The legal and policy framework for scientific data sharing, mining and reuse

Melanie Dulong de Rosnay

The legal and policy framework for scientific data sharing, mining and reuse*

Mélanie Dulong de Rosnay

Mélanie Dulong de Rosnay is an Associate Research Professor at the Institute for Communication Sciences, CNRS - Paris Sorbonne - UPMC. She is also a Visiting Researcher in the Department of Media & Communications at the London School of Economics. She studies the interaction between law and code, the techno-legal infrastructure for knowledge commons, and peer to peer law.

Text and Data Mining, the automatic processing of large amounts of scientific articles and datasets, is an essential practice for contemporary researchers. Some publishers are challenging it as a lawful activity and the topic is being discussed during European copyright law reform process. In order to better understand the underlying debate and contribute to the policy discussion, this article first examines the legal status of data access and reuse and licensing policies. It then presents available options supporting the exercise of Text and Data Mining: publication under open licenses, open access legislations and a recognition of the legitimacy of the activity. For that purpose, the paper analyses the scientific rational for sharing and its legal and technical challenges and opportunities. In particular, it surveys existing open access and open data legislations and discusses implementation in European and Latin America jurisdictions. Framing Text and Data mining as an exception to copyright could be problematic as it de facto denies that this activity is part of a positive right to read and should not require additional permission nor licensing. It is crucial in licenses and legislations to provide a correct definition of what is Open Access, and to address the question of pre-existing copyright agreements. Also, providing implementation means and technical support is key. Otherwise, legislations could remain declarations of good principles if repositories are acting as empty shells.

Keywords: copyright law, Text and Data Mining, open access

* Revised version post peer-review was submitted in April 2014. This is the updated version from October 2015.

Dulong de Rosnay, Mélanie 2017. « The legal and policy framework for scientific data sharing, mining and reuse » [en ligne], *in* C. Mabi, J.-C Plantin, et L. Monnoyer-Smith (dir.), *Ouvrir, partager, réutiliser. Regards critiques sur les données numériques*, Paris, Éditions de la Maison des sciences de l'homme, disponible sur <http://books.openedition.org/editionsmsh/9082> doi : 10.4000/books.editionsmsh.9082

L'extraction de textes et de données (Text and Data Mining), le traitement automatique ou la fouille de grandes quantités d'articles scientifiques et de jeux de données, est une pratique essentielle pour les chercheurs et les chercheuses contemporains. Certains éditeurs et éditrices contestent la légitimité de cette activité et le sujet est discuté lors du processus de réforme du droit d'auteur en Europe. Afin de mieux comprendre le débat sous-jacent et de contribuer à la discussion politique, cet article examine d'abord le statut juridique de l'accès aux données et des politiques de réutilisation et d'octroi de licences. Il présente ensuite les options disponibles pour autoriser et faciliter l'exercice d'extraction des textes et des données : publication sous licences ouvertes, lois sur l'accès libre et reconnaissance de la légitimité de l'activité. À ces fins, l'article analyse la rationalité scientifique du partage et ses défis et opportunités juridiques et techniques. En particulier, il examine les lois existantes en matière d'accès ouvert et de données ouvertes et examine la mise en œuvre dans les juridictions européennes et latino-américaines. Le traitement de l'exploration de textes et de données en tant qu'exception au droit d'auteur pourrait être problématique puisqu'il équivaudrait à nier de fait que cette activité fait partie d'un droit positif de lecture, et ne devrait par conséquent pas nécessiter d'autorisation supplémentaire ou de licence. Il est crucial que les licences et les législations fournissent une définition correcte de ce qu'est l'accès libre et règlent la question des accords de droits d'auteur préexistants. En outre, la mise en œuvre des moyens et le soutien technique sont essentiels. Autrement, les législations pourraient rester des déclarations de bons principes si les dépôts servent de coquilles vides.

Mots-clés : droit d'auteur, fouille de texte et de données, accès ouvert

Introduction

Legal aspects of data sharing matter to at least three decision-making areas, all depending on access to publicly funded research: scientific and innovation policy; databases, publishing platforms, repository and data mining applications producers; public sector information and open data movements.

The topic of open data was discussed in the European Parliament with the vote in March 2013 of the *Horizon 2020* EU program for research and innovation, which contained a part on open access to publication and

scientific results. The automated processing of large amounts of scientific articles and databases deriving from Open Access and Open Data allows to detect connections which could not have been made manually. Called Text and Data Mining, this practice is contested by some right holders as infringing on their copyright. Text and Data Mining has been more specifically discussed since January 2015 in the report by MEP Julia Reda on the revision of the copyright directive as a future exception. The European Commission consultation *Licenses for Europe*, with stakeholders proposing to create a new exception for users, and others to create a new revenue stream for publishers, had revealed similar opposition to other copyright-related issues. At the same time, the UK made progress towards open access by developing the *Gateway to Research* portal, requiring Open Access for certain research outputs to be considered in evaluation policies, and introducing an exception for Text and Data Mining while Spain, Argentina, Italy, Germany and Peru voted laws to mandate Open Access.

