idea is to cross-reference geo-tagged data in order to identify the characteristics of geographical areas. To do this, “the descriptive techniques used in data mining and more precisely Agglomerative Hierarchical Clustering (AHC) were used to identify the number of homogeneous groups based on Ward’s aggregation criterion and Euclidean distance, using the R Data Mining Software.”¹¹ The data processed by TDM can thus be used to support Geographic Information Systems (GIS).
- 33 Econometrics is another data-driven science that could benefit from progress in TDM. This discipline enables observers to monitor the way the economy is really functioning, which is useful in situation analysis and public and private decision-making. However, there are currently few links between econometrics and TDM/machine learning/Big Data. A possible explanation is the difficulty in determining causal links and in quantifying the impact of a variable on an observed phenomenon.¹²
- 34 Finally, the research activities and publications resulting from the use of TDM require a high degree of transparency and rigour as regards methods, peer reviews and external influences. This information helps researchers better understand the systemic aspects of the topic studied and also provides a means of assessing the reliability of the results. The analysis of empirical facts through TDM technologies thus offers – in some cases – the possibility of analysing the veracity of political statements and providing support for decision-making.¹³

NOTES

1. https://en.wikipedia.org/wiki/Abductive_reasoning.
2. OECD, *Governance of Public Research: Toward Better Practices*, 2003.
3. The expression “water tower effect” refers to an economic concept. This is an echo of the “trickle-down” theory, extended to include innovations and not just tax policy. To summarise, TDM is a set of techniques that can be used in numerous sectors; it “flows” over them and can therefore “irrigate” different economic activities.
4. COMETS, “The ethical issues of scientific data sharing”, 2015, p. 4.
5. Article R.413-5-1 of the French Penal Code: “restricted regime areas, as referred to in Article R. 413-1, are those for which protection is imperative in order to prevent essential elements of the

scientific or technical potential of the Nation from:

- 1° being appropriated in a manner that could lead to weakening the Nation's defences, compromising its security or prejudicing its other fundamental interests;
- 2° or being diverted for purposes of terrorism, proliferation of weapons of mass destruction and their means of delivery or contribution to the increase of military arsenals.

The restricted regime areas may include, within their scope, premises for which enhanced protection is justified by the storage of products or by the execution of activities involving specific risks with regard to the imperatives mentioned in the first three sub-paragraphs."

6. <https://archives.limsi.fr/Actualites/GoogleAward.html>

7. <https://perso.limsi.fr/pz/>

8. <http://www.sciencedirect.com/science/article/pii/S0166497211000113>

9. Catherine Laurent, Jacques Baudry, Marielle Berriet-Sollic, Marc Kirsch, Daniel Perraud, Bruno Tinel, Aurélie Trouvé, Nicky Allsopp, Patrick Bonnafous, Françoise Burel, Maria José Carneiro, Christophe Giraud, Pierre Labarthe, Frank Mastose and Agnès Ricoch, "Pourquoi s'intéresser à la notion d'« evidence-based policy » ?" *Revue Tiers Monde*, 4(200), 2009, <http://www.cairn.info/revue-tiers-monde-2009-4-page-853.htm>

10. Overseas Development Institute, *Evidence-Based Policymaking: What is it? How does it work? What relevance for developing countries?*, November 2005.

11. Marwa Chalgham, Abderrahmane Fadil and Abdelaziz Dammak, *Le Data Mining pour l'aide à la décision en géomarketing* (Datamining as a decision support tool in geomarketing), ROADEF - 15^e congrès annuel de la Société française de recherche opérationnelle et d'aide à la décision, February 2014, Bordeaux, France, <hal-00946452>

12. Quora, "Why is econometrics isolated from the big data/machine learning revolution?", 2013 <https://www.quora.com/Why-is-econometrics-isolated-from-the-big-data-machine-learning-revolution>;

Hal R. Varian, "Big Data: New Tricks for Econometrics", *Journal of Economic Perspectives*, 28(2), Spring 2014.

13. NESTA, *Using Research Evidence: A Practice Guide*, Section A: "What is evidence-informed decision-making, and why focus on research?", 2015.

How to organise TDM?

TDM structures and research centres in other countries

United Kingdom: The National Centre for Text Mining

- ¹ In the United Kingdom, the National Centre for Text Mining (NaCTeM) is a body financed by the state¹ and overseen by the University of Manchester. It was set up in 2004, initially to address the needs of the university community.
- ² NaCTeM was set up to develop tools and services related to text mining.² It offers text-mining services, software developed by NaCTeM teams or by others, seminars and workshops on TDM, tutorials and demonstrations, as well as publications resulting from text mining.
- ³ Although this site was initially intended to assist the academic world in data mining, it now has a much wider scope and audience, including industry.³
- ⁴ NaCTeM proposes to develop tailor-made tools and services to meet the needs of researchers in the academic or industrial worlds, when those available on the website are not suitable in the context of a particular project.⁴ An example of the services proposed by NaCTeM is FACTA+, which stands for Finding Associated Concepts with Text Analysis in the biomedical field:⁵



- 5 NaCTeM strives to develop fully interoperable software tools, as the lack of interoperability is one of the greatest challenges facing researchers.⁶ The NaCTeM website does not list only the services and tools it has developed itself. TDM services and tools developed by third parties are also listed.
- 6 NaCTeM also oversees projects⁷ relating to text and data mining and lists national and international events on the subject.⁸
- 7 NaCTeM was the first national centre devoted to text and data mining.⁹

United States

- 8 In the United States, the National Science Foundation (NSF) is an independent agency of the United States Government, whose mission is to fund basic scientific research. In 2002, \$6 million were made available to support research in data mining in particular.¹⁰
- 9 As for federal agencies, the Data Mining Reporting Act¹¹ requires those who use or develop data mining to submit an annual report to Congress.¹²
- 10 There seems to be no structure in the United States for sharing tools and data for purposes of text and data mining.

International Council for Science (ICSU) and World Data System (WDS)

- 11 The International Council for Science (ICSU), created in 1931, is the largest non-governmental scientific organisation in the world. It is made up of 121 national members and 32 international scientific unions, and is responsible for encouraging international scientific and technological activities and advocating access to science for all, as well as preserving universal access to good quality scientific data for the long term.
- 12 The ICSU created the World Data System in 2008¹³ to promote and facilitate the exchange of data between the members of the WDS. The ICSU World Data System aims to make the transition from separate data centres to a global and interoperable data system, which would incorporate emerging technologies and the new data sciences.

The mission of the ICSU WDS is to promote fair and universal access to reliable and good quality scientific information and data, in all disciplines.¹⁴

- 13 In September 2016, the ICSU WDS had 98 members.¹⁵ In order to become a member, each candidate must be certified,¹⁶ in particular regarding the criteria of integrity and confidentiality:¹⁷

“As part of the process of developing WDS, a certification procedure for evaluating candidates for membership was developed by the Scientific Committee to ensure the trustworthiness of WDS Members in terms of authenticity, integrity, confidentiality and availability of data and services.”

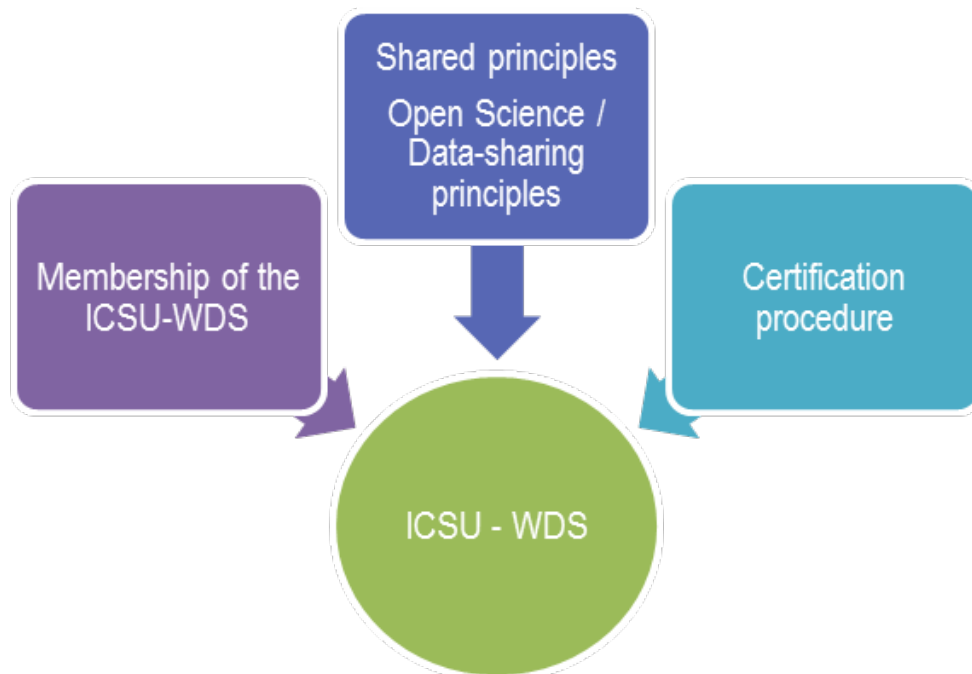
- 14 “Data-Sharing Principles”, a sort of ethics charter, were drawn up, and include the principle of respect for the data source and its integrity:¹⁸

“Data, metadata, products, and information should be fully and openly shared, subject to national or international jurisdictional laws and policies, including respecting appropriate extant restrictions, and in accordance with international standards of ethical research conduct.

