



<i>Project Acronym</i>	<i>SoBigData</i>
<i>Project Title</i>	<i>SoBigData Research Infrastructure Social Mining & Big Data Ecosystem</i>
<i>Project Number</i>	<i>654024</i>
<i>Deliverable Title</i>	<i>IP principles and business models</i>
<i>Deliverable No.</i>	<i>D2.7</i>
<i>Delivery Date</i>	<i>01 September 2016</i>
<i>Authors</i>	<i>Iryna Lishchuk (LUH), Stefanie Hännold (LUH), Nikolaus Forgó (LUH)</i>



DOCUMENT INFORMATION

PROJECT	
Project Acronym	SoBigData
Project Title	SoBigData Research Infrastructure Social Mining & Big Data Ecosystem
Project Start	1st September 2015
Project Duration	48 months
Funding	H2020-INFRAIA-2014-2015
Grant Agreement No.	654024
DOCUMENT	
Deliverable No.	D2.7
Deliverable Title	IP principles and business models
Contractual Delivery Date	31 August 2016
Actual Delivery Date	01 September 2016
Author(s)	Iryna Lishchuk (LUH), Stefanie Hänold (LUH), Nikolaus Forgó (LUH)
Editor(s)	Theresia Rasche (LUH), Iryna Lishchuk (LUH)
Reviewer(s)	
Contributor(s)	
Work Package No.	WP2
Work Package Title	WP2 – NA1_Legal and Ethical Framework
Work Package Leader	TU Delft
Work Package Participants	CNR, UNIPI, LUH
Dissemination	Public
Nature	Report
Version / Revision	V1.0
Draft / Final	Final
Total No. Pages (including cover)	84
Keywords	Copyright, Intellectual Property Rights (IPR), data ownership

DISCLAIMER

SoBigData (654024) is a Research and Innovation Action (RIA) funded by the European Commission under the Horizon 2020 research and innovation programme.

SoBigData proposes to create the Social Mining & Big Data Ecosystem: a research infrastructure (RI) providing an integrated ecosystem for ethic-sensitive scientific discoveries and advanced applications of social data mining on the various dimensions of social life, as recorded by “big data”. Building on several established national infrastructures, SoBigData will open up new research avenues in multiple research fields, including mathematics, ICT, and human, social and economic sciences, by enabling easy comparison, re-use and integration of state-of-the-art big social data, methods, and services, into new research.

This document contains information on SoBigData core activities, findings and outcomes and it may also contain contributions from distinguished experts who contribute as SoBigData Board members. Any reference to content in this document should clearly indicate the authors, source, organisation and publication date.

The document has been produced with the funding of the European Commission. The content of this publication is the sole responsibility of the SoBigData Consortium and its experts, and it cannot be considered to reflect the views of the European Commission. The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

The European Union (EU) was established in accordance with the Treaty on the European Union (Maastricht). There are currently 27 member states of the European Union. It is based on the European Communities and the member states’ cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice, and the Court of Auditors (<http://europa.eu.int/>).

Copyright © The SoBigData Consortium 2015. See <http://project.sobigdata.eu/> for details on the copyright holders.

For more information on the project, its partners and contributors please see <http://project.sobigdata.eu/>. You are permitted to copy and distribute verbatim copies of this document containing this copyright notice, but modifying this document is not allowed. You are permitted to copy this document in whole or in part into other documents if you attach the following reference to the copied elements: “Copyright © The SoBigData Consortium 2015.”

The information contained in this document represents the views of the SoBigData Consortium as of the date they are published. The SoBigData Consortium does not guarantee that any information contained herein is error-free, or up to date. THE SOBIGDATA CONSORTIUM MAKES NO WARRANTIES, EXPRESS, IMPLIED, OR STATUTORY, BY PUBLISHING THIS DOCUMENT.

GLOSSARY

ABBREVIATION	DEFINITION
IP	Intellectual Property
IPR	Intellectual Property Rights
CoPhIR	Content-based Photo Image Retrieval
CJEU	Court of Justice of European Union
EU	European Union
UK	United Kingdom
RI	Research Infrastructure
API	Application Programming Interface
IoT	Internet of Things
NDA	Non-Disclosure Agreement
WIPO	World Intellectual Property Organisation

TABLE OF CONTENT

DOCUMENT INFORMATION	2
DISCLAIMER	3
GLOSSARY	5
TABLE OF CONTENT	6
DELIVERABLE SUMMARY.....	8
EXECUTIVE SUMMARY	9
1 Relevance to SoBigData.....	10
1.1 Purpose of this document	10
1.2 Relevance to project objectives	10
1.3 SOBIGDATA project description	10
1.4 Relation to other workpackages	11
1.5 Structure of the document.....	11
2 Potential protection of datasets by IPR	12
2.1 Protection of creative content by copyright	13
2.1.1 Legal framework	13
2.1.2 Requirements for protection	14
2.1.3 Protection of investment in producing the data by related rights	16
2.1.4 Copyright relevant actions.....	17
2.1.5 Exceptions and limitations	20
2.1.6 Conclusions for SoBigData	21
2.2 The sui generis database right	22
2.2.1 Definition of “database”	24
2.2.2 Object of sui generis protection.....	25
2.2.3 Scope of protection and exceptions	27
2.2.4 Term and beneficiaries of protection	29
2.2.5 Conclusions for SoBigData	29
2.3 Protection of data as know-how	30
2.4 Ownership of IP rights	32
3 Responsible IP principles.....	35
4 Data ownership issues.....	38
4.1 A notion of data ownership	38
4.2 Governing use of data by contractual relations	39
4.3 Ownership of personal information	41

5 Conclusions43

Appendix A. IP relevant metadata.....46

Appendix B. SoBigData datasets, applicable rights and terms of use50

REFERENCES.....81

DELIVERABLE SUMMARY

This deliverable was produced in the result of preliminary analysis of IPR issues in the SoBigData project and reflects the results reached so far. The three main aspects, which we consider here, concern protectability of SoBigData datasets by IPR (Section 2), responsible IP principles (Section 3) and data ownership issues (Section 4).

We start Section 2 with an overview of datasets collected in the SoBigData project and seek to determine whether such datasets may be protected by IPR. In particular, copyrights and sui generis database rights come into focus. Since most datasets have been produced by post-processing already existing data, the IP protection may apply to the existing data, from which the SoBigData datasets have been produced (and in which case the terms of use of such data must be observed) or the IP protection may also be acquired by processing the data in SoBigData project. We consider the requirements for protection by copyrights and sui generis rights in application to the SoBigData datasets and data collections.

Section 3 concerns the issue of data ownership. The relations of ownership over the data are to be assessed in the complex of factors. If the dataset is protected by IPR, then it is the question of ownership over the IPR, and not the data as such, which is important. The ownership over the data, which does not reach the level of IP protection, has to be considered from perspective of the legal and contractual relations, under which such data has been generated. The ownership of personal information is to be considered from the perspective of data protection law.

Compliance with the rights of third parties and terms of use of external datasets makes one of the essential IP principles for the SoBigData RI. The legal analysis revealed that a major number of datasets may be protected by copyrights and the use of such data is subject to the terms of use of the right holder or external source, where the data is taken from. The overview of data types in SoBigData, applicable rights and terms of use is presented in Appendix B. It is advised that the information on IP aspects, surrounding each particular dataset, should be integrated into the dataset metadata and be completed by the data providing party. The advised list of parameters of IP relevant metadata (i.e. type of IP, source, license/terms of use, basic rights, etc.) is provided in Appendix A.