Considering the scientific data ecosystem in its entirety gives the opportunity to study the question of scientific data, from its creation by researchers to its access and reuse by students, citizens, public bodies, NGOs and companies. Therefore, this paper will combine a presentation of the legal framework governing the creation and the usage of data, the policy options from all-rights-reserved to unlimited reuse, and the requirements of platforms and applications to process scientific data, perform queries, data mining, visualization or other analysis tasks without restrictions. Abovementioned examples from the European and Latin American countries moving forward Open Access to scientific publication and in some cases data will together illustrate tendencies and controversies around scientific data sharing and reuse policies.

As for methodology and definition of scope, the legal and policy framework is understood not only as the set of laws and contracts governing the access to and reuse of data (regulation by law), but also the opportunities and restrictions embedded in the technical architecture (regulation by technology) hosting the data. While the article focuses on scientific data, such analysis and conclusions are also applicable to public sector information and citizen data as they can also be used by researchers.

The legal framework for data

Data creation, access and reuse

The mere generation of data is not covered by most copyright-related legislation. Copyright provides an exclusive right on ideas, facts or data only when they are formalized, for instance under the form of an article. Even if raw data is in principle not to be protected, some jurisdictions recognized a specific right to compilations or databases.[1] This is the case in Europe, with the EC 2006 Database Directive[2] granting a *sui generis* right to the producer of a database, defined as "a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means." The Directive, which had to be transposed into Member States legislation, grants to the data producer, the entity responsible for the investment in time and resources, an exclusive right on access and reuse of data. According to this right, any person willing to access to and reuse the data will have to request the authorization of the database producer. Only non-substantial reuse may fall under the scope of exceptions to the Directive and be performed freely by potential users. Any update or new investment will lead to the renewal of the 15 year protection allowing a possible perpetuity of exclusivity in the case of maintenance of the database.

In the US however, collections of facts lacking creativity and originality[3] are outside of the scope of copyright and remain free to reuse for researchers, libraries, consumers and companies. Most other countries follow that model and do not grant protection to the database in addition to the content and its elements which may fall under copyright or not fulfill

1. For a more detailed overview of the legal status of research data mining regarding licensing, database and copyright law, see Lucie Guibault, "Licensing research data under open access condition," ***in*** D. Beldiman (ed.), *Information and Knowledge: 21st Century Challenges in Intellectual Property and Knowledge Governance*, Cheltenham, Edward Elgar, 2013, available at <http://www.ivir.nl/publicaties/download/Open_Research_Data.pdf>; Andres Guadamuz and Diane Cabell, "Data mining in UK higher education institutions: Law and policy", *Queen Mary Journal of Intellectual Property*, 4(1), 2014: 3–29, 2014, available at <http://www.elgaronline.com/view/journals/qmjip/4-1/qmjip.2014.01.01.xml>; and European Commission, *Standardisation in the area of innovation and technological development, notably in the field of Text and Data Mining, Report from the Expert Group*, Luxembourg, European Union, 2014, 80 p., available at <http://ec.europa.eu/research/innovation-union/pdf/TDM-report_from_the_expert_group-042014.pdf>.

2. Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, Official Journal L 077, March 27, 1996: 20–28.

3. *Feist Publications, Inc., v. Rural Telephone Service Co.*, 499 U.S. 340 (1991).

the requirements of intellectual creation. Countries recognizing a right to database makers including compilations of facts lacking creativity include Mexico[4] and Korea.[5] Treaty 1996 proposal has left the agenda of WIPO, the World Intellectual Property Organization, but in countries lacking database rights, legislations for unfair competition may reach the same effect.[6]

Specific status for scientific data

Scientific data and databases can also be populated by copyrightable items, such as photos, or notices drafted from observation results and comments. Metadata and underlying taxonomy and ontology structure will fall under the definition of a database, and are protected differently than the item they describe, a raw data which will not be protected by itself, or a photo which will be copyrightable but useless out of context and lacking metadata. "This discrepancy reveals an epistemological gap between copyright law and scientific effort conceptions of a creative or original effort, the threshold of protection."[7] In some cases, to add to complexity, scientific data can be considered as public sector information and/or as geographic or environmental data and therefore, if produced in Europe, submitted to additional Directives offering more possibilities to exclude documents held by research institutions[8] or to restrict access for reasons related to Intellectual Property Rights or the protection of endangered species.[9]

4. "Summary on existing legislation concerning intellectual property in non-original databases," text prepared by the World Intellectual Property Organization Secretariat for the Standing Committee on Copyright and Related Rights: Eighth Session (Geneva, November 4–8, 2002), document SCCR/8/3, September 23, 2002.

5. Survey led by Creative Commons among its affiliates: <http://wiki.creativecommons.org/4.0/Sui_generis_database_rights>.

6. Catherine Colston, "*Sui Generis* database right: Ripe for review?," *Journal of Information, Law & Technology*, 3, 2001, available at <https://www2.warwick.ac.uk/fac/soc/law/elj/jilt/2001_3/colston/>.

7. Melanie Dulong de Rosnay and Andrés Guadamuz, "Open access to biodiversity scientific data: A comparative study," paper presented at the 17th International Consortium on Applied Bioeconomy Research ICABR Conference on Innovation and the Policy for the Bioeconomy (Ravello [Italy], June 18–21, 2013).

8. Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information, OJ L 345, December 31, 2003: 90–96, available at <http://eur-lex.europa.eu/eli/dir/2003/98/oj>.

9. Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE), OJ L 108, April 25, 2007: 1–14, available at <http://data.europa.eu/eli/dir/2007/2/oj>; Directive 2003/4/EC of the European Parliament and of the Council of 28 January 2003 on public access to

Contractual and restrictive implementation

Legislation enacted by States is not the only legal instrument to govern the availability of data for access and reuse. Databases can also be regulated by private ordering as data producers have the possibility to apply a license, a contract, or terms of use to their database. Producers have therefore the freedom to reserve all rights on their databases, and disregard potential leeway or users' rights existing in their legislation which would have allowed researchers or just anyone to perform data mining on data they would have had legal access to. Access to data is not sufficient in an area of digital processing. Data can only be effectively used and reused if it can be mined. Data mining is understood as the process by which software is scanning and crossing data to detect patterns or other interesting feature or knowledge.[10]

The *Licensing for Europe* Text and data mining Working Group[11] at the European Commission has been following that direction. Indeed, right holders have been asking text and data mining to be submitted to re-licensing for an additional remuneration of texts to libraries, researchers or the public for that purpose. The assumption that re-licensing for text and data mining purposes of already licensed content led consumer, research and library organizations to express their disagreement and leave the consultation process.[12] They advocate clarifying that text and data mining can be undertaken for free by those who already benefit from a lawful access. The European Database Directive does indeed allow some legislation to treat content mining as an infringement or as a grey area.[13] The exception

environmental information and repealing Council Directive, OJ L 41, February 14, 2003: 26–32, available at <http://data.europa.eu/eli/dir/2003/4/oj>.

10. Usama Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth, "From data mining to knowledge discovery in databases," *AI Magazine*, 37(3), 1996, available at <http://dx.doi.org/10.1609/aimag.v17i3.1230>.

11. <http://ec.europa.eu/licences-for-europe-dialogue/node/7>.

12. Paul Keller, "Open letter regarding the Commission's stakeholder dialogue on text and data mining," *Communia blog*, February 27, 2013, available at <http://www.communia-association.org/2013/02/27/open-letter-regarding-the-commissions-stakeholder-dialogue-on-text-and-data-mining/> and Paul Keller, "Research sector, SMEs, civil society groups and open access publishers withdraw from Licences for Europe dialogue on text and data mining," *Communia blog*, May 25, 2013, available at <http://www.communia-association.org/2013/05/25/research-sector-smes-civil-society-groups-and-open-access-publishers-withdraw-from-licences-for-europe-dialogue-on-text-and-data-mining/>.

13. Andres Guadamuz and Diane Cabell, "Data mining in UK higher education institutions...," *op .cit*.

in the database directive article 6 to perform non-substantial extraction and reuse can be limited to non-substantial reuse and granted only "where there is use for the sole purpose of illustration for teaching or scientific research, as long as the source is indicated and to the extent justified by the non-commercial purpose to be achieved." Anyway, Database Directive article 5[14] allows right holders to maintain exclusivity for data mining to the extent that can be considered as a "repeated and systematic extraction."

Open data policy

The rational for sharing

Open Access to data is complementary to Open Access for publications. Authors and their institutions benefit from Open Access to data because they will be able to extract, parse and analyze data collected by others and potentially process much more information than they would have been able to produce themselves or for which they would have the time and resource to request permission and eventually pay royalties. Funding agencies and governments will avoid duplication of funding for the collection of similar datasets. Companies and NGOs can develop services and applications from the same data, and citizens can increase their scientific knowledge and education. Open Access has economic, cultural and democratic benefits, but the main scientific reason to share data is to allow more researchers to check findings, correct possible mistakes, edit and update knowledge. Open Access to data as a complement to Open Access to articles they are associated with will allow other researchers to reproduce the results.

Besides result reproducibility, Open Access also contributes to data archiving. According to a study on the availability of research data based on 516 studies,[15] chances to find the dataset fall by 17% every year from the third year after publication. Most data related to studies of the 1990s would be permanently lost, due to change of authors contact information and obsolescence of storage, making it impossible to produce long-term or comparative studies. Archiving and preservation would be better performed

14. "The repeated and systematic extraction and/or re-utilization of insubstantial parts of the contents of the database implying acts which conflict with a normal exploitation of that database or which unreasonably prejudice the legitimate interests of the maker of the database shall not be permitted."

15. Timothy H. Vines *et al.*, "The availability of research data declines rapidly with article age," *Current Biology*, 24(1), 2014: 94–97, available at <http://dx.doi.org/10.1016/j.cub.2013.11.014>.

at institutional or publishing levels than by the researchers themselves. But guidelines are needed to ensure that data will be reusable, avoiding scientists to be "piling their data in fairly unsearchable data repositories because they are forced to by journal editors or funders."[16]

Solutions and recommendations for sharing

Open licenses

In order to circumvent possible legislation granting an exclusive right to control the extraction and the reuse of data, producers may choose to apply terms of use to their database to indicate that they will renounce such rights. Open Access tools such as Creative Commons licensing suite will allow marking websites with a set of permissions. But the greatest scope of acts can only be performed if no rights are attached to the data, and placing them into the public domain will fulfill these conditions and allow interoperability.[17]

Open Access has been defined by the Budapest Open Access Initiative (BOAI) as the free availability of and unrestricted access to research results, meaning without financial, legal or technical barriers. The revised BOAI recommendations state that research results should therefore be made available without payment, without contractual, legal, or licensing restrictions on use or reuse other than integrity of the data and attribution of the author or contributors. "Libre Open Access" (which combines free access as well as liberal open licensing will be achieved for publications preferably under a Creative Commons Attribution license[18] or equivalent and for research data with a CCo[19] or equivalent.[20]