Data, metadata, products, and information produced for research, education, and public-domain use will be made available with a minimum time delay and free of charge, or for no more than the cost of dissemination, which may be waived for lower-income user communities to support equity in access.

All who produce, share, and use data and metadata are stewards of those data, and have responsibility for ensuring that the authenticity, quality, and integrity of the data are preserved, and respect for the data source is maintained by ensuring privacy where appropriate, and encouraging appropriate citation of the dataset and original work and acknowledgement of the data repository.

Data should be labelled ‘sensitive’ or ‘restricted’ only with appropriate justification and following clearly defined protocols, and should in any event be made available for use on the least restrictive basis possible.”



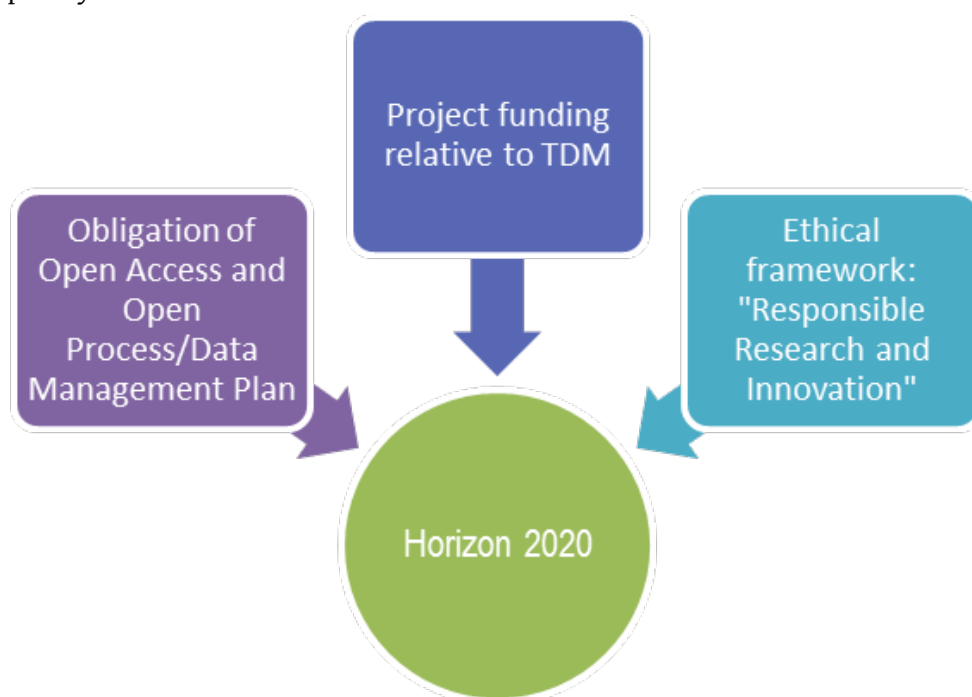
The Research Data Alliance (RDA)

- 15 The RDA is a joint initiative of the European Commission, the American National Science Foundation (NSF) and National Institute of Standards and Technology (NIST), as well as the Department of Industry, Innovation and Science of the Australian Government. Its purpose is to build a social and technical infrastructure to foster the openness and sharing of research data. The initiative is driven by the research communities themselves, and currently has more than 4,300 members from 111 countries. The members of the RDA are divided into Working Groups or dedicated Interest Groups, responsible for developing and validating an infrastructure and supporting the growth of a data-sharing community, including contributors from every scientific discipline and every geographical origin¹⁹.
- 16 The RDA, launched in 2012, has already issued recommendations and tested and harmonised standards in different aspects of data science. These first results relate to aspects of infrastructure concerning the reproducibility of scientific output, long-term preservation and good practice in the maintenance of data repositories, the training of data specialists, citations of data, etc.

The framework proposed for European projects under H2020

- 17 A political awareness of the need to provide a legal framework for TDM operations in the field of research has emerged:
 - in its strategy for a Digital Single Market, the European Commission had referred to the need to adapt the Directive on copyright (the InfoSoc Directive) to technological advances in order to avoid legal disparities within the single market;
 - a report of the European Commission from July 2016 stated that the absence of legal provisions at EU level concerning TDM for research purposes causes uncertainties with adverse effects on research and innovation (R&I).²⁰
- 18 These positions favourable to TDM have contributed to guiding Horizon 2020 towards open access and TDM, and the proposed Directive on Copyright in the Digital Single Market introduces an exception in favour of TDM.
- 19 Horizon 2020 contributes indirectly to the development of TDM by fostering free access to research publications and data. This is because this European programme for research and innovation requires that all research publications and some data stemming from funded projects be released for free access after a certain period of time. Transparent and long-term management of research data must be ensured through the drafting of a Data Management Plan (for projects participating in the pilot scheme). To accompany this approach, the European Commission provides indications, published as guidelines, on good practice for managing research data.²¹ The European Commission's strategy in favour of free access to research data is laid out in a report (Open Research Data Pilot²²), published in December 2013. According to this document, Horizon 2020 projects must make every effort to enable third parties to access, mine, exploit, reproduce and disseminate the data from their research (free of charge for users).

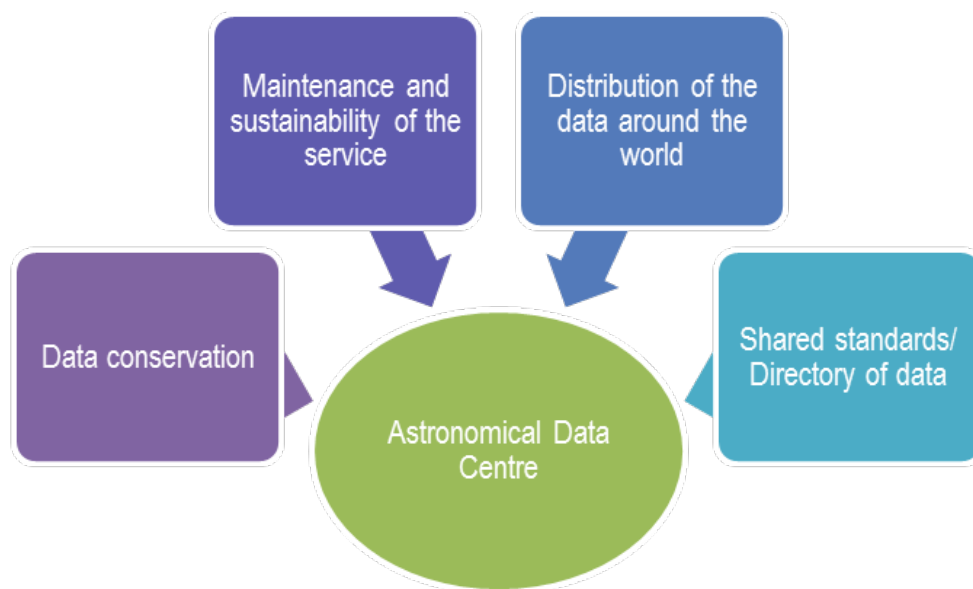
- 20 Horizon 2020 also contributes directly to the development of TDM by funding European projects on the subject. The European Commission invests in research on TDM with, for example, the H2020 call for projects GARRI-3-2014 “Scientific Information in the Digital Age: Text and Data Mining (TDM)”, issued in 2014. This call has given rise to several projects for the period 2015–2018, including:
- FutureTDM,²³ “The Future of Text and Data Mining”, which aims to spread the use of TDM in Europe. In particular, the project organises forums to obtain opinions and feedback concerning TDM from researchers, developers, publishers and other stakeholders.
 - OpenMinTeD,²⁴ “Open Mining INfrastructure for TExt and Data”, which aims to facilitate the use of TDM technologies on scientific publications by making software and platforms interoperable via standardisation.
- 21 The proportion of research projects on TDM among the projects funded by the European Commission has been calculated following a TDM analysis. Among more than 30,000 projects funded by the European Commission during the period 2007–2016 (FP7 and H2020), approximately 2.90% were related to TDM (n=885 projects).²⁵ Some projects (0.38%, n=115) deal specifically with TDM. They all have the terms “Text Mining”, “Data Mining” or “Text and Data Mining” in their descriptions. The other projects (2.53%, n=770) contain similar terms, such as “Big Data”, “Data Analysis” or “Machine Learning”. This analysis also highlighted a semantic shift, with the expression “Text and Data Mining” being used increasingly.
- 22 Horizon 2020 also proposes an ethical framework for research. Good research practice goes under the umbrella term of Responsible Research and Innovation (RRI). The goal is to take into account the potential impact of each innovation on society, anticipating future change and thinking in terms of sustainable development and sharing. In particular, personal data must be processed with full consideration of the right to privacy.