EXECUTIVE SUMMARY

This deliverable considers the aspects of Intellectual Property (IP), associated with the data collected within the SoBigData project and sharing such data. This report reflects the preliminary results of the initial legal analysis on the IPR issues in the SoBigData project. The SoBigData infrastructure is aimed as the European Research Infrastructure for Big Data and Social Mining and accumulates various datasets from different sources, including social media content (like tweets, blogs, etc), call graphs from mobile phone call data, networks crawled from many online social networks, including Facebook and Flickr, etc. Collecting, using and sharing of such data raise questions about intellectual property and data ownership.

In this deliverable, we define the types of protection, which may be applicable to the datasets in the SoBigData project, describe the legal requirements for protection, which the datasets need to satisfy in order to be protected, cover the IP issues and legal requirements for sharing the datasets, which have already been accumulated in the project and consider the issues of data ownership. In the center of consideration are such aspects as: whether social media content can achieve IP protection (“high level creativity”), what type of protection and what legal requirements for protection apply; the conditions under which such IP protected content can be shared in a (virtual) research environment for scientific purposes; the ownership of new data and personal information generated through the analysis of social media, sui generis protection of databases. The basic standards of IP and copyright law that provide protection and grant exclusive rights to the works in return for intellectual input and/or economic investment of owned resources, energy or assets (e.g. algorithms, software, models) are taken into account and balanced against the basic ideas of data protection and privacy principles.

1 RELEVANCE TO SOBIGDATA

1.1 PURPOSE OF THIS DOCUMENT

This deliverable outlines the preliminary IP principles and practices in SoBigData, that have evolved in and outside SoBigData so as to accommodate both legitimate IP claims in data and privacy related rights. The basic requirements of IP law, relevant to datasets in SoBigData, the conditions for sharing IP protected content via digital media, the protection of investment in collection, arrangement and presentation of data by sui generis database rights, the issue of data ownership have been analyzed and are considered here.

1.2 RELEVANCE TO PROJECT OBJECTIVES

The SoBigData is aimed as the European Research Infrastructure for Big Data and Social Mining. The SoBigData infrastructure already contains datasets, collected and provided by the project partners. The data collection is going to be enriched by the new datasets, which may either be collected for the project by the partners or uploaded by the users. Such new datasets may include Social Media Content (like tweets, blogs, etc), data taken from the public domain (such as hotel or airport statistics, banking and stock exchange information, etc.), geolocation data collected by the apps, images, bibliographies, to name a few. Creative content, shared over social media (like blogs, comments, photos, etc.) may fall under copyright and data protection law. Data collected by private companies or governmental bodies by virtue of economic investments may be protected by sui generis database rights. Collecting, using and sharing such content in digital environment may raise questions about the applicable IP protection and conditions, under which such content may be used and shared within SoBigData project. Also, the issue of data ownership is relevant for defining who has what rights in the accumulated data.

1.3 SOBIGDATA PROJECT DESCRIPTION

“SoBigData proposes to create the Social Mining & Big Data Ecosystem: a research infrastructure (RI) providing an integrated ecosystem for ethic-sensitive scientific discoveries and advanced applications of social data mining on the various dimensions of social life, as recorded by “big data”. Building on several established national infrastructures, SoBigData will open up new research avenues in multiple research fields, including mathematics, ICT, and human, social and economic sciences, by enabling easy comparison, re-use and integration of state-of-the-art big social data, methods, and services, into new research. It will not only strengthen the existing clusters of excellence in social data mining research, but also create a pan-European, inter-disciplinary community of social data scientists, fostered by extensive training, networking, and innovation activities. In addition, as an open research infrastructure, SoBigData will promote repeatable and open science. Although SoBigData is primarily aimed at serving the needs of researchers, the openly available datasets and open source methods and services provided by the new research

infrastructure will also impact industrial and other stakeholders (e.g. government bodies, non-profit organisations, funders, policy makers).¹

1.4 RELATION TO OTHER WORKPACKAGES

The IPR issues and defining IP principles responsible for the data sharing are of particular importance to workpackages, dealing with data management in the SoBigData project. In particular, Task 2.4 correlates with WP 8 Big Data Ecosystem, WP10 SoBigData e-Infrastructure. WP8 considers strategies for data sharing with the scientific community, inside the consortium, on a bilateral basis, or allowing data access within secure environments. In this respect, the IP issues, associated with each particular dataset, which need to be observed by data sharing, have been identified and incorporated into metadata as part of Task 8.1. WP10 supports the creation and operation of the SoBigData e-infrastructure. In particular, Task 10.1 considers types of resources (e.g. datasets, services, algorithms) and metadata properties of the resources (e.g. classifications, access parameters, hardware parameters, IPR issues, input/output data types).

1.5 STRUCTURE OF THE DOCUMENT

The rest of this deliverable is organized as follows. Section 2 deals with the potential protection of SoBigData datasets by IPR and describes the requirements for protection. In Section 3 we describe some responsible IP principles and practices of data managing in SoBigData. Section 4 elaborates on the data ownership issues. Conclusions finalize this deliverable in Section 5.

¹ Horizon 2020 Call: H2020-INFRAIA-2014-2015 Topic: INFRAIA-1-2014-2015 Type of action: RIA Proposal number: 654024 Proposal acronym: SoBigData, Abstract.

2 POTENTIAL PROTECTION OF DATASETS BY IPR

SoBigData is aimed as the European research infrastructure for big data and social mining. The SoBigData platform is supposed to aggregate a large amount of data from various sources. A number of datasets are already available and provided to the project by the project partners. This data includes call graphs from mobile phone call data, networks crawled from many online social networks, including Facebook and Flickr, transaction micro-data from diverse retailers, query logs both from search engines and e-commerce, societywide mobile phone call data records, GPS tracks from personal navigation devices, survey data about customer satisfaction or market research, billions of tweets, and data from location-aware social networks. The dataset catalog is available at: <http://www.sobigdata.eu/private/ouhnr3inebchcerhncurhebcnbn8749892/datasets>.

Some datasets have been generated in the result of analyzing the data from social networks (e.g., networks crawled from social networks, including Facebook² and Twitter³). Some data come in the result of post-processing digital images (e.g. CoPHiR provides an XML structure with 5 extracted standard MPEG-7 image features: Scalable Colour, Colour Structure, Colour Layout, Edge Histogram, Homogeneous Texture. The data collected consist of 106 million processed images from Flickr)⁴. Some data come from databases with biological data (e.g. DNA: collection of families of genomes (type: highly repetitive biological data). Some data reflect statistic data on hotels, airport and railway traffic⁵, banking transactions⁶, etc.

At the present stage of the European legislation, there are several potential schemes of legal protection that may come into question as options for protecting datasets, collected within the SoBigData research infrastructure. These include the sui generis database right; know how protection, copyrights and related rights. These types of protection are tailored to specific objects. The central question whether such datasets or data collections may achieve intellectual property or sui generis protection depends on whether the one or another dataset/data collection satisfies the legal requirements for protection. Thus, social media content, which achieves “high level creativity” and shows originality may be protected by copyrights. Data collected and arranged into a data repository by investment of considerable resources (e.g. financial, technical, human power, etc.) can be protected by sui generis database rights. Data, which possesses certain economic value and has been handled subject to the confidentiality measures, may

² SoBigData, Data Catalog, Facebook EuroSys '09, available at: <http://www.sobigdata.eu/content/facebook-eurosys-09>.