As for technical availability, open data should be offered with no technical restrictions which might prevent data mining and any other automatic

16. Rene J. Melis *et al.*, "Sharing of research data," *The Lancet*, 378(9808), 2011: 1995, available at <http://dx.doi.org/10.1016/S0140-6736(11)61870-9>.
17. Science Commons, "Science Commons Protocol for implementing open access data", 2011, available at <http://sciencecommons.org/projects/publishing/open-access-data-protocol/>.
18. Such as the Creative Commons Attribution 3.0 unported <http://creativecommons.org/licenses/by/3.0/>.
19. CC0 1.0 Universal (CC0 1.0) Public Domain Dedication <http://creativecommons.org/publicdomain/zero/1.0/>.
20. Melanie Dulong de Rosnay, "Communia association Position on EC Horizon 2020 open access policy," *Communia blog*, November 20, 2012, available at <http://www.communia-association.org/2012/11/20/position-on-ec-horizon-2020-open-access-policy/>.

processing to download, analyze, filter, index, search, connect and map datasets in order to detect patterns and results leading to scientific discovery or correlation of facts. It appears that most data, even in fields claiming to be Open Access and practice an Open Data policy, remain locked behind legal or technical barriers. A study led in 2008 by the author on a set of 200 databases of life science claiming to be in the public domain revealed that less than 20% were both legally and technically clearly available for access and reuse.[21] A more recent research published in 2013 analyzed 11,000 datasets of the Global Biodiversity Information Facility (GBIF),[22] a model institution for the collection and sharing of biodiversity data, showing that only 10% of them were carrying a license and only 1% an Open Data license. It could be that datasets could be reused even in the absence of a Public Domain statement, but the presence of a standard Open Data license is making it easier for the reuser to assume that any action, including data mining, can be performed on the data without having to analyze applicable law or check and understand possibly contradictory or unclear terms of use.

Open Access legislation

The contractual solution is based on voluntary contributions. It requires authors to make the decision to use an Open Data license, or institutions or databases to include in the terms of contributions that authors agree to deposit data under such a license. Relying on voluntary efforts is not a complete solution, and ends up in fragmenting scientific data, because some will be all-rights-reserved, some will be in the public domain, and some will be under possible incompatible terms of use, making it impossible for researchers to mix different sources without asking a lawyer to try to clear rights, or exposing them to possible legal risks if right holders found out, for instance in a publication, that they had reused a database without authorization and decided to sue. Together with the development

21. Melanie Dulong de Rosnay, "Check your data freedom: A taxonomy to assess life science database openness," *Nature Preceedings*, July, 2008, available at <http://sciencecommons.org/wp-content/uploads/npre20082083-1.pdf> and Melanie Dulong de Rosnay, "From free culture to open data: Technical requirements for access and authorship," *in* D. Bourcier, P. Casanovas, M. Dulong de Rosnay and C. Maracke (eds.), *Intelligent Multimedia: Managing Creative Works in a Digital World*, Florence, European Press Academic Publishing, 2010: 47–66.

22. Peter Desmet, "Analyzing the licenses of all 11,000+ GBIF registered datasets" [online], November 22, 2013, available at <http://peterdesmet.com/posts/analyzing-gbif-data-licenses.html>.

of accompanying measures providing effective support and incitation, the best way to ensure data can be reused by researchers is to go beyond contractual solutions and adopt legislation which would be applicable to all. Although there is so far no open data legislation in the world requiring authors to share their data, this section will present efforts which are going in that direction. Mandates can come from different sources: the scientific institution, the funding institution, a recommendation or a law enacted by the state or the European Union.

Open Access institutional mandates require researchers to make the final drafts of their publications available in a repository. Many universities[23] and research funding institutions[24] are developing such policies.[25] So far, they cover scientific articles, but not the underlying data. The perspective of funding mandates for research data is announced with the Open Data pilot of the European Commission Horizon 2020 published in December 2013.[26] "'Research data' refers to information, in particular facts or numbers, collected to be examined and considered and as a basis for reasoning, discussion, or calculation. In a research context, examples of data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images," therefore not only data and databases but also copyrightable elements such as text and images. Metadata associated with the research data and describing it are also included. It is not a mandate, but an experimentation encouraging the deposit of underlying research data, while there is an obligation to deposit the article. Underlying data are defined as the data necessary to validate the results presented in the scientific publications, including the metadata which it should be possible to access, mine, exploit, reproduce and disseminate free of charge. A suggested way to ensure this is to attach a Creative

23. Including Harvard University, Massachusetts Institute of Technology, University College London, Queensland University of Technology, University of Minho, University of Liege and ETH Zürich.

24. Such as National Institutes of Health, Research Councils UK, National Fund for Scientific Research, Wellcome Trust and European Research Council. For more information of the NIH policy, see Michael W. Carroll, "Complying with the National Institutes of Health Public Access Policy: Copyright considerations and options," A joint SPARC/Science Commons/ARL White Paper, February, 2008, available at <http://sparc.arl.org/resources/papers-guides/nih-copyright>.

25. They are listed at <http://roarmap.eprints.org/>, see also <http://aoasg.org.au/statements-on-oa-in-australia-the-world/>.