Data repositories in France: A few examples

The TDM framework in a pioneer discipline: Astronomy

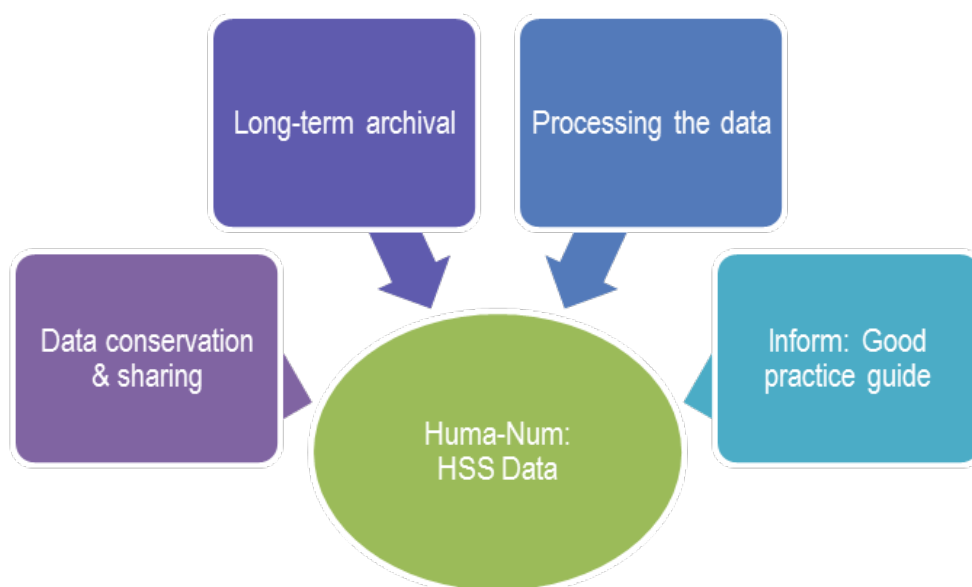
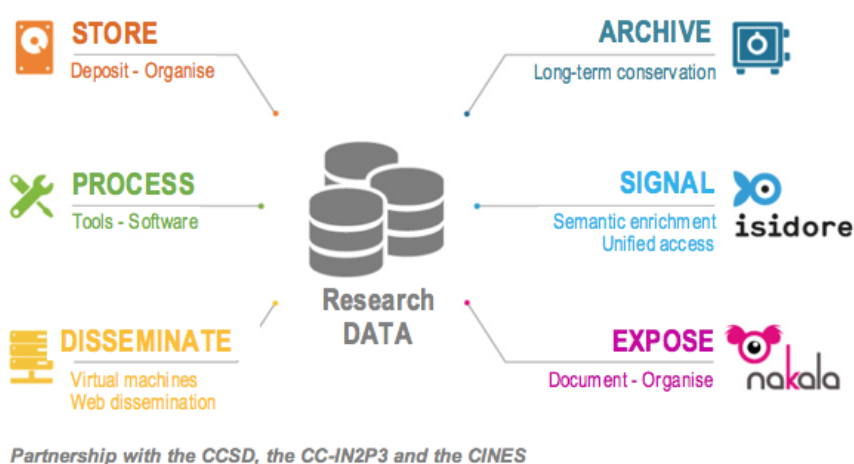
- 23 The abovementioned article “Préserver les données de la recherche à l’ère du Big Data” (Preserving research data in the age of Big Data) takes the example of the organisation of astronomy, a pioneer discipline in the conservation and sharing of data.
- 24 The article quotes Françoise Genova, a researcher at the CNRS National Institute of Sciences of the Universe and the Strasbourg Astronomical Data Centre (CDS), who states that:
- “To understand the physical phenomena operating in astronomy, we need to collect observations obtained by different techniques and to work from data obtained by other instruments and other teams.”
- “To respond to this need to exchange and preserve data, the international astronomical community is structured around the Virtual Observatory, a set of services that enables researchers to find useful information among all the astronomical data available to them through a large directory and shared standards.”²⁶
- 25 The Strasbourg Astronomical Data Centre (CDS) hosts the SIMBAD, the world astronomical reference database for the identification of astronomical objects. Its mission is to collect and exploit astronomical data and related information and make sure it is available all over the world. It is responsible for conserving data, maintaining the service and ensuring its long-term viability.



Huma-Num in the field of the human and social sciences

- 26 “Huma-Num is a very large research infrastructure (VLRI) aiming to help research in the human and social sciences successfully navigate the move to digital.
- 27 To fulfil this mission, the Huma-Num VLRI is built on a new type of organisation that involves mobilising both human and technological resources (by collective consultation and sustainable digital services, respectively) at the national and European levels, with the help of a large network of partners and operators.

- 28 The Huma-Num VLRI aims to help consortia of players from within the scientific communities to coordinate the reasoned and collective production of source corpora (scientific recommendations, best technological practices). It is also developing a single technological structure for the processing, preservation, access and interoperability of research data. This set-up consists of a grid of dedicated services, a single access platform (ISIDORE) and a procedure for long-term archival.
- 29 The Huma-Num VLRI also offers general good practice guides for researchers on the technology involved. It can provide expertise and training services on request. It participates in the DARIAH project on behalf of France as the National Coordinating Institution.
- 30 The Huma-Num VLRI is run by the Joint Service Unit 3598, involving the CNRS, Aix-Marseille University and the Campus Condorcet.”²⁷



The PREDON project

- 31 The PREDON project (for *PRE*servation de *DON*nées, or Preservation of Data), an offshoot of the CNRS's MASTODONS programme seeking to encourage cooperation between different disciplines around the concept of "Big Data", was set up by a small group of researchers from the IN2P3.²⁸ Its mission is to federate all French initiatives in the field of the preservation of scientific data. The project proposes a new approach based on the scientific, technical and organisational capacity of research units, international partnerships and major computing centres.
- 32 The PREDON group works closely with similar French and international initiatives, especially with the International Committee for Future Accelerators (ICFA) panel for the preservation of data in high energy physics.
- 33 In 2014, the PREDON Working Group produced a summary document ("Scientific Data Preservation 2014") briefly describing the contributions of the participants in the workshops it has organised. The document is divided into three parts corresponding to the complementary aspects of the preservation of scientific data: scientific potential, methodology and technologies.