³ SoBigData, Data Catalog, Italian Twitter Dataset, available at: <http://www.sobigdata.eu/content/italian-twitter-dataset>.

⁴ SoBigData, Data Catalog, CoPHiP, available at: <http://www.sobigdata.eu/content/cophir>.

⁵ SoBigData, Data Catalog, Pisa Airport Statistics, available at: <http://www.sobigdata.eu/dataset/pisa-airport-statistics>.

⁶ SoBigData, Data Catalog, e-mid, available at: <http://www.sobigdata.eu/dataset/e-mid>.

achieve protection under the legal regime of knowhow. Questionable is whether the investment in processing the data by enriching or enhancing already available datasets may attract protection by IPR. The preliminary analysis of SobigData datasets and applicable is presented in Appendix B.

These types of IP regimes, which may be applicable for protecting the datasets, and the legal requirements for protection we consider below.

2.1 PROTECTION OF CREATIVE CONTENT BY COPYRIGHT

As noted above, a number of datasets in SoBigData are generated by social media analysis or crawling. Examples include SoBigData dataset Aalto Twitter, containing about 418 million tweets from June 25, 2015 - September 19, 2015, containing tweets from trending hashtags gathered through the public Twitter api⁷, Facebook - EuroSys '09 - social and interaction graphs representing two large-scale Facebook regional networks⁸.

A number of items, which may usually appear in one or another form on the platform services, be it Twitter or Facebook, may fall into the category of copyright works. Such content items, as images, melodies, videos, commentaries, blogs, etc. require and are typically produced by some intellectual or technical effort, such as like designing a picture, composing a melody, singing or recording a song. Such works, which relate to original intellectual creations in the literary or artistic domain constitute works protected by copyright.

A range of copyright works is large and may include writings, lectures, musical compositions with or without words, cinematographic works, works of drawing, painting, photographic works, illustrations, maps, plans, sketches, etc⁹. Such works may be protected either by traditional copyright or related rights. Related rights usually subsist in fixations of copyright works, such as phonograms, fixations of films, performances and broadcasts¹⁰.

2.1.1 LEGAL FRAMEWORK

The primary legal framework of copyright law is formed by the international treaties, such as Berne Convention for the Protection of Literary and Artistic Works¹¹, TRIPS Agreement¹², WIPO Copyright

⁷ SoBigData, Data Catalog, Aalto Twitter, available at: <http://www.sobigdata.eu/dataset/aalto-twitter>.

⁸ SoBigData, Data Catalog, Facebook - EuroSys '09, available at: <http://www.sobigdata.eu/content/facebook-eurosys-09>.

⁹ Article 2 Berne Convention.

¹⁰ International Convention for the Protection of Performers, Producers of Phonograms and Broadcasting Organizations, Rome, 26 October 1961.

¹¹ Berne Convention for the Protection of Literary and Artistic Works of 9 September 1886.

¹² Agreement on Trade-Related Aspects of Intellectual Property Rights, the TRIPS Agreement, Appendix 1C of the Marrakesh Agreement Establishing the World Trade Organization, Marrakesh, Morocco, 15 April 1994.

Treaty¹³. At the European law level copyright law is harmonized by a number of Directives, such as: Directive 2001/29/EC of 2001 on the harmonisation of certain aspects of copyright and related rights in the information society¹⁴, Directive 2004/48/EC of 2004 on the enforcement of intellectual property rights¹⁵, Directive 2006/116/EC of 2006 on the term of protection of copyright and certain related rights (codified version)¹⁶. The copyright law and requirements for copyright, applicable in the Member States, are defined by the national copyright laws.

2.1.2 REQUIREMENTS FOR PROTECTION

Traditional copyright law has distinct requirements for protection according to the type of work at issue. Traditionally, copyright protects original works *“in the literary, scientific and artistic domain, whatever may be the mode or form of its expression”*¹⁷. As regards copyright in the works of writing, the standard of copyright, established by the CJEU applies at the level of European law. According to the CJEU, which dealt with copyrightability of text extracts from articles in the case *Infopaq International A/S*, *“it is only through the choice, sequence and combination of those words that the author may express his creativity in an original manner and achieve a result which is an intellectual creation.”*¹⁸ And this standard applies not only to copyright works as a whole, but also to portions, extracted from such protected works, which may also be copyrighted. According to the court various parts of a work enjoy copyright protection, *“provided that they contain elements which are the expression of the intellectual creation of the author of the work”*¹⁹. It means, that even portions of a copyright work are protected and reproduction, distribution and communication of such works via digital media requires consent of the right holder, unless exemptions apply. Along with that, as clarified by the CJEU, isolated items, be they words, keywords, syntax, figures or mathematical concepts will not attract copyright, since they do not amount to a creative expression. Such items, *“considered in isolation, are not as such an intellectual creation of the author who employs them. It is only through the choice, sequence and combination of those words that the author may express his creativity in an original manner and achieve a result which is an intellectual creation.”*²⁰

¹³ WIPO Copyright Treaty (WCT), Geneva, 20 December 1996.

¹⁴ Directive 2001/29/EC of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society, Official Journal of the European Communities, L 167/10, 22.6.2001.

¹⁵ Directive 2004/48/EC of 2004 on the enforcement of intellectual property rights, OJEU, L 195/16, 2.6.2004.

¹⁶ Directive 2006/116/EC of 2006 on the term of protection of copyright and certain related rights (codified version), OJEU, L 372/12, 27.12.2006.

¹⁷ Berne Convention for the Protection of Literary and Artistic Works of 9 September 1886.

¹⁸ CJEU, Judgment of 16 July 2009, Case C 5/08, *Infopaq International A/S v Danske Dagblades Forening*, Recital 45.

¹⁹ *Ibid*, Recital 39.

²⁰ CJEU, Judgment of 2 May 2012, Case C 406/10, *SAS Institute Inc v World Programming Ltd*, ECLI:EU:C:2012:259.

Although the European Directives provide some guidelines and set standards for copyright protection, the national laws, however, have different interpretation criteria of copyright works. Thus, under the UK law “*anything worth copying is prima facie worth protecting*”²¹. The German law requires personal intellectual creation for the work to attract copyright protection²². The French law grants copyright to a work, which is a result of intellectual creation: “*l’oeuvre de l’esprit résultat d’une activité creative*”²³.

Also, the criteria of individuality/originality vary from jurisdiction to jurisdiction. The UK law gives much importance to protect skill and labour, which the authors deploy in their works²⁴. Thus, the UK law measures originality by the “*skill, judgment, labour*”, invested into a work, and may grant copyright protection to a work “*not copied from another source*”²⁵. The German law places the copyright protection upon the “*Prägetheorie*”, requiring reflection of individuality of the author in a copyright work. Traits of the author’s personality constitute margin for creativity under the German law. The French copyright law also operates with the criteria of originality and traces of personality in a copyright work (originalité, “*l’empreinte de la personnalité*”), however, has a low standard for protection²⁶.

Against these observations it may be said that despite the European standard of copyright, set by the CJEU, providing that the works (both as elements of such works), which constitute an expression of the intellectual creation of the author, are protected by copyright, specific criteria for copyright works apply in national jurisdictions. Whereas the German or French law would protect a work which amounts to an original creative expression, the UK law is more flexible and may grant copyright to a work to remunerate the author for the investment, which he made in creating a work, regardless whether such work reflects his original creativity or not. It has a result that assessment whether one or another piece of social media content is protected by copyright shall be done on a case by case basis and depending on the national requirement of copyright law applicable for the work at issue.