26. European Commission, "Guidelines on open access to scientific publications and research data in Horizon 2020," Version 1.0, December 11, 2013, available at <http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf>.

Commons license (CC BY or CC0 tool) to the data deposited. The inclusion of the activity of *mining* has an unfortunate side effect as one can assume that it had to be included in the list of possible actions because it is part of the right holders' exclusive rights. Therefore, a legalist interpretation could be that the European Commission acknowledges that data mining is not always an activity outside of the scope of copyright.

Opting out is possible in many cases, some of which can be subjected to rather broad interpretations, possibly defeating the purpose of the pilot (for confidentiality or security reasons, for personal data, if there is an obligation to protect results if they can be commercially exploited, if the principal objective of the project is jeopardized, or also for any other legitimate reason). Grantees will be asked to produce a *Data Management Plan* explaining which data are concerned and how they will be collected, shared and archived. While these constitute positive accompanying steps of 20% of the research funded under Horizon 2020 scheme and have been adopted after tough negotiations and opposition of many stakeholders fearing this would "interfere with the decision to exploit research results commercially, e.g. through patenting"[27] (one has to choose patenting or publishing as a first step), they are not sufficient to ensure that all funded data can be accessible and reused.

Besides data mandates or encouragements by institutions funding the research, recommendations and binding legislation can also be enacted by states. The European Union published several recommendations to support open data.[28] Policy recommendations to reform European and Member-States copyright legislation include proposals to revoke the database directive and to include content and data mining in the list of exceptions to exclusive rights.[29] After Spain in 2011, Argentina, Italy, Germany and Peru

27. *Ibid.*

28. European Commission High Level Group on Scientific Data, "Riding the Wave: How Europe can gain from the rising tide of scientific data," Final Report to the European Commission, October, 2010, available at <http://ec.europa.eu/information_society/newsroom/cf/itemlong-detail.cfm?item_id=6204>.
European Commission, Recommendation 2012/417/EU on Access to and preservation of scientific information, OJ L 194, July 21, 2012: 39–43, available at <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2012:194:0039:0043:EN:PDF>.

29. Andres Guadamuz and Diane Cabell, "Data mining in UK higher education institutions…," ***op.*** *cit.*; House of Commons Business, Innovation and Skills Committee, "The hargreaves review of intellectual property: Where next?," First Report of Session 2012–13, June 21, 2012, available at <http://www.publications.parliament.uk/pa/cm201213/cmselect/cmbis/367/367.pdf>.

voted in 2013 legislation mandating open access, while the UK addressed Text and Data Mining in 2014. They contain restrictions, and in some cases no implementation issues, but as they are new legislations, there is room for improvement and extension.

In Spain, the June 2011 National Law of Science[30] established a self-archiving requirement not later than 12 months after publishing. Researchers primarily funded by public institutions were expected to follow it from December 2011. However, it had never been applied in any project call until November 2013.[31] The law contains a final article potentially canceling the effects of this Open Access archiving mandate. Indeed, it is without prejudice of the agreements which can have transferred to third parties the rights on the publications, typically the publishers, or when the results are susceptible of protection. Data are not addressed.

The Peruvian legislation[32] adopted in June 2013 also created a central national repository for Open Access to publications, but also data and statistics. The information should be in Open Access, free to read, reuse, mine and all necessary acts, but for non-commercial purposes, which excludes commercial users, and with respect to copyright law, which leaves it unclear whether authors may deposit their work. In the latter case, metadata should still be deposited.

The Argentine legislation[33] of November 2013 requires public research institutions to develop repositories, and publicly funded research to be made available in Open Access repositories within 6 months after publication for the article, and 5 years after collection for the primary data so that other researchers might reuse them. There are exceptions in the case of intellectual property, prior agreements with third parties, confidentiality. The Ministry is expected to provide technical assistance and technical support and institutions which would do not comply will risk losing financial support.

30. Ley 14/2011, de 1 de junio, de la Ciencia, la Tecnología y la Innovación, article 37.3, available at <http://noticias.juridicas.com/base_datos/Admin/l14-2011.html>.

31. Ignasi Labastida, Responsable de la Oficina de Difusión del Conocimeinto de la Universidad de Barcelona, Creative Commons Europe mailing list, December 17, 2013.

32. Peru. Ley Nº 30035 que regula el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto fue publicada, available at <http://etd2012.blogspot.com/2013/06/ley-n-30035-que-regula-el-repositorio.html>.

33. Argentina. Ley 26899: Creación de Repositorios Digitales Institucionales de Acceso Abierto, Propios o Compartidos, available at <http://repositorios.mincyt.gob.ar/recursos.php>.

The German law[34] of October 2013 provides a mandate of self-archiving for non-commercial purposes of the author's final version of articles appearing in journals published at least twice a year (excluding all other formats) and at least 50% publicly funded, and declares contradictory publishers' agreements void. This last provision is good, but may apply to national publishers only.[35] The embargo is for a maximum of 12 months after publication. Data are not addressed.

The Italian law[36] of October 2013 also only targets articles (as opposed to books or other formats) which are publicly available in journals published at least twice a year and at least 50% publicly funded. They must be deposited in a non-commercial institutional or disciplinary repository within 18 months after first publication for scientific, technical, and medical disciplines and 24 months for humanities and social sciences, which is longer than acceptable recommendations by the Open Access scientific community. It leaves the implementation to institutions, and does not address the copyright question nor define Open Access. Data are not addressed.