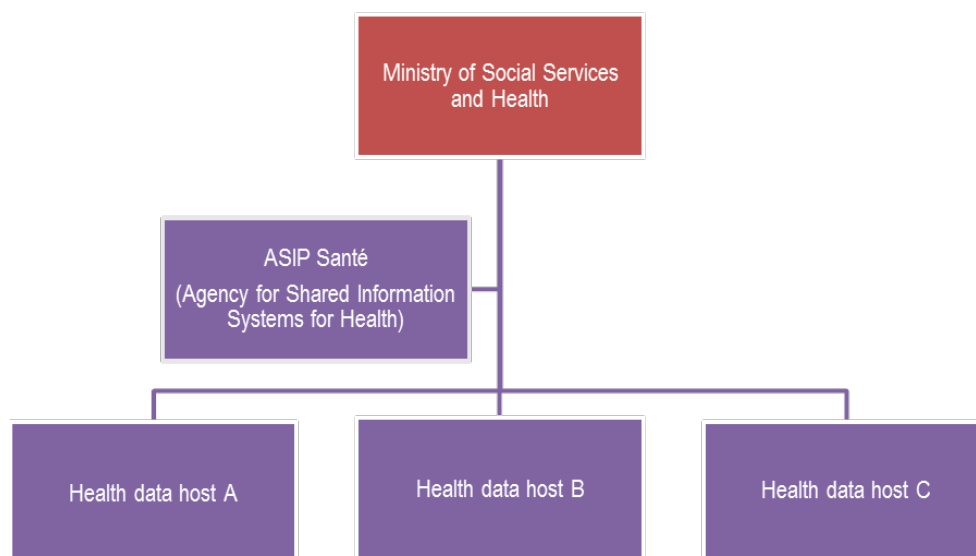
The expert contribution of the CINES

- 34 The Centre Informatique National de l'Enseignement Supérieur or CINES (National Informatics Centre for Higher Education) is a national public body under the supervision of the Ministry of Higher Education and Research. The CINES has three strategic national missions:
 - supercomputing,
 - long-term archival of digital data,
 - hosting computing platforms on a national scale.
- 35 As part of its second mission, the CINES is responsible for archiving the data and digital documents produced by the French higher education and research community, including scientific data from observation and experimentation, or resulting from calculation.
- 36 "It provides shareable, economic and customisable solutions for digital archiving in the medium and long term",²⁹ and "on the strength of its dual experience in supercomputing and long-term archiving, and also its excellent IT infrastructure", the CINES "assists the producers and managers of data with their archiving problems".³⁰
- 37 The main objective is to ensure the accessibility, integrity, readability and comprehensibility of the data for as long as necessary, by adapting the level of requirements necessary for the archiving as appropriate for the duration in question.
- 38 In this framework, the CINES lobbies the players on the market to ensure that their file formats will remain readable for many years. A team of engineers is engaged in a permanent race against obsolescence by ensuring that software and hardware platforms can continuously access data. The FACILE web service³¹ lists the range of formats currently supported by the CINES.

A parallel with the way the hosting of healthcare data is organised

39 To decide on the best way to organise the “preservation and communication” of files resulting from the processing of research data, lessons can be learned from the way the system for hosting healthcare data is organised in France.

40 This is organised as follows:



41 **Creation.** The ASIP Santé is a public interest group set up between the state, the CNAM (national health insurance fund) and the CNSA (national solidarity fund for autonomy).

³² In 2009 the public interest group for personal medical records (GIP-DMP) became the ASIP Santé (Agency for Shared Information Systems for Health).³³ The ASIP Santé is overseen by the Ministry of Health.³⁴

42 **Role.** The role of the ASIP Santé is to promote and supervise the development of shared systems in the areas of health and welfare sector.

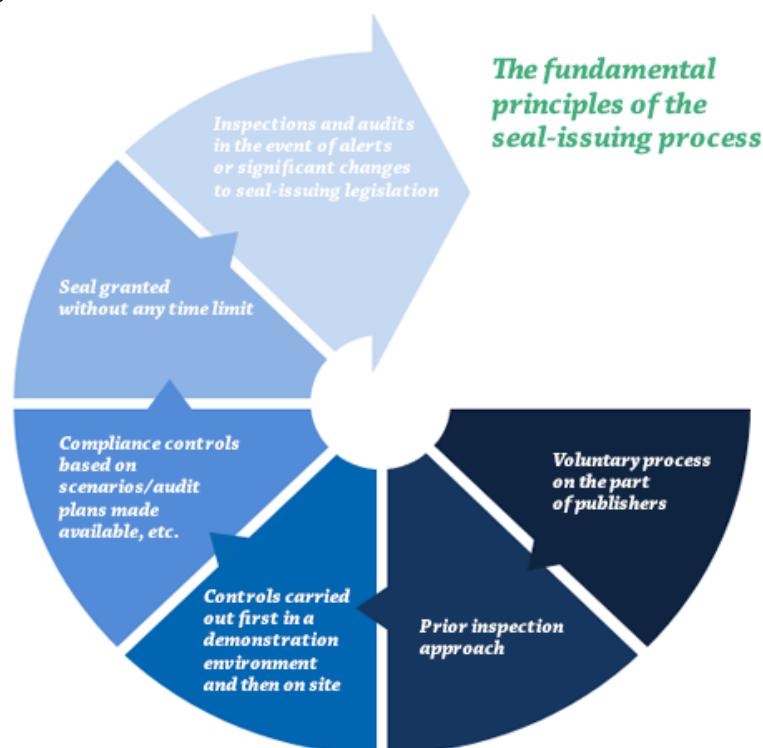
43 **The need for interoperability.** In order to facilitate and increase the exchanges between the different professionals in the areas of health and the welfare sector, interoperability must be achieved at the semantic, syntactic and technical levels.³⁵ Making information entirely digital requires “the definition of languages common to all the information systems that will manipulate them, so that there is no need to develop new languages every time two information systems need to exchange or share data”.

44 In compliance with its founding agreement, the ASIP Santé provides reference frameworks, standards, products or services contributing to the interoperability, security and use of health and telehealth information systems, and monitors their proper application. The Health Information Systems Interoperability Framework (HIS-IF) “sets the rules for a fully communicating health information system”.³⁶

45 **Accreditation.** A licence must be obtained for the hosting of health-related personal data, because of their sensitive nature.³⁷ The ASIP Santé is responsible for the pre-appraisal of requests for accreditation and has developed a reference framework for completing the application forms.³⁸ During this pre-appraisal, the ASIP Santé considers three aspects, detailed in the forms that the applicant has to complete:

- an ethical and legal aspect (including the actions taken to obtain the consent of the persons concerned and the conditions relating to requests for corrections of information);

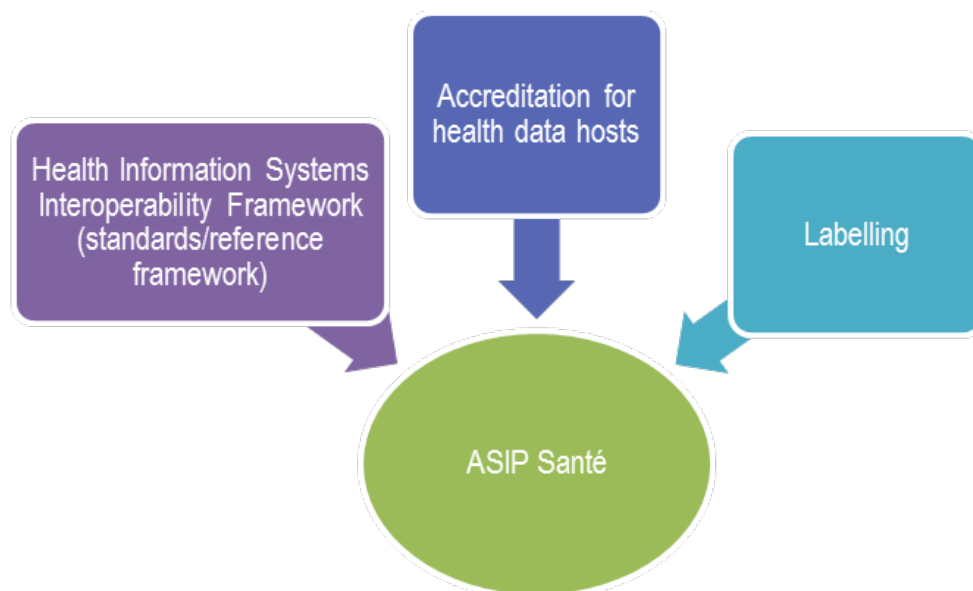
- a security and technical aspect (including measures to guarantee security of access and data transmission, measures to control access rights and the traceability of access and processing, the conditions for the verification of hacking attempts and unauthorised access, and the means for verifying the registers of accredited persons and how they are kept up to date);
 - an economic and financial aspect.
- 46 Accreditation is issued by the Ministry of Health for a duration of three years and is renewable.³⁹
- 47 **Seals.** The ASIP Santé also issues seals,⁴⁰ which are issued as explained in the following diagram:⁴¹