A similar standard of copyright applies to photographic works. According to Recital 16 Directive 2006/116/EC, a photographic work is protected by copyright, if it is original. A work “*is to be considered original if it is the author’s own intellectual creation reflecting his personality*”. Other criteria such as merit or purpose are not relevant for copyright. According to the CJEU decision in the case C 145/10 REC of Eva-Maria Painer, copyright protects pictures taken by an individual exercising free and creative choices thus

²¹ University of London Press Ltd v University Tutorial Press Ltd ([1916] 2 Ch 60).

²² Article 2 “Urheberrechtsgesetz vom 9. September 1965 (BGBl. I S. 1273), das zuletzt durch Artikel 7 des Gesetzes vom 4. April 2016 (BGBl. I S. 558) geändert worden ist”/Act on Copyright and Related Rights (Copyright Act).

²³ Article L. 1121 Code de la propriété intellectuelle (CPI)/French Intellectual Property Code.

²⁴ S. Stokes, “Digital Copyright, Law and Practice,” 4. Edition, Hart Publishing, Oxford and Portland, Oregon, 2014, p. 121.

²⁵ Ibid.

²⁶ Article L. 1121 CPI/French Intellectual Property Code.

stamping a picture with his personal touch²⁷. It means, pictures from social media, which were taken by the users with some level of creativity may be protected by copyright. On the other hand, images, generated automatically, will lack the creative input and may not be protectable by copyright.

Against this legal background, once items from social networks or media, such as blogs, photographs, comments, messages, etc. are supposed to be used and shared via SoBigData platform, one must count that such items may be protected by copyrights and their use may be subject to the terms, stipulated by the right holder and/or the platform.

The preliminary analysis of IP rights which may be applicable to the datasets collected within the SoBigData platform along with the terms, under which such datasets may be used, is provided in Appendix B.

2.1.3 PROTECTION OF INVESTMENT IN PRODUCING THE DATA BY RELATED RIGHTS

Apart from the copyrights considered so far, there are in the field of intellectual property law a number of other emerging rights granted as a response to investment in producing the works. These rights are normally provided to the person, who invests in producing the protectable information. Such rights are referred to as related rights. Protection by related rights does not necessarily link to the intellectual creation (as the case is with traditional copyright), but rather to the economic investment.

The major rationale for protection by related rights tends to shift between intellectual creation and investment²⁸. A mixture of artistic creation and investment attracts exclusive rights to performers in fixations of their performances. The economic investment constitutes a major factor, which renders exclusive rights to phonogram producers in their phonograms, to the film producers in respect of first fixations of their films, to broadcasting organizations in fixations of their broadcasts²⁹. However, the number of related rights as of now is rather limited (mostly to those, indicated above). Hence, provided phonograms, broadcasts or movies are to be used in SoBigData (such as the case with IMDb dataset may be, which accumulates movie, TV and celebrity content³⁰), protectability of such content by related rights shall be taken into account. The preliminary assessment of which datasets may be protected by related rights and under what conditions such datasets may be used is provided in Appendix B.

Once a content item transmitted via platform services satisfies the requirements relating to one or other of the above kinds of works, any use of such items, in particular in the context of digital media, may

²⁷ CJEU, Judgment of 7 March 2013, Case C 145/10 REC, *Eva-Maria Painer v. Standard VerlagsGmbH, Axel Springer AG, Süddeutsche Zeitung GmbH, Spiegel-Verlag Rudolf Augstein GmbH & Co. KG, Verlag M. DuMont Schauberg Expedition der Kölnischen Zeitung GmbH & Co. KG*, Recital 94.

²⁸ H. Zech, *Information als Schutzgegenstand*, at p. 142.

²⁹ DIRECTIVE 2001/29/EC of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society, OJEU L 167/10 - L 167/19, 22.6.2001.

³⁰ SoBigData, Data Catalog, IMDb, available at: <http://www.sobigdata.eu/content/imdb>.

constitute a copyright relevant action and would require authorization of the right holder. What uses of copyright works constitute copyright relevant actions we consider in the next section below.

2.1.4 COPYRIGHT RELEVANT ACTIONS

Whereas reading a hardcopy of a book or listening to music via radio does not create a copyright relevant action, upload of a book in digital form, download of a music mp3 file or display of an avi file via digital media may produce a copyright relevant action and require the authorization of the right holder. The simple reason is that, in contrast to where there is simple perception of the work by a viewer or hearer, technical actions of these kinds involve a degree of copying or (re)publication.

The Directive 2001/29/EC provides authors of copyright works, performers, phonogram producers, producers of the first fixations of films, broadcasting organizations with the exclusive rights on reproduction, distribution and communication to the public³¹. In particular, the right holders have the exclusive right to authorize or prohibit reproduction of their works either directly or indirectly, temporary or permanently, by any means and in any form, in whole or in part; communication to the public by wire or wireless means, including the making available to the public of their works in such a way that members of the public may access them from a place and at a time individually chosen by them.

Apart from these three basic rights, the Berne Convention attributes to exclusive rights of the right holders and subjects to their authorization such acts as: translation³², adaptation, arrangement and other alterations³³.

Relevant to the processing of data on the SoBigData platform in the context of copyright law may be the CJEU case Infopaq International³⁴. In this case, the court qualified the data capture process performed on the newspaper articles as reproduction, belonging to exclusive rights of the right holders and subject to their authorization.

In the present case, Infopaq operated a media monitoring and analysis business, which consists primarily in drawing up summaries of selected articles from Danish daily newspapers and other periodicals. The articles are selected on the basis of certain subject criteria agreed with customers and the selection is made by means of a 'data capture process'. The summaries are sent to the customers by email. The data capture process, conducted by Infopaq, comprises five stages. First, the relevant publications are registered manually by Infopaq employees in an electronic registration database. Secondly, the publications³³ are scanned. The section to be scanned is selected from the registration database, scanning

³¹ Directive 2001/29/EC of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society, Official Journal of the European Communities, L 167/10, 22.6.2001.

³² Article 8 Berne Convention.

³³ Ibid, Article 12.

³⁴ CJEU, Judgment of 16 July 2009, Case C 5/08, Infopaq International A/S v Danske Dagblades Forening.

produces a TIFF ('Tagged Image File Format') file for each page, after scanning the TIFF file is transferred to an OCR ('Optical Character Recognition') server. In the third step, the OCR server translates the TIFF file into data that can be processed digitally. Fourthly, the text file is processed to find a search word defined beforehand. Each time a match for a search word is found, data is generated giving the publication, section and page number on which the match was found. A value in percentage between 0 and 100 indicates how often the text it is to be found in the article. At the end of the data capture process a cover sheet is printed out noting the pages where the relevant search word was found. For example, the text of a cover sheet may look as follows: '4 November 2005 – Dagbladet Arbejderen, page 3: TDC: 73% "a forthcoming sale of the telecommunications group TDC which is expected to be bought"'.

DDF, a professional association of Danish daily newspaper publishers, complained that Infopaq was processing articles for commercial purposes without authorisation from the relevant rightholders. DDF took the view that such data capture activity amounted to reproduction of copyright works, constituted a copyright relevant action and required consent from the right holders. Infopaq argued that the acts in question, are covered by the exemption from the right of reproduction provided for in Article 5(1) Directive 2001/29.