The UK law of 2014[37] is the only example of an exception to copyright for Research, private study and text and data analysis for non-commercial research. However, framing text and data mining as an exception to private entitlement by default could be problematic as it de facto denies that this activity is part of a positive right to read and should not require additional permission nor licensing.

It is crucial in these laws to provide a correct definition of the scope of the research results covered, of what Open Access is, and address the question of pre-existing copyright agreements and confidentiality. Also, providing implementation means and technical support is key in these laws.

34. Gesetz zur Nutzung verwaister und vergriffener Werke und einer weiteren Änderung des Urheberrechtsgesetzes vom 1.10.2013, BGBl I 2013, 3728, available at <http://www.rechtliches.de/info_UrhG.html>.

35. Valentina Moscon, "Open access to scientific articles: Comparing Italian with German law," Kluwer Copyright blog, December 3, 2013, available at <http://kluwercopyrightblog.com/2013/12/03/open-access-to-scientific-articles-comparing-italian-with-german-law/>.

36. Legge 7 ottobre 2013, n. 112 Conversione in legge, con modificazioni, del decreto-legge 8 agosto 2013, n. 91, recante disposizioni urgenti per la tutela, la valorizzazione e il rilancio dei beni e delle attivita' culturali e del turismo. (13G00158) (GU n. 236 del 8-10-2013), available at <http://www.lexitalia.it/leggi/2013-112.htm>.

37. Section 29A and Schedule 2(2)1D of the UK Copyright Designs and Patents Act 1988, available at <http://www.legislation.gov.uk/uksi/2014/1372/regulation/3/made>.

Otherwise, they can remain declarations of good principles supported by repositories acting as empty shells.

The proposed revision of the 2001 European Union Copyright Directive (the 2015 Julia Reda report[38]) suggests to "allow the automatical analysis of large bodies of text and data (text & data mining)." A clarification that "lawful access to data includes the right to mine it through automated analytical techniques" as suggested by MEP Julia Reda would not suffer the argumentative drawback opened by framing the right of Text and Data Mining as an exception instead of a lawful use part of copyright. The compromise amendment voted on 17 June 2015 is unclear on the solution which will be chosen as it "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, provided that permission to read the work has been acquired." Nevertheless, it is a clear support to authorizing this activity regardless of the vehicle.

Technical platforms for Open Data

Technical platforms for Open Data are being developed by the EC, in the Netherlands and in the UK. The EC funded the development of the platform OpenAIRE[39] to host articles and datasets produced by its FP7 and Horizon 2020-funded projects. In the Netherlands, a data center[40] and Data Archiving and Networked Services[41] have been available since May 2013 for the deposit and permanent archival of underlying research data while some universities have developed another open source repository for short-term archiving by researchers themselves, the Dutch Dataverse Network.[42]

Gateway to research is a UK portal intended to provide information on all research funded in the UK. It contains data about projects, but not data from projects. It may however be including links to Open Access repositories and data catalogs where they exist. Technical API seem efficient and open, open licensing is addressed with an Open Government Licence v2.0.

38. Report on the implementation of Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society (2014/2256(INI)), Committee on Legal Affairs, Rapporteur: Julia Reda, June 24, 2015, available at http://www.europarl.europa.eu/sides/getDoc.do?type=REPORT&reference=A8-2015-0209&language=EN.

39. <https://www.openaire.eu/>.

40. <http://data.3tu.nl/repository/>.

41. <https://dans.knaw.nl/nl>.

42. <https://www.dataverse.nl/dvn/>.

There is no obligation to deposit data, but rather a declaration of Common Principles on Data Policy: "Publicly funded research data […] should be made openly available with as few restrictions as possible in a timely and responsible manner that does not harm intellectual property." Metadata, "legal, ethical and commercial constraints on release" and "a limited period of privileged use" are being considered and data sources acknowledged.

Other private initiatives exist to host research data, mostly repositories at the publishers' level. Journals can host data in content management systems linked with publications, and require authors to deposit underlying data and code in order to assess submissions' validity and quality during the publication submission process.[43] The evolution of this procedure has been studied for the computer science discipline towards result reproducibility, as more and more journals provide repositories for data and/or mandate the deposit of underlying data at the same time as the submission of the article.[44] The Joint Data Archiving Policy[45] requires as a condition to be published in several evolution journals including Nature and PLoS to deposit underlying data in a repository. Data repositories[46] and the publication of data papers[47] are being developed in other disciplines such as biodiversity studies, in order to recognize the contribution to databases and not only the publication of scientific papers. Data citation protocols are expected to provide an incentive for authors to share their data and for reusers to attribute them correctly and seamlessly.

43. <http://www.communia-association.org/2012/11/20/position-on-ec-horizon-2020-open-access-policy/>.

44. Victoria Stodden, Peixuan Guo and Zhaokun Ma, "How journals are adopting open data and code policies," paper submitted to The First Global Thematic IASC Conference on the Knowledge Commons: Governing Pooled Knowledge Resources (Louvain-la-Neuve, Belgium, September 12, 2012), available at <http://dlc.dlib.indiana.edu/dlc/handle/10535/9584>.

45. <http://datadryad.org/pages/jdap>.