- 48 **Observations.** The creation of a governance structure for the ASIP Santé has been welcomed by professionals in the sector of health and social information systems: “At last there is a clearly identified governing body working in a spirit of consultation.”⁴² Operators and actors within the sector are invited to participate in the development of the HIS-IF by expressing their needs, to ensure the relevancy of the repositories established by the ASIP Santé.
- 49 The accreditation procedure for organisations wishing to host health data may seem onerous at first sight. Yet, on the one hand, it is necessary to ensure the security of the stored information in the light of its sensitive nature, and, on the other, the ASIP Santé attempts to make the process easier for applicants by making filing forms available. It is worth emphasising that this is simply a declaratory procedure, since only documents are examined. It should be noted that the Act for the Modernisation of the French Health System⁴³ potentially provides for an assessment of technical compliance by a certifying body accredited by the COFRAC, instead of the approval procedure. Accordingly, on 16 September 2016 the ASIP Santé proposed a certification reference

framework for consultation.⁴⁴ Operators in the sector are thus invited to submit comments and proposals. It draws attention to the following points:⁴⁵

- the absence of any possibility of a *a posteriori* control accorded by the Act to the General Inspectorate of Social Affairs;
- the need for the reference framework to define the resources to be deployed and not only the goals to be achieved;
- the consequent lack of visibility for applicants as to their chances of obtaining accreditation;
- the need for human and technical resources for examining these applications;
- the absence of outside audits by qualified auditors;
- the need to take into account technical developments and the availability of commercial hosting services.



- 50 An analysis of all the existing frameworks and actors and their mission and organisation enables us to sketch out and propose an organisational framework for the preservation and sharing of scientific research data for the particular purpose of TDM.

NOTES

1. <http://www.nactem.ac.uk/>
2. <http://openminted.eu/about/partners/univ-of-manchester-nactem/>
3. <http://www.nactem.ac.uk/requestaccess.php>
4. <http://www.nactem.ac.uk/customised.php>
5. <http://www.nactem.ac.uk/facta/>
6. <http://www.nactem.ac.uk/uima.php>
7. <http://www.nactem.ac.uk/research.php>
8. <http://www.nactem.ac.uk/news.php>
9. <http://www.nactem.ac.uk/>
10. https://www.nsf.gov/news/news_summ.jsp?cntn_id=103047

11. Section 804, *Implementing the Recommendations of the 9/11 Commission Act of 2007*, entitled *The Federal Agency Data Mining Reporting Act of 2007* (Data Mining Reporting Act).
12. <https://www.dni.gov/index.php/newsroom/reports-and-publications/94-reports-publications-2011/619-data-mining-report>
13. <https://www.icsu-wds.org/organization>
14. <http://www.icsu-wds.org/services/data-sharing-principles>
15. <http://www.icsu.org/what-we-do/interdisciplinary-bodies/wds/about>
16. <http://www.icsu-wds.org/files/wds-certification-summary-11-june-2012.pdf>
17. <http://www.icsu-wds.org/services/certification>
18. http://www.icsu-wds.org/files/WDS_Data_Sharing_Principles_2015.pdf
19. <https://www.rd-alliance.org/groups/interest-groups>.
20. European Commission, *Toward a modern, more European Copyright Framework*, Brussels, 9.12.2015, COM(2015) 626 final.
21. European Commission, *Guidelines on FAIR Data Management in Horizon 2020*, 26 July 2016 (FAIR = findable, accessible, interoperable and reusable).
22. <https://www.openaire.eu/opendatapilot>
23. <http://www.futuretdm.eu/>
24. <http://openminted.eu/>
25. All these data are taken from FutureTDM, Deliverable D4.1, *European Landscape of TDM: Applications Report*, May 2016, p. 38.
26. Guillaume Garvanèse, "Préserver les données de la recherche à l'ère du Big Data", *CNRS Le Journal*, 9 September 2016, <https://lejournel.cnrs.fr/articles/preserver-les-donnees-de-la-recherche-a-lere-du-big-data>
27. <http://www.huma-num.fr/la-tgir-en-bref>
28. Institut national de physique nucléaire et de physique des particules (French National Institute of Nuclear and Particle Physics).
29. <https://www.cines.fr/archivage/>
30. <https://www.cines.fr/archivage/typologies/donnees-scientifiques/>
31. <https://facile.cines.fr/>
32. <http://esante.gouv.fr/partenaires/nationaux/339>
33. Decree of 8 September 2009 approving the founding agreement for a public interest group: <https://www.legifrance.gouv.fr/eli/arrete/2009/9/8/SASC0917305A/jo> ; <http://esante.gouv.fr/asip-sante/espace-presse/communiqués-de-presse/le-gip-dmp-devient-l-asip-sante>
34. <http://esante.gouv.fr/actus/politique-publique/l-asip-sante-publie-un-etat-des-lieux-des-maitrises-d-ouvrage-regionales>
35. <http://esante.gouv.fr/services/referentiels/referentiels-d-interoperabilite/cadre-d-interoperabilite-des-systemes-d>
36. <http://esante.gouv.fr/services/referentiels/referentiels-d-interoperabilite/cadre-d-interoperabilite-des-systemes-d>
37. Article L. 1111-8 of the French Public Health Code: <https://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000021941353&cidTexte=LEGITEXT000006072665>
38. <http://esante.gouv.fr/services/reperes-juridiques/le-role-de-l-agence-des-systemes-d-information-partages-de-sante-dans-la>
39. Article R. 1111-15 of the French Public Health Code: https://www.legifrance.gouv.fr/affichCodeArticle.do?sessionId=60C3A828598DFE6EA728751C24A32AF5.tpdjo15v_2?cidTexte=LEGITEXT000006072665&idArticle=LEGIARTI000006908152&dateTexte=20130118&categorieLien=id#vig
40. <http://esante.gouv.fr/services/label-e-sante-logiciel-maisons-et-centres-de-sante>

41. <http://esante.gouv.fr/services/labellisation/editeurs-comment-obtenir-le-label#P1>
42. Annual Report 2009 of the ASIP Santé, p. 29,
http://esante.gouv.fr/sites/default/files/ASIP_Sante_Rapport_d_activite_2009.pdf
43. Act No. 2016-41 of 26 January 2016 for the modernisation of the French health system (1), Article 204, 5th c: “I. - Under the conditions laid down in Article 38 of the Constitution and no later than two years after the promulgation of this Act, the Government is authorised to take the measures by ministerial order for the improvement and simplification of the health system falling within the scope of the Act, to: ... c) Replace the accreditation provided for in the same Article L. 1111-8 by an assessment of technical compliance carried out by a certifying body accredited by the national accreditation body referred to in Article 137 of Act No. 2008-776 of 4 August 2008 on the modernisation of the economy, or by the competent body of another Member State of the European Union. This certification of compliance focuses in particular on the control of procedures, the organisation and the material and human resources as well as on the processes for the qualification of the hosted applications.”
44. <http://esante.gouv.fr/actus/services/agrement-des-hebergeurs-de-donnees-de-sante-publication-du-referentiel-de>
45. http://esante.gouv.fr/sites/default/files/asset/document/asip_sante_-_vue_densemble_referentiel_hds_-_v0.3.0.pdf

Proposals for applying the act

Introduction

- 1 The proposed Directive on Copyright in the Digital Single Market encourages the holders of rights and research organisations to jointly define best practices relating to the security and integrity of the networks and databases where the works are stored.
- 2 Furthermore, Article 38 of the Digital Republic Act stipulates that “A decree lays down the conditions under which the exploration of texts and data is implemented, as well as the terms for storage and communication of the files produced on conclusion of the research activities for which they were produced”;
- 3 The above discussion of the concept of TDM and the issues surrounding it, together with the observation of organisational models already existing in France, in the H2020 programme and in other countries, can help refine a proposal for a framework for the practice of TDM and, more broadly, Open Science.

Defining standards

A reference framework for interoperability specific to Open Science

- 1 To enable communication between different data and networks, methodologies and standards should be drawn up to ensure interoperability between the data produced by any field of science.
- 2 The network of curator entities must create a set of data repositories that are interoperable and speak the same language.
- 3 Similarly to the *Référentiel Général d'Interopérabilité* (General Interoperability Framework) or RGI, Open Science could acquire a standard to promote interoperability between digital data curation entities (version 2.0 of the RGI was officially promulgated by the Order of 20 April 2016, published in the *Journal officiel de la République française* (JORF) No. 0095 of 22 April 2016).
- 4 The RGI is a framework of recommendations referencing norms and standards that promote interoperability within the information systems of government bodies. These recommendations set out the objectives to be achieved to promote interoperability. They allow actors seeking to interact and therefore to develop the interoperability of their information systems to go beyond simple bilateral arrangements.
- 5 The RGI is defined in Order No. 2005-1516 of 8 December 2005 concerning electronic exchanges between users and government bodies and between government bodies. In Article 11 of this order, the “RGI lays down the technical rules to ensure the interoperability of information systems. In particular, it specifies which data directories, norms and standards must be used by government bodies.”