Under Article 5(1) of Directive 2001/29, an act of reproduction may be exempted from the reproduction right if it satisfies five conditions:

- the act is temporary;
- it is transient or incidental;
- it is an integral and essential part of a technological process;
- the sole purpose of that process is to enable a transmission in a network between third parties by an intermediary of a lawful use of a work or protected subject-matter; and
- the act has no independent economic significance.

These conditions are cumulative meaning that if one condition is not fulfilled, the act would amount to the reproduction right not exempted by the Directive.

The court found that an act may count as temporary and transient reproduction only if such act forms an integral part of a technological process, enables its completion and does not exceed what is necessary for the proper completion of that technological process³⁵. Also, in light of Recital 33 of the Directive, the court clarified that the exemption provided by Article 5(1) of the Directive covers only those acts of temporary reproduction, which are created and deleted automatically and without human intervention, such as caching. Deletion of copies upon discretionary human intervention, particularly by the user of protected works, would remove the actions of reproduction from the exemption. The act of reproduction counts as transient, only "*if its duration is limited to what is necessary for the proper completion of the technological*

³⁵ Infopaq International, *supra*, Recital 61.

*process in question*³⁶. The copies are transient as long as they are deleted automatically from the computer memory. In those circumstances, reproduction of copyright works with the possibility to delete the copies dependent on the will of the user and making copies outside the sphere of computer technology (such as printing or extracting copies to external medium) does not fulfill the second condition and may not be covered by the exemption from the reproduction right. Accordingly, reproducing the copies of copyright works to external media upon control of the user having the power to decide about their deletion cannot be carried out without the authorization of the copyright holder.

This case has a direct bearing to such acts of data processing within SoBigData platform, which include or allow extraction (download) of datasets for further analysis. An example may be the data procession conducted for the SoBigData story Monitoring Topics across Time and Space³⁷. This story foresees the extraction of datasets from the web archive. In this story, the test processing was made on the publication of a heavily discussed book, but may include larger and more heterogeneous collections in the future. The story also foresees embedding images - the process, which includes copying and insertion of selected image fragments to another media or memory space or virtual environment. Provided the datasets or images, selected for extraction, are copyrighted (according to the requirements for protection described in Section 2.1.2 above), such actions may qualify as copyright relevant reproduction and be subject to the authorization of the right holder/data provider.

In the given story (Monitoring Topics across Time and Space), the data processing is made on the data from the German Web Archive. The terms of use of the dataset German Web Archive in the SoBigData project are governed by the Memorandum of Understanding (MoU) between the Leibniz Universität Hannover and data provider Internet Archive. The MoU provides for the creation of a topical subset of the .de domain from the .de Web Data collection, upload of the subset into a Gate Cloud (a text processing infrastructure provided by the University of Sheffield), procession of the data subset to extract entities, topics, sentiments and keywords and transfer of the results to other data processing services in the SoBigData RI. Provided the selected datasets include copyrighted materials and such materials are subject to extraction or reproduction, the processing within SoBigData RI may be deemed as performed within the contractually allowed framework.

As regards the data processing for the other stories and whether such data processing may qualify as copyright relevant depends (a) whether the data items qualify as copyright works, (b) whether the actions of data processing qualify as copyright relevant. Such actions may include i.a. reproduction of datasets at control of the user (or processing party), extending the scope of automatic temporary reproduction, modification (such as processing of images in the CoPhIR dataset³⁸), arrangement of existing datasets into new subsets, placing such datasets available to the users via SoBigData RI in a way that they may access them from a place and at a time individually chosen by them.

³⁶ Ibid, Recital 64.

³⁷ https://support.d4science.org/projects/sobigdata-eu/wiki/Story_Monitoring_Topics_across_Time_and_Space.

³⁸ <http://www.sobigdata.eu/content/cophir>.

For instance, the case on Migration Studies³⁹ involves social network analysis to understand the migration phenomenon and includes opinion mining analysis to understand how people perception on migration is changing. The data procession is done on the ISTAT 2011 dataset (processed Italian census data, per commune, incl. population, immigrant counts but also immigrant integration measures (i.e. fraction of mixed marriages, fraction of work contracts given to immigrants, school attendance, and others) and INE dataset (Spanish referendum and marriage datasets, similar to ISTAT data). Such statistic datasets shall not have the quality of copyrighted materials. Hence, the procession of such data should not qualify as copyright relevant and require authorization of copyright holder. The data procession may, however, be subject to the terms of the data providers (Istituto Nazionale Di Statistica).

The case City of Citizens⁴⁰ aims to produce an overview of the city and the people living in it. The goal is to describe a city by a set of basic and complex statistics. The story deals i.a. with Flickr geo-localized photos. From this Flickr data a set of sequences of common visited places is extracted (done by the Trip Builder tool). In the given scenario, these are not images as potential copyrighted works, which are extracted (copied), but rather the technical geo-information, which due to the absence of creative input does not have the quality of copyrighted material. Second, the use of geo-metadata for tracking the trajectories shall not qualify as a copyright relevant action, because the data procession is not done on the copyrighted images as such.

The story Integrated Business and Economics⁴¹ uses micro-scale social data to understand macro-scale business and economic phenomena. The story is conducted on the economic datasets: FED curated data of balance sheet + asset class capitalization (publicly available) and market transaction data (commercial dataset, ThomsonReuters). The methods of data procession involve i.a. credit risk scoring, shock propagation, segregation discovery, etc. Also, in the present case, the data, on which the data processing is performed should not have the quality of copyrighted works and the analysis of such data for extraction of knowledge by software tools shall not qualify as copyright relevant. The use of such data may, however, be subject to the terms of data providers.

For further details on the terms, governing the use of datasets in SoBigData RI please see Section 3 and Annex B.

2.1.5 EXCEPTIONS AND LIMITATIONS

Some exemptions, provided for the procession of copyrighted material by the Directive 2001/29/EC may be brought to consideration in SoBigData as well. The Directive foresees some exceptions for copyright relevant actions in Article 5. I.a., the mentioned temporary acts of reproduction, which are transient or

³⁹ https://support.d4science.org/projects/sobigdata-eu/wiki/Migration_Studies.

⁴⁰ https://support.d4science.org/projects/sobigdata-eu/wiki/Story_City_of_Citizens_.

⁴¹ https://support.d4science.org/projects/sobigdata-eu/wiki/Story_Societal_Well-being_&_Economic_Performance.

incidental [and] an integral and essential part of a technological process, use of copyrighted materials for use for the sole purpose of illustration for teaching or scientific research may be relevant.

Article 5 (1) Directive 2001/29/EC provides:

“Temporary acts of reproduction referred to in Article 2, which are transient or incidental [and] an integral and essential part of a technological process and whose sole purpose is to

enable:

(a) a transmission in a network between third parties by an

intermediary, or

(b) a lawful use

of a work or other subject-matter to be made, and which have no independent economic significance, shall be exempted from the reproduction right provided for in Article 2.”

As we considered in Section 2.1.4 above, the acts of processing the data upon responsibility of the user with the user having the possibility to extract the copies to external media and exercising control over deletion of the copies may not be covered by the exemption from reproduction right and performance of such actions requires consent of the right holder.

Also, the exemption from the right of reproduction and communication to the public for the purpose of illustration for teaching and scientific research, provided by Article 5(3)(a) Directive 2001/29, is not applicable to the data procession within the SoBigData project. The data procession, facilitated by the SoBigData infrastructure, goes beyond the goal of pure illustration, and is rather intended to deriving new knowledge via data procession.