46. <http://www.figshare.com>.

47. In the biodiversity community: Vishwas S. Chavan and Peter Ingwersen, "Towards a data publishing framework for primary biodiversity data: Challenges and potentials for the biodiversity informatics community," *BMC Bioinformatics*, 10, Suppl. 14, 2009: S2, available at <http://www.biomedcentral.com/1471-2105/10/S14/S2>; Vishwas S. Chavan and Lyubomir Penev, "The data paper: A mechanism to incentivize data publishing in biodiversity science," *BMC Bioinformatics*, 12, Suppl. 15, 2011: S2 available at <http://www.biomedcentral.com/1471-2105/12/S15/S2>.

Big data and privacy: the risks of sharing

In the context of citizen science and open data sharing quantified self-practices, some users knowingly and voluntarily share their own data, on health or on other topics. The aggregation and mining of data contributed by the users themselves create a risk of reidentification,[48] from correlation to profile deduction, making privacy and confidentiality difficult to enforce legally. Contextualized privacy solutions[49] and consent protocols[50] are being developed. But the risk of exclusion, for instance of insurance companies, remains. The 2012 project of European regulation on data protection foresees that information given to citizens on the processing of their personal data should be transparent and in clear language in order to guarantee an informed consent to share within a specific context; requirements on data portability are also planned.[51]

Conclusion

The discrepancies between the techno-legal framework and the requirements of researchers' applications to process data, perform queries, mining, visualization or other analysis tasks without restriction indicate points of frictions which should be solved. The framework and opportunities for data sharing show that the legal and policy measures requiring the deposit of data must be accompanied by a technical infrastructure to host research data. In the years following the first experiments, it is likely that copyright and technical obstacles to data sharing will have been corrected. Most important current issues of Text and Data Mining identified by the author

48. Kate Crawford, "The hidden biases in big data," Harvard Business Review, April 01, 2013, available at <https://hbr.org/2013/04/the-hidden-biases-in-big-data>; Latanya Sweeney, Akua Abu and Julia Winn, "Identifying participants in the personal genome project by name," Data Privacy Lab, IQSS, Harvard University, April 04, 2013, White paper, available at <http://dataprivacylab.org/projects/pgp/1021-1.pdf>.

49. Helen Nissenbaum, "A contextual approach to privacy online," *Daedalus*, 140(4), Fall 2011: 32–48, available at <http://www.amacad.org/publications/daedalus/11_fall_nissenbaum.pdf>.

50. For health data, see Consent to research at <http://weconsent.us/>, for web data, see the Algopol project protocol to collect Facebook data: Irène Bastard *et al.*, "Travail et travailleurs de la donnée," *InternetActu.net*, December 13, 2013, available at <http://www.internetactu.net/2013/12/13/travail-et-travailleurs-de-la-donnee/>.

51. Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), Brussels, 25.1.2012, COM(2012) 11 final, available at <http://ec.europa.eu/justice/newsroom/data-protection/news/120125_en.htm>.

in a response to the Public Consultation on the review of the EU copyright rules[52] are attribution, non-commercial and share-alike licensing requirements, the lack of definition of data, the framing of Text and Data Mining as an exception instead of a right and technical restrictions.

Some of the ethical risks of data sharing have been identified by legislation promoting or mandating open data, except when confidential or personal data are concerned. The risks of these exceptions to open access principles are the usage of intellectual property or confidentiality reasons without more details, leaving room for too much interpretation and legal insecurity. A chilling effect can also be caused by an over extensive interpretation of confidentiality, causing an impossibility to take advantage of the knowledge to be deduced from big data. Legal solutions to preserve personal rights against the collection and processing of their own data could be the extension of moral rights of personality and destination, towards the control of one's own data, associated with the dedication (through a copyleft license attached to personal data) of the research results of data mining to the commons.

52. <http://www.communia-association.org/2014/03/05/communia-responds-to-eu-consultation-on-new-copyright-rules/>; <http://ec.europa.eu/internal_market/consultations/2013/copyright-rules/index_en.htm>.

References

Bastard, Irène *et al.*, "Travail et travailleurs de la donnée," *InternetActu.net*, December 13, 2013, available at <http://www.internetactu.net/2013/12/13/travail-et-travailleurs-de-la-donnee/>.

Carroll, Michael W., "Complying with the National Institutes of Health Public Access Policy: Copyright considerations and options," A joint SPARC/Science Commons/ARL White Paper, February, 2008, available at <http://sparc.arl.org/resources/papers-guides/nih-copyright>.

Chavan, Vishwas S. and Ingwersen, Peter, "Towards a data publishing framework for primary biodiversity data: Challenges and potentials for the biodiversity informatics community," *BMC Bioinformatics*, 10, Suppl. 14, 2009: S2, available at <http://www.biomedcentral.com/1471-2105/10/S14/S2>.

Chavan, Vishwas S. and Penev, Lyubomir, "The data paper: A mechanism to incentivize data publishing in biodiversity science," *BMC Bioinformatics*, 12, Suppl. 15, 2011: S2, available at <http://www.biomedcentral.com/1471-2105/12/S15/S2>.

Colston, Catherine, "*Sui Generis* database right: Ripe for review?," *Journal of Information, Law & Technology*, 3, 2001, available at <https://www2.warwick.ac.uk/fac/soc/law/elj/jilt/2001_3/colston/>.