Certification or accreditation procedure

- 6 A procedure for the certification or accreditation of digital data curation entities could be defined.

- 7 This procedure for assessing digital data curation entities should ensure that they abide by their obligations, in particular as regards:
- data security;
 - compliance with standards and data formats;
 - compliance with obligations relating to technical infrastructures and those forming networks (maintenance of platforms for the collection and provision of data, maintenance of hubs for network management);
 - provision of good behaviour guides and an Ethics Charter.

Creating a network of digital data curation entities

Conceptualising a typical digital data curation entity

- 1 There are already bodies in France that curate digital data and make it available: the Strasbourg Astronomical Data Centre, and Huma-Num for data in the human and social sciences.
- 2 In order to conceptualise a typical digital data curation entity, we propose undertaking a precise analysis of these existing data repositories, especially in terms of structure, security, power, capacity, staff, features, use by the scientific community, and indicators.

Forming a network of digital data curation entities

- 3 We propose building on these existing structures to create similar structures in each of the scientific disciplines.
- 4 These structures must have the following features in common:
 - preservation of the discipline's scientific data;
 - distribution and availability of the discipline's scientific data via a platform;
 - enrichment of the data and the provision of processing tools;
 - maintenance in operational condition;
 - maintenance of a catalogue of the discipline's data.
- 5 In addition, in order to enable transdisciplinary and multidisciplinary research and analysis, priority should be given to the establishment of what has been called "networked science."¹ Although there are already some exchanges of theory and corpora, if data repositories can be created for each scientific discipline and networked together, this will facilitate the sharing of knowledge.
- 6 Finally, these digital data curation entities could be responsible for archiving data and ensuring its long-term preservation, with the support of the CINES.

- 7 This networking and the construction of computer hubs between the digital data curation entities of each scientific discipline implies the definition of standard formats and decisions on how to ensure the interoperability of the data.
-

NOTES

1. Florence Millerand, “La science en réseau. Les gestionnaires d’information ‘invisibles’ dans la production d’une base de données scientifiques” (Networked science: The ‘invisible’ information managers in the production of a scientific database), *Revue d’anthropologie des connaissances*, 6(1), 2012, pp. 163–190.

An ethical framework for TDM via an “Ethics Charter”

- 1 The results obtained by TDM must of course be interpreted and analysed critically. The risk exists, for example, of erroneously diagnosing causal links to explain the statistics obtained by TDM, leading to false correlations. TDM thus raises new methodological issues.
- 2 The integrity of research using TDM also depends on compliance with ethical rules. In particular, personal data must be processed with full consideration of the right to privacy. More broadly, the European Commission has decided to use the umbrella term Responsible Research and Innovation (RRI) to refer to good research practice. The goal is to take into account the potential impact on society of each innovation, anticipating future change, and thinking in terms of sustainable and shared development. For example, “Assessment of scientific and technological choices” (better known by the expression “technology assessment”) is an interactive scientific process whose objective is to contribute to the emergence of political and public opinions on new technologies. It is therefore necessary to predict the implications of TDM to ensure a beneficial impact on society.
- 3 There will consequently need to be serious discussions about the use of these tools. As the CNRS Ethics Committee has stated on this subject:

“It is not always possible to apply the basic principles of how to deal with personal data, such as informing people about the fate and use of the data, or obtaining their consent. The research process may require obtaining information without the knowledge of the person being investigated. In such cases, it would be necessary to stipulate the principles to be observed in the absence of consent, such as a commitment to inform this person after the event. The question of consent also arises when the research uses information resulting from data mining on social networks. These data, publicly available, are considered by the CNIL as personal data.”¹

- 4 The Ethics Committee has also argued that “confronted by this dynamic movement of data relayed by their supervisory authorities and by their community, researchers must:
- be aware of their individual, deontological² and ethical responsibilities, with respect to the community to which they belong;
 - be aware of the international undertakings of the institutions on which they depend;
 - participate in the definition of ethical principles specific to their discipline in the field of data sharing and of Big Data in general.”³
- 5 This need for ethical rules was expressed by the researchers themselves during the survey on the uses and needs of STI carried out by the CNRS in March 2015.
- 6 Setting up an “Ethics Charter for STI” laying down “Ethical principles to transcend the different categories of instruments and to affirm the goals of public research in a global context of Open Science” may partially satisfy this need.⁴
-

NOTES

1. COMETS, “The ethical issues of scientific data sharing”, 2015, p. 6.
2. “Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences”, Committee on Responsibilities of Authorship in the Biological Sciences, National Research Council, National Academy of Sciences.
3. Self-referral at the initiative of the COMETS, “Les enjeux éthiques du partage des données scientifiques” (The ethical issues of scientific data sharing), by the Data Sharing Group, 12 December 2014.
4. Results of the survey on the uses and needs of STI for the Research Units, carried out among CNRS Unit Directors – March 2015, p. 59.

Training researchers and research personnel on TDM practice

- 1 Researchers must also be trained in the correct use of text and data mining. TDM is currently mostly taught as part of the computer science curriculum, but it is clearly of use to research and its teaching should be extended. In order for all the sciences and every area of research to benefit, it should be taught in as many different curricula as possible.
- 2 TDM can make the workforce of any country more productive. All the techniques it involves require qualified, or indeed highly qualified, personnel.

Training for the different specialities

- 3 Concerning infrastructures, TDM will require far more data centres and consequently more technical personnel to develop them: in the construction trades, computing and technical maintenance specialities, and digital engineering.

The emergence of new trades and qualifications

- 4 Apart from the infrastructures, all the technical trades specific to TDM should also be developed. These professions, specialists in Big Data or database management, are the essential human capital for national economic productivity. The training they receive is therefore crucial.

The initial training of researchers

- 5 Currently, there are not enough specialised training courses concerning the knowledge resulting from TDM. At university level, the University of Lyon 2 offers a Master's in Data Mining,¹ and the University of Paris 8 a Master's in Big Data and Data Mining.² These master's courses involve training computer engineers competent in "deep learning", data analysis and management, visual representation of the results of text

mining, etc. In addition to the opportunities of becoming a researcher or lecturer-researcher in these IT disciplines, these training courses also provide entry into the professions of data scientist or study engineer, or statistical study director. The Universities of Paris 6, Nice Sophia-Antipolis and Paris 13 also offer specialised master's courses. However, these courses have a limited number of places (20 in the second-year master's course at the University of Paris 8, for example).

- 6 Data-analysis specialities are available as part of certain economics curricula, for instance at the ESSEC (Data Science and Business Analytics) or at the Toulouse Business School (Digital Intelligence and Marketing Intelligence).
- 7 However, the demand is increasing and it seems necessary to make these TDM courses available more widely.
- 8 Alongside these university or specialised courses, some engineering schools also offer specialist training. For example, the École Polytechnique (in Paris) set up a “data science” department in 2014, in partnership with companies like Keyrus, Orange and Thales.³ Télécom Paritech is moving in the same direction, as are other schools, but they are distributed too unevenly across France to provide training pathways as rich and popular as in the United States, Germany or the United Kingdom.
- 9 Other countries, especially in the English-speaking world, seem to be much more aware of the importance of TDM for the employment market. France has a lot of catching up to do in terms of training, despite the excellence of the existing structures. In Germany there are more master's in computational and data science or in data and knowledge engineering (in Dortmund and Berlin in particular). The United Kingdom and the United States nevertheless have the most advanced infrastructures and training centres in these areas. There are at least twenty different training courses in the United Kingdom, in Coventry, Leicester, London, Edinburgh, Lancaster, Nottingham, Oxford, Bristol, Norwich, Glasgow and Sheffield. Such wide geographical distribution, all over Britain, is echoed in its technical range: theoretical and generalist training (Data Science, Machine Learning), applied training (Applied Data Analytics), and specialised training (Business Intelligence Systems), etc.
- 10 In the United States, there are also many training courses in research centres specialising in TDM and Data Science. In particular, this is the case of the University of New York and its Center for Data Science, or the University of Harvard and its Berkman Center for Internet and Society. Several other universities also offer more advanced training in TDM than in France: Stanford, MIT, Northwestern University, Penn State, Indiana, University of Maryland, Carnegie Mellon, Berkeley, etc.
- 11 These training courses are essential for better preparing our human capital and also for fostering the increase of knowledge.

Occupational training and awareness-raising activity

- 12 It is also necessary to train and raise awareness among doctoral students, researchers and lecturer-researchers about depositing and sharing data and about data-processing techniques.
- 13 This awareness-raising could be achieved by drafting a good practices guide for Open Science, and by online tutorials and training (e-learning).

- 14 This overall training offer will need to be structured and coordinated with the organisations already working in France (Huma-Num, INIST, Persée, CCSD, URFIST, OST, etc.) to provide a national STI training programme for all scientific communities.
-

NOTES

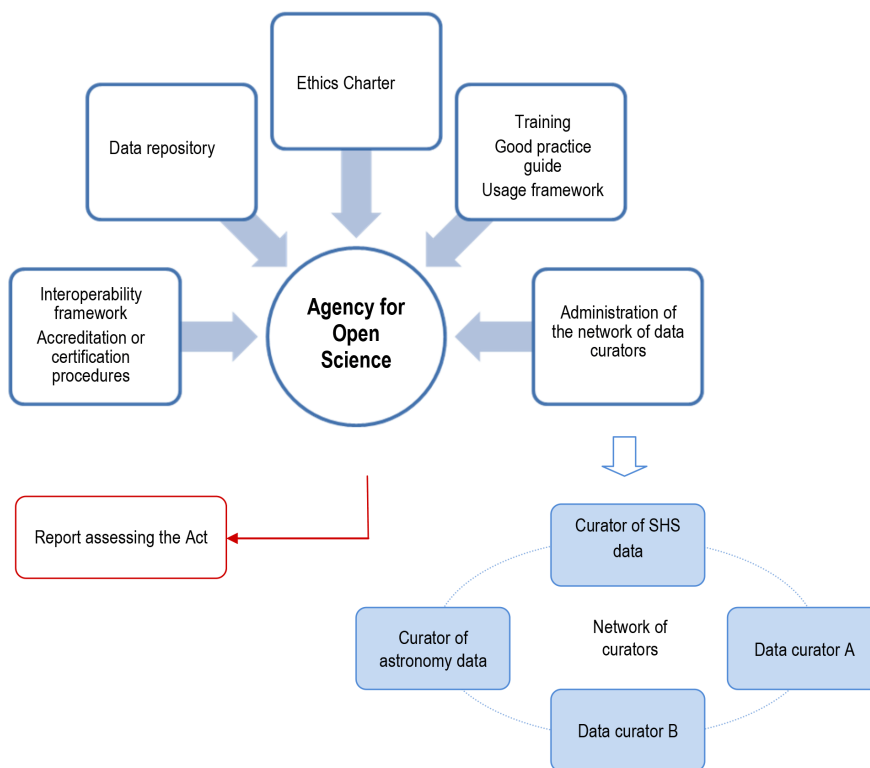
1. <http://master-datamining.univ-lyon2.fr/>
2. <http://www.univ-paris8.fr/Master-MIASHS-Big-Data-et-fouille-de-donnees>
3. <https://www.polytechnique.edu/fr/polytechnique-keyrus-orange-thales-creent-une-chaire-data-scientists>

Creation of a national agency for Open Science

- 1 In other countries, even in the English-speaking world, we found no evidence of national agencies or other authorities being set up to organise the collection and provision of scientific data. This lack of a central structure can be explained by the common-law tradition pertaining in the English-speaking world.
- 2 In view of the structure of the French administration and the multitude of missions necessary to implement Open Science, we propose the creation of a French national agency for Open Science. This agency could take the form of an Independent Administrative Authority. It could serve as a platform where scientific publishers and scientific communities could all express their views.
- 3 This national agency for Open Science could be given the following missions:
 - drafting and imposing the interoperability framework specific to scientific data;
 - creating a network of digital data curation entities for each scientific discipline, and ensuring that they accomplish their missions correctly;
 - training researchers;
 - drafting and imposing a good practice guide;
 - drafting and imposing an Ethics Charter;
 - drawing up a procedure for the certification or accreditation of digital data curation entities and monitoring its application;
 - maintaining a catalogue of scientific data.
- 4 This national agency would ensure the efficacy of Open Science and monitor and measure the corresponding practices. Indicators could be developed to monitor, for example, the sharing of data by discipline, the use of analysis tools, the emergence of new knowledge through the use of these tools, and the impact of these tools on the world of scientific publishing.
- 5 Such a national agency could contribute to the assessment of the progressive enforcement of the legislation by reporting on the impact of the provisions relating to Open Science in the Digital Republic Act and its implementing decrees.

Summary diagram of the overall framework

- 1 The following diagram summarises all the framework proposals laid out above:



Roadmap for the implementing decree for Article 38

Statement of the principles of Open Science

- 1 The Research Code lays down an institutional framework for the organisations that participate in scientific research in France, but there is no text defining the principles or values of the scientific community.
- 2 A specific body of law for science, established by consensus among scientists on public research, would involve incorporating the values of the scientific communities such as:
 - knowledge sharing;
 - open access to scientific data;
 - open processing of scientific data.
- 3 A text establishing the principles for Open Science would enable France to be a pioneer in this field.
- 4 This implementing decree could state these values in its Article 1 and propose definitions in Article 2 of the core terms used in the sector, including concepts such as:
 - text and data exploration and mining;
 - which versions of manuscripts may be subject to an embargo period;
 - the scope of the notion of “text and data included in or associated with scientific texts”.

The creation of “networked science”

- 5 The decree could establish the principle of a network architecture connecting digital data curation entities within each scientific discipline.
- 6 These digital data curation entities could be subject to an accreditation procedure.
- 7 They would be responsible for collecting, curating and providing access to scientific data, for ensuring the security of their information systems, and for ensuring compliance with the standards and interoperability framework necessary for network communications.

A model contract for the transfer of copyright between authors and publishers

- 8 In order to guarantee the rights of researchers regarding their published material and to mitigate the risks of contractual asymmetry, the implementing decree could lay down a model contract for transferring copyright for use in public research.
- 9 This contract would lay down the rules governing the relationship between the parties and protect researchers in their relationship with publishers. It would in particular ensure that there was no exclusive transfer, and guarantee the rights of researchers to:
 - authorise the immediate depositing and reproduction in open archives of the publication in the author's version, and in the publisher's version after expiry of an embargo period;
 - allow the immediate exploration of the content of the article using digital data-processing tools;
 - prevent all forms of private retention or appropriation of ownership concerning the content of the article.
- 10 This contract could be promulgated by decree and thus have a regulatory status that could be imposed on publishers for any scientific publication resulting from public research.

Creating an interoperability framework and standards

- 11 The implementing decree could state the principle of an interoperability framework specific to scientific data.
- 12 To set up networks between digital data curation entities within scientific disciplines requires communication between their information systems and compliance with data formats and standards.

An Ethics Charter for digital science

- 13 The CNRS Ethics Committee, supported by the CNRS Scientific Board, is in favour of the introduction of an Ethics Charter for digital science. This charter would define the values associated with accessing and sharing scientific data, as well as good practice for researchers, such as:
 - depositing scientific data on Open Science platforms;
 - ensuring that authorship is clearly mentioned.
- 14 An ethics committee would guarantee that this charter is respected, in particular by:
 - ensuring that its content is disseminated and understood;
 - ensuring that researchers are aware of the importance of ethics: "Researchers and all personnel involved in research must be trained to understand the ethical dimensions of data management, in particular in respect of privacy, intellectual property, and the quality and integrity of data. They must be informed as to the current status and evolution of the legal rules concerning responsible sharing of data used;"¹
 - issuing opinions with recommendations to clarify the good practice guidelines laid down in the charter.

Creation of a national agency for Open Science

- 15 The decree could provide for the creation of a national agency for Open Science, as a forum for different points of view, with responsibility for securing and monitoring practices, and expressing the opinions of the actors of Open Science.
- 16 Among its principal roles it would:
 - enforce compliance with the principles and values of Open Science;
 - ensure observance of the ethical rules defined in the Ethics Charter;
 - administer the network of digital data curation entities;
 - create an interoperability framework for data, setting standards and data formats;
 - establish the accreditation procedure for digital data curation entities;
 - draft a good practice guide;
 - maintain a catalogue of data;
 - monitor the training of researchers;
 - track technical developments and changing practices, and also manage the good practice guidelines;
 - play an advisory role;
 - propose changes to the existing legal framework in the light of the evolution of practices and needs.
- 17 The agency could also be responsible for writing a report on the impact of the principle of free access to scientific data on the scientific publishing market and on the circulation of ideas and scientific data.
- 18 France could very well propose an international convention for universal Open Science.

Creation of a European agency for Open Science

- 19 A European agency for Open Science could be created in line with the spirit of Article 3 of the proposed Directive on Copyright in the Digital Single Market introducing an exception to copyright and to the right of database producers.
- 20 The French agency could serve as a model for a European version, and France could be at the origin of this initiative.
- 21 After all, the problems facing Open Science are universal and call for a harmonised approach to the values of sharing and access to knowledge, at least at EU level.
- 22 Such a European agency could also be the vehicle of a discourse addressed to the Organisation for Economic Co-operation and Development (OECD), for example.

NOTES

1. Opinion issued by COMETS, “The ethical issues of scientific data sharing”, 7 June 2015.

Appendix

Analytical table comparing different TDM legislation

	France	European Union	United Kingdom	United States	Japan
Legal basis	Exception to copyright and the rights of database producers: right of digital copying and reproduction for purposes of TDM	Exception to copyright and the rights of database producers: right to reproduce or extract for purposes of TDM	Exception to copyright for purposes of “computational analysis”	Court ruling Fair use exception	Exception to copyright for purposes of “information analysis” for comparison, classification or statistical analysis
Scope	Mining of text and data included in or associated with scientific texts	TDM on works or other objects	Works and all associated data	Works	Information of all types
Beneficiary	/	Research organisations (the notion is broadly defined in Article 2 of the proposed Directive)	/	Parties to the dispute	/

Limit	TDM limited to the needs of scientific research/in a research context Non-commercial purpose Lawful source/ lawful access to the texts and data to be mined	TDM limited to the needs of scientific research Non-commercial purpose Lawful access to the material to be mined	Limited to the needs of research Lawful access Sufficient acknowledgement No copy may be transferred to any other person/No copy may be transferred or licensed by contract	Non-commercial purpose Nature of work protected by the copyright Portion of the work used No financial effect of the use	Not limited to public research or to non-commercial purposes
-------	--	--	--	---	--

References

Legal texts and laws on TDM

Digital Republic Act, including impact assessment of 9 December 2015¹.

EPRS, Tambiama Madiega, *EU Copyright Reform: Revisiting the Principle of Territoriality*, 2015.

HM Government [British Government], *Modernising Copyright: A Modern, Robust and Flexible Framework*, 2012.

Proposal for a Directive of the European Parliament and of the Council on Copyright in the Digital Single Market, COM(2016) 593 final, 14 September 2016.

Reda Report, 2015.

Report from the Expert Group of the European Commission, “Standardisation in the area of innovation and technological development, notably in the field of text and data mining”, April 2014.

Sirinelli Report for the CSPLA, *Rapport de la mission sur la révision de la directive 2001/29/CE sur l’harmonisation de certains aspects du droit d’auteur et des droits voisins dans la société de l’information* (Report of the mission on the revision of Directive 2001/29/EC on the harmonisation of certain aspects of copyright and neighbouring rights in the information society), December 2014.

Wolf & Partners, *Study on the Legal Framework of Text and Data Mining*, study for the European Commission, March 2014.

Institutional analyses of TDM

S. Ananiadou, *The National Centre for Text Mining: A Vision for the Future*, 2007;

APRIST (Association of STI Managers of Research Organisations), *Le TDM comme outil innovant de recherche scientifique* (TDM as an innovative tool for scientific research) (note on text and data mining).

J. Clark, *Text Mining and Scholarly Publishing*, study commissioned by the Publishing Research Consortium (PRC), Amsterdam, 2013.

COUPERIN and ADBU, *Mission relative au data mining : l’analyse de Couperin et de l’ADBU* (Mission relating to data mining: The analysis of Couperin and the ADBU), 2014.

CNRS White Paper, *Open Science in a Digital Republic*, 2016;

De wolf & Partners, *Study on the Legal Framework of Text and Data Mining*, 2014.

Guillaume Garvanese, “Préserver les données de la recherche à l’ère du Big Data” (Preserving research data in the age of Big Data), *CNRS Le Journal*, 9 September 2016, <https://lejournel.cnrs.fr/articles/preserver-les-donnees-de-la-recherche-a-lere-du-big-data>.

JISC, *The Value and Benefits of Text Mining: Digital Infrastructure Directions Report*, Doc#811, 2012.

J. Kelly, *The text and data mining copyright exception: Benefits and implications for UK higher education*, Jisc Publications, 2016.

NESTA – Alliance for Useful Evidence, *Using Research Evidence: A Practice Guide*, 2015.

OUTSELL, *Text and Data Mining: Technologies Under Construction*, 2016.

Science Europe, *Text and Data Mining and the Need for a Science-Friendly EU Copyright Reform*, Briefing Paper, 2015

UK IPO (Intellectual Property Office), *Impact Assessment (IA): Exception for Copying of Works for Use by Text and Data Analytics*, 2012.

Research on TDM

M. Borghi and S. Karapapa, *Copyright and Mass Digitization: A Cross-Jurisdictional Perspective*, Oxford University Press, 2013

W. Fan, L. Wallace, S. RICH and Z. ZHANG, *Tapping into the Power of Text Mining*, 2005;

U. Fayyad and R. Uthurusamy, “Data mining and knowledge discovery in databases: Introduction to the special issue”, *Communications of the ACM*, 39(11), 1999.

W. J. Frawley, G. Piatetsky-Shapiro and C. J. MATHEUS, *Knowledge Discovery in Databases: An Overview*, 1992;

C. Handke, L. Guibault and J. J. Vallbé, “Is Europe Falling Behind in Data Mining? Copyright’s Impact on Data Mining in Academic Research”, June 2015 ;

M. A. Hearst, *Untangling Text Data Mining*, School of Information Management & Systems, University of California, Berkeley, 1999.

M. A. Hearst, “What is Text Mining?”, SIMS, UC Berkeley, 17 October 2003.

F. Ibekwe-Sanjuan, *Fouille de textes : méthodes, outils et applications* (Text mining: Methods, tools and applications), coll. Systèmes d’information et organisations documentaires, Hermès, 2007, 352 p.

S. Jusoh and H. M. Alfawareh, *Techniques, Applications and Challenging Issue in Text Mining*, November 2012.

C. Laurent, J. Baudry ET AL., “Pourquoi s’intéresser à la notion d’« evidence-based policy » ?” (Why be interested in the concept of “evidence-based policy”?), *Revue Tiers Monde* 4(200), 2009, pp. 853–873.

F. Millerand, *La science en réseau. Les gestionnaires d’information « invisibles » dans la production d’une base de données scientifiques* (Networked science: The “invisible” information managers in the production of a scientific database), *Revue d’anthropologie des connaissances*, 6(1), 2012, pp. 163–190.

Y. Toussaint, “Extraction de connaissances à partir de textes structurés” (Extracting knowledge from structured texts), *Document Numérique*, 8(3), 2004, pp. 11–34.

NOTES

1. <http://www.assemblee-nationale.fr/14/projets/pl3318-ei.asp>

Acknowledgements

- 1 We would like to thank the contributors to these Study Review and Proposals for Implementing the Act:

Cabinet Alain Bensoussan

- Alain Bensoussan
- Sarah Lenoir

The staff of the CNRS DIST

- Renaud Fabre, Director of the CNRS DIST
- Laurence El Khouri, Deputy Director of the CNRS DIST
- Francis André, national expert in research data
- Stéphanie Dos Santos, Research Officer
- Quentin Messerschmidt-Mariet, Research Officer
- Marc Roux, Project Leader

Credits

- 2 English translation by Jackie Godfrey and John Kerr – Coup de Puce Expansion (27, allée Edouard Branly, F-31400 Toulouse – www.coupdepuce.com) under the supervision of Jean-François Nominé – INIST CNRS (2, allée du Parc de Brabois, F-54519 Vandoeuvre-lès-Nancy – www.inist.fr).