Also, the exemption, provided by Article 5(3)(n) of the Directive, is related to research. This exemption allows *“use by communication or making available, for the purpose of research or private study, to individual members of the public by dedicated terminals on the premises of establishments referred to in paragraph 2(c) of works and other subject-matter not subject to purchase or licensing terms which are contained in their collections”*. This exemption is directed at and is limited to public libraries, educational establishments, museums or archives, which do not pursue direct or indirect economic or commercial advantage and applies to the uses of copyright works on the premises of such establishments. The project partners consist not only from the institutions, to whom this exemption may apply. Also, the uses of copyrighted materials in SoBigData exceed the making available or communication of copyright works to the public and such acts are not necessarily performed by dedicated terminals on the premises of such establishments. Since these copyright exemptions are limited and subject to narrow interpretation, this exemption may not be considered directly applicable to SoBigData as well.

2.1.6 CONCLUSIONS FOR SOBIGDATA

First, we consider whether the datasets generated for SoBigData project may be protected by copyrights. SoBigData datasets are mostly generated by one or another type of digital data processing (such as

crawling, data mining, post-processing, etc.). Although the copyright law apart from the protection of intellectual creation also affords protection for the economic investment in producing some type of works by related rights, the list of objects protected by related rights is limited to phonograms, fixations of films, fixations of broadcasts and performances. The related rights do not cover and may not be applicable to protecting the investment, deployed in producing the datasets by software tools so far. The traditional copyright law protects intellectual creation, expressed in a work. Such expression of original intellectual creativity may not be found in datasets, produced by digital means of data processing. Thus, protectability of datasets, which have been generated for SoBigData by software tools and digital means of data processing, by copyright may hardly be an option.

However, related rights and copyrights may well apply to the data, from or on top of which the SoBigData datasets have been produced.

As may be observed from the SoBigData data catalog, the SoBigData project is supposed to make use of the data, made available on social platforms, like Twitter, Facebook, Flickr, etc. The content items, distributed over social media may fall under one or another item, protected by the copyright law, be it a comment, a blog, a piece of writing, an image or a melody. Such items, produced by some intellectual input and reflecting some scope of original creativity of the author may qualify as one or another copyright work and be protected by the law of copyright.

Against the legal background considered above, we can infer that the services of data procession, which are offered by the SoBigData infrastructure to the users, may qualify as one or another copyright relevant action, such as reproduction, communication or making available of copyrighted elements to the public. As we observed above, the actions of processing the data by means of the SoBigData facilities may not be considered as covered by the exemptions, provided by the Directive 2001/29. Unless released from the authorization requirement of the right holder by the Directive, such actions may not be carried out without the consent of the right holder. Indeed, the use of datasets with copyrighted elements may be allowed by the terms of data providers.

In the result of preliminary analysis of IPR issues in SoBigData, some datasets, potentially covered by copyrights have been identified. The legal analysis on the type of datasets, applicable rights and terms, under which protected items may and shall be used, is provided in Appendix B. The integration of the IPR issues to the SoBigData is further explored in Section 3 below.

2.2 THE SUI GENERIS DATABASE RIGHT

One of the main goals of the SoBigData project is to be able to more efficiently collect and organize large amounts of datasets from various data sources and make them available to research, as well as being able to quickly process the data within the research infrastructure.

Consequently, data repositories play a major role in the project. Technically speaking, data repositories are collections of individual data strings arranged in a certain manner.

In legal terms, data repositories are generally classified as *databases*. This part of the deliverable will therefore focus on the protection afforded to databases.

Databases are protected according to one of three basic systems that differ in whether only originality and creativity are recognized as worthy of protection or also other types of investments, such as time and money⁴².

Granting copyright protection even where creativity is lacking, in recognition of the investments which were made in order to compile a database – the so-called “sweat of the brow principle”. This is typical for most common law countries with the exception of the United States⁴³.

The European law grants a separate *sui generis* protection for databases, which may lack originality, but are compiled using other types of investments⁴⁴.

Granting copyright protection to works which possess a certain creativity and owe their existence to an “act of authorship”⁴⁵, as in the United States.

Some jurisdictions also apply the rules of unfair competition and also competition law⁴⁶ to the protection of databases⁴⁷. Their analysis will be excluded from this deliverable, as such protection is regulated on a purely national basis.

The main piece of relevant legislature in the EU is Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases (“Database Directive”)⁴⁸. The Database Directive grants two types of protection depending on the type of effort that went into producing the database: copyright on the one hand and a *sui generis* right on the other. The protection of databases by copyright applies to the database structure and may not be considered relevant to protecting the datasets in the SoBigData project. Therefore, protection of databases by copyright is outside the scope of analysis presented in this deliverable. The protection of databases by *sui generis* rights we consider below.

This section will first consider the definition of the term database, then describe the *sui generis* database rights and finally analyze the *sui generis* rights as applicable to SoBigData.

⁴² Such a categorization is by design rather simplistic, but it nevertheless grants a good overview of the systems in place through the world. See Mark J. Davison, *The Legal Protection of Databases*, Cambridge Studies in Intellectual Property Rights (Cambridge: Cambridge Univ. Press, 2003), p. 10.

⁴³ Ibid.

⁴⁴ Ibid.

⁴⁵ David I. Bainbridge, *Introduction to Information Technology Law*, 6th ed. (Harlow: Pearson Longman, 2008), p. 73.

⁴⁶ Understood as abuse of market power.

⁴⁷ See Davison, *The Legal Protection of Databases*, pp. 37 et seqq, pp. 43 et seqq.

⁴⁸ DIRECTIVE 96/9/EC of 11 March 1996 on the legal protection of databases, *Official Journal L 077*, 27/03/1996 P. 0020 - 0028.

2.2.1 DEFINITION OF “DATABASE”

Article 1(1) clarifies the scope of the Database Directive by stating that it “concerns the legal protection of databases in any form” and thus applies equally to electronic and non-electronic databases.

The Database Directive defines a database very broadly 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.”⁴⁹

“Works” is meant in the traditional copyright sense and refers to literary, artistic and other such works. “Data” refers to “information”, i.e. data understandable to humans, as contrasted to pure data, i.e. data which is independent of meaning and has not yet reached the status of human intelligible information⁵⁰.

What the term “other materials” refers to is unclear. Some commentators argue that it is meant to include physical objects, so that a collection of stamps could achieve protection under the Database Directive, while others argue that nowhere does the Database Directive list any examples of physical objects, but only intangible ones, so that “other materials” must be interpreted in the same vein⁵¹.

Various national courts have made use of the broad definition and, in different decisions, granted database protection to: “telephone directories, collections of legal materials, real estate information websites, radio and television guides, bibliographies, encyclopedias, address lists, company registries, exhibition catalogues, tourism websites, collections of hyperlinks, and hit parades.”⁵²

In considering the database definition, the European Court of Justice affirmed the broad scope of the Database Directive’s wording and explained that a database is “any collection of works, data or other materials, separable from one another without the value of their contents being affected, including a method or system of some sort for the retrieval of each of its constituent materials.”⁵³

However, the Court placed great emphasis on the requirement that the materials collected are “independent”, i.e. “separable from one another without their informative, literary, artistic, musical or other value being affected.”⁵⁴

⁴⁹ See Estelle Derclaye, *The Legal Protection of Databases*, p. 58 (2008).

⁵⁰ Ibid, p. 58 et seq. (2008).

⁵¹ See Estelle Derclaye, p.59.

⁵² Paul Goldstein and P. Bernt Hugenholtz, *International Copyright: Principles, Law, and Practice* (Oxford: Oxford University Press, 2012), p. 242.

⁵³ CJEU, Case C-444/02 *Fixtures Marketing Ltd v Organismos prognostikon agonon podosfairou AE (OPAP)*, para 32.

⁵⁴ Ibid, para 29.

Consequently, there are few restrictions to what collection of information actually constitutes a database. In particular, data need not be organized with the primary purpose of creating a database. It is sufficient that the data is organized and individually accessible and separable, even if the creation of the database was incidental to the actual purpose of, for example, accomplishing a certain result with the individual data⁵⁵.

It has been argued, that the broad definition awarded to databases is at odds with the Database Directive's purpose of encouraging investments in information storage and processing systems to develop the information market within the EU⁵⁶, because even biographies might be considered databases under the Database Directive⁵⁷.

This view seems to forget the requirement that individual data be separable from the whole without impacting their individual value – a quite powerful restrictor. For example, while a biography can indeed be seen as a collection of individually accessible data (looking either at words or chapters), individual words cannot be separated without substantially losing their value, and individual chapters probably only in exceptional circumstances. Consequently, classifying a biography as a database seems a stretch⁵⁸.

Finally, while the definition of a database is broad, the Database Directive does not automatically award protection to all databases simply because they are databases. Only databases that fulfill certain requirements as to the effort that went into producing the database qualify for protection.

2.2.2 OBJECT OF SUI GENERIS PROTECTION

Sui generis protection to databases is granted in recognition of the fact that constructing a database requires *“investment of considerable human, technical and financial resources”*⁵⁹. The directive aims to reward and protect such investment by providing the maker of a database with a sui generis database right that places him in a position to prevent unauthorized access and copying of the database contents, which he compiled. In this regard, Article 7 Database Directive states:

“Member States shall provide for a right for the maker of a database which shows that there has been qualitatively and/or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents to prevent extraction and/or re-utilization of the whole or of a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database.”

⁵⁵ Davison, supra, p. 70 et seq.

⁵⁶ Recital 7 et seqq. of the Database Directive.

⁵⁷ See Davison, p. 72.

⁵⁸ See also Estelle Derclaye, *The Legal Protection of Databases*, p. 62 et seq. (2008).

⁵⁹ Recital 7 Database Directive.

Consequently, the Database Directive demands that “*there has been qualitatively and/or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents*”⁶⁰. The type of investment required can be time, financial resources, personnel, or technical means invested, or indeed any other “sweat of the brow”-type resource, as distinct from creative, intellectual efforts.

The CJEU is very strict in its understanding that the investment must be made to *obtain* the contents. A database that is a mere spinoff/by-product from another investment/activity (such as scientific data resulting from research) does not typically qualify for protection under the Database Directive’s *sui generis* regime. There must additionally be a further substantial investment in obtaining, verifying or presenting the data⁶¹.

In other words, the CJEU demands that the investment be made specifically to “*seek out existing independent materials and collect them in the database*”⁶². An investment in “*the creation of materials which make up the contents of a database*”⁶³ is deemed insufficient. As a result, creators of data rarely enjoy a *sui generis* right of protection for any non-original database constructed out of that data – so-called “single source databases”⁶⁴ – unless there is also a substantial investment in the verification or presentation of the contents.

Such an additional substantial investment is often the case for data resulting from clinical trials. Such data must first undergo an extensive verification process before it can be used in research and entered into a database. Importantly, the data verification process is subsequent and separate from the obtaining/creation of the original data, as otherwise it would be excluded from protection.

“*Verification*” is understood to mean steps taken to ensure the information is reliable. As with the requirement of “obtaining” data, an investment in verifying information during the information’s creation is excluded⁶⁵.

⁶⁰ Article 7(1) Database Directive.

⁶¹ See Estelle Derclaye, *The Legal Protection of Databases*, pp. 92 et seqq. (2008).

⁶² Case C-203/02 *The British Horseracing Board Ltd and Others v William Hill Organization Ltd.*, para 42.

⁶³ Case C-203/02 *The British Horseracing Board Ltd and Others v William Hill Organization Ltd.*, para 42. See also Jasper A. Bovenberg, *Property Rights in Blood, Genes & Data: Naturally Yours?* p. 176 et seq. (2006).

⁶⁴ See European Commission, DG Internal Market and Services Working Paper – First evaluation of Directive 96/9/EC on the legal protection of databases, 2005, p. 14, available at:

http://ec.europa.eu/internal_market/copyright/docs/databases/evaluation_report_en.pdf.

⁶⁵ See Estelle Derclaye, *The Legal Protection of Databases*, p. 97 (2008).

“*Presentation*” is defined as the way data is structured and made accessible to others, so that the creation of an index or the design of a user interface can all be seen to fulfill the requirements of an investment in the presentation of the contents⁶⁶.

Finally, the investment must also be of a “*qualitatively and/or quantitatively*” substantial nature⁶⁷. The Database Directive does not define “*substantial*” and neither has the CJEU ruled on the matter. However, the Preamble of the Directive indicates that, “*as a rule, the compilation of several recordings of musical performances on a CD (...) does not represent a substantial enough investment to be eligible under the sui generis right*”⁶⁸. Member States generally adopt a low level approach to the requirement, and the Advocate General has taken the same stance⁶⁹.

As regards the quantitative and/or qualitative qualification, these are understood to mean investments quantifiable and not-quantifiable, respectively, such as money on the one hand and intellectual effort on the other⁷⁰.

2.2.3 SCOPE OF PROTECTION AND EXCEPTIONS

The protection extends against unauthorized extraction and re-utilization of a substantial part, evaluated qualitatively and/or quantitatively, of the contents of the database⁷¹, or the repeated and systematic use of insubstantial parts that infringes upon the legitimate interests of the database maker⁷².

The CJEU has held that a substantial part, evaluated quantitatively, refers to the volume of data in relation to the database as a whole⁷³. Evaluated qualitatively, a substantial part refers to the scale of investment in obtaining the contents⁷⁴.

“*Extraction*” is defined as “*the permanent or temporary transfer (...) to another medium by any means or in any form*”⁷⁵. Extraction does not require physical copying of contents, and neither does it require the

⁶⁶ See Estelle Derclaye, *The Legal Protection of Databases*, p. 97 et seq. (2008).

⁶⁷ Article 7(1) Database Directive.

⁶⁸ Recital 19 Database Directive.

⁶⁹ See Estelle Derclaye, *The Legal Protection of Databases*, pp. 75 et seqq. (2008).

⁷⁰ See Case C-338/02 *Fixtures Marketing Ltd v Svenska Spel AB*, para 28.

⁷¹ See Article 7(1) Database Directive.

⁷² See Article 7(5) Database Directive.

⁷³ See Case C-203/02 *The British Horseracing Board Ltd and Others v William Hill Organization Ltd.*, para 70.

⁷⁴ *Ibid*, para 71.

⁷⁵ Article 7(2)(a) Database Directive.

intention of creating a competing product⁷⁶. For example, consulting a protected database, assessing its contents individually and then using it for the creation of a new database that expands upon the contents of the first database can constitute extraction. Specifically, the CJEU held that consulting a database of poetry titles in order to create a database of poems can constitute extraction⁷⁷.

“Re-utilization” is defined as “any form of making available to the public (...) by the distribution of copies, by renting, by on-line or other forms of transmission”, albeit subject to the principle of exhaustion⁷⁸.

As regards the above mentioned repeated and systematic extraction/reutilization of insubstantial parts, the CJEU has held that such use is prohibited, “the cumulative effect of which would be to seriously prejudice the investment made by the maker of the database just as the extractions and/or re-utilisations”⁷⁹ – which is the case where either the database as a whole or in substantial parts is basically reconstituted by the systematic and repeated efforts of the user⁸⁰.

In parallel, protection is granted without prejudice to any rights residing in the contents themselves⁸¹, but with certain exceptions possible, such as for scientific research⁸². The details are the same as *supra*⁸³.

As regards the scope of the database right, it would protect the collected data from being copied (extracted) as a whole or in substantial part, evaluated “qualitatively and/or quantitatively”⁸⁴ and either conducted in one action or step by step⁸⁵. In this respect, the database right does not protect the data or isolated datasets per se; Article 7 (4) makes explicit that database right: “shall apply irrespective of eligibility of the contents of that database for protection by copyright or by other rights. Protection of databases [...] shall be without prejudice to rights existing in respect of their contents”.

Thus, the database right may be an option to protect the data as a whole or in substantial part, but it will not extend to protect isolated data items. Therefore, if it is the datasets per se, or the investment

⁷⁶ See Diane Rowland, Uta Kohl and Andrew Charlesworth, *Information Technology Law*, p. 390 et seq. (4th ed, 2012).

⁷⁷ See Case C-304/07 *Directmedia Publishing GmbH v Albert-Ludwigs-Universität Freiburg*, paras 9 et seqq.

⁷⁸ See Article 7(2)(b) Database Directive.

⁷⁹ Case C-203/02 *The British Horseracing Board Ltd and Others v William Hill Organization Ltd.*, para 86.

⁸⁰ See Case C-203/02 *The British Horseracing Board Ltd and Others v William Hill Organization Ltd.*, para 87.

⁸¹ See Article 7(4) Database Directive.

⁸² See Article 9 Database Directive.

⁸³ See *supra*, section 3.2.2.2.

⁸⁴ See Article 7 (1) Database Directive.

⁸⁵ See Article 7(5) Database Directive.

deployed in processing the data, which are to be protected, the database right would not provide such protection.

2.2.4 TERM AND BENEFICIARIES OF PROTECTION

The *sui generis* right begins upon completion or being made publically available of the database and runs for 15 years from the first of January of the year following completion; any substantial change or accumulated changes to the database that require a substantial new investment renew the term⁸⁶. There is no limit to the number of renewals possible, so that databases that require ongoing and substantial updating are effectively protected for as long as they are kept alive.

The *sui generis* right is extended to natural persons who are makers of or right holders to databases and are nationals or habitual residents of a Member State. Legal persons must be formed according to the laws of a Member State and also have their registered office, central administration or principal place of business within and a genuine link to the economy of a Member State⁸⁷.

The maker of a database is seen as “*the person who takes the initiative and the risk of investing*”⁸⁸, but excluding subcontractors.

This definition means that databases made by the State arguably fall outside the scope of *sui generis* protection. Because taxpayer’s money is used, the State does not undertake a financial risk when making databases⁸⁹ and is therefore not the maker of a database according to the Database Directive. Nevertheless, national courts have granted State-made databases protection, albeit before CJEU decisions on the requirement of truly “obtaining” data⁹⁰.

2.2.5 CONCLUSIONS FOR SOBIGDATA

As noted above, there is no central data warehouse in SoBigData. The datasets, available for SoBigData, are typically stored and maintained by the parties at their premises/servers. These datasets normally have been generated, are arranged and managed by the parties themselves according to the volume, type and format of the datasets at issue. Once a particular dataset is called by the user, this dataset is supposed to

⁸⁶ See Article 10 Database Directive.

⁸⁷ See Article 11(1) and 11(2) Database Directive.

⁸⁸ See Recital 41 Database Directive.

⁸⁹ See Estelle Derclaye, *The Legal Protection of Databases*, p. 74 (2008).

⁹⁰ *Ibid.*

be retrieved from the server of the respective party, who provides the dataset at issue. The SoBigData infrastructure is not supposed to integrate a central repository, aggregating all the datasets, provided by the project parties.

In light of a big number of SoBigData datasets, which are maintained by the parties themselves, the applicability of sui generis protection to the parties repositories depends on how the data has been collected and how the data is arranged and organized and how much investment the respective party deployed into collecting, verifying and presenting the data. It is in the responsibility of each RI-Partner to assess the data collection and to determine whether such data collection qualifies as a database in the meaning of Database Directive and satisfies the necessary requirements for the sui generis rights. The above legal considerations shall serve as guidelines for assessing the protectability of the repositories by sui generis database rights.

Provided the repository of the party qualifies as database in the meaning of the Database Directive and the party holds the sui generis database rights, this party may stipulate the terms of using the repository as a whole, grant the rights of use under contractual license, prevent and enforce the unauthorized extraction/reutilization of the repository contents as a whole or in substantial part. The party, who holds the sui generis rights, may leverage how the contents of its repository may be used, whether the data items may be extracted and in what scope, whether the data may be transferred to the user or whether the analysis should be done by the RI-Partner holding the right or another RI-Partner to whom the data gets transferred.

However, the sui generis protection applies to the contents of the repository as a whole or in substantial part and may apply separately and irrespective of protectability of data items by other rights, such as copyrights. Thus, the party may manage the use of the repository as a whole. However, the use of separate data items, protected by other rights, may remain governed by the terms, stipulated by the holders of rights in such items. Once, the SoBigData RI allows the use of some data items, the processing of such data may in principal be subject to their individual terms of use and does not necessarily depend on whether the repository, where such data item is maintained, is covered by sui generis rights or not. Therefore, if it is the datasets per se, or the investment deployed in processing the data, which are to be protected, the database right would not provide such protection.

2.3 PROTECTION OF DATA AS KNOW-HOW

A further means of affording protection to the SoBigData data may be the rules dealing with confidential undisclosed information as 'know-how'. Such protection is provided by Section 7, Article 39 et seq. TRIPS Agreement⁹¹, and enables natural and legal persons to prevent "*information lawfully within their control from being disclosed to, acquired by, or used by others without their consent in a manner contrary to honest commercial practices*"⁹². Unfair practices for these purposes would include the acquisition of

⁹¹ Agreement on Trade-Related Aspects of Intellectual Property Rights, the TRIPS Agreement, Annex 1C of the Marrakesh Agreement Establishing the World Trade Organization, Marrakesh, Morocco, 15 April 1994.

⁹² Article 39 (2) TRIPS Agreement.

information via violation of contractual duties, breach of confidentiality obligations, inducement to breach, etc.⁹³

In order to be protectable, the relevant information should have the quality of protectable information within the meaning of Article 39 TRIPS Agreement. Article 39 (2) TRIPS Agreement provides protection to information, which:

“(a) is secret in the sense that it is not, as a body or in the precise configuration and assembly of its components, generally known among or readily accessible to persons within the circles that normally deal with the kind of information in question;

(b) has commercial value because it is secret; and

(c) has been subject to reasonable steps under the circumstances, by the person lawfully in control of the information, to keep it secret.”

The first weak point of protecting the processed data as know-how is that as of now the legal framework on know-how protection in the EU is not harmonized. Although, there is a proposal for a draft directive on the protection of undisclosed information in the EU⁹⁴ (the Draft Directive), before it is adopted and implemented, protection of know how remains dispersed through the national states of the EU Member States, and subject to varying requirements for and scope of protection⁹⁵.

The Draft Directive, which is intended to harmonize the national laws in relation to know-how protection, in many aspects repeats the provisions of