Crawford, Kate, "The hidden biases in big data," Harvard Business Review, April 01, 2013, available at <https://hbr.org/2013/04/the-hidden-biases-in-big-data>.

Desmet, Peter, "Analyzing the licenses of all 11,000+ GBIF registered datasets" [online], November 22, 2013, available at <http://peterdesmet.com/posts/analyzing-gbif-data-licenses.html>.

Dulong de Rosnay, Melanie, "Check your data freedom: A taxonomy to assess life science database openness," *Nature Preceedings*, July, 2008, available at <http://sciencecommons.org/wp-content/uploads/npre20082083-1.pdf>.

Dulong de Rosnay, Melanie, "From free culture to open data: Technical requirements for access and authorship," *in* D. Bourcier, P. Casanovas, M. Dulong de Rosnay and C. Maracke (eds.), *Intelligent Multimedia. Managing Creative Works in a Digital World,* Florence, European Press Academic Publishing, 2010: 47–66.

Dulong de Rosnay, Melanie, "Communia association Position on EC Horizon 2020 open access policy," *Communia blog*, November 20, 2012, available at <http://www.communia-association.org/2012/11/20/position-on-ec-horizon-2020-open-access-policy/>.

Dulong de Rosnay, Melanie and Guadamuz, Andrés, "Open access to biodiversity scientific data: A comparative study," paper presented at the 17th International Consortium on Applied Bioeconomy Research ICABR Conference on Innovation and the Policy for the Bioeconomy (Ravello [Italy], June 18-21, 2013).

European Commission High Level Group on Scientific Data, "Riding the Wave: How Europe can gain from the rising tide of scientific data," Final Report to the European Commission, October, 2010, available at <http://ec.europa.eu/information_society/newsroom/cf/itemlongdetail.cfm?item_id=6204>.

Fayyad, Usama, Piatetsky-Shapiro, Gregory and Smyth, Padhraic, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine*, 37(3), 1996, available at <http://dx.doi.org/10.1609/aimag.v17i3.1230>.

Guadamuz, Andres and Cabell, Diane, "Data mining in UK higher education institutions: Law and policy," *Queen Mary Journal of Intellectual Property*, 4(1), 2014: 3–29, 2014, available at <http://www.elgaronline.com/view/journals/qmjip/4-1/qmjip.2014.01.01.xml>.

Guibault, Lucie, "Licensing research data under open access condition," *in* D. Beldiman (ed.), *Information and Knowledge: 21st Century Challenges in Intellectual Property and Knowledge Governance*, Cheltenham, Edward Elgar, 2013, available at <http://www.ivir.nl/publicaties/download/Open_Research_Data.pdf>.

House of Commons Business, Innovation and Skills Committee, "The hargreaves review of intellectual property: Where next?," First Report of Session 2012–13, June 21, 2012, available at <http://www.publications.parliament.uk/pa/cm201213/cmselect/cmbis/367/367.pdf>.

Keller, Paul, "Open letter regarding the Commission's stakeholder dialogue on text and data mining," *Communia blog*, February 27, 2013, available at <http://www.communia-association.org/2013/02/27/open-letter-regarding-the-commissions-stakeholder-dialogue-on-text-and-data-mining/>.

Keller, Paul, "Research sector, SMEs, civil society groups and open access publishers withdraw from licences for Europe dialogue on text and data mining," *Communia blog*, May 25, 2013, available at <http://www.communia-association.org/2013/05/25/research-sector-smes-civil-society-groups-and-open-access-publishers-withdraw-from-licences-for-europe-dialogue-on-text-and-data-mining/>.

Melis, Rene J. *et al.*, "Sharing of research data," *The Lancet*, 378(9808), 2011: 1995, available at <http://dx.doi.org/10.1016/S0140-6736(11)61870-9>.

Moscon, Valentina, "Open access to scientific articles: Comparing Italian with German law," *Kluwer Copyright blog*, December 3, 2013, available at <http://kluwercopyrightblog.com/2013/12/03/open-access-to-scientific-articles-comparing-italian-with-german-law/>.

Nissenbaum, Helen, "A contextual approach to privacy online," *Daedalus*, 140(4), Fall 2011: 32–48, available at <http://www.amacad.org/publications/daedalus/11_fall_nissenbaum.pdf>.

Stodden, Victoria, Guo, Peixuan and Ma, Zhaokun, "How journals are adopting open data and code policies," paper submitted to The First Global Thematic IASC Conference on the Knowledge Commons: Governing Pooled Knowledge Resources (Louvain-la-Neuve, Belgium, September 12, 2012).

Sweeney, Latanya, Abu, Akua and Winn, Julia, "Identifying participants in the personal genome project by name," Data Privacy Lab, IQSS, Harvard University, April 24, 2013, White paper, available at <http://dataprivacylab.org/projects/pgp/1021-1.pdf>.

Vines, Timothy H. *et al.*, "The availability of research data declines rapidly with article age," *Current Biology*, 24(1), 2014: 94–97, available at <http://dx.doi.org/10.1016/j.cub.2013.11.014>.

World Intellectual Property Organization Secretariat, "Summary on existing legislation concerning intellectual property in non-original databases," text prepared for the Standing Committee on Copyright and Related Rights: Eighth Session (Geneva, November 4–8, 2002), document SCCR/8/3, September 23, 2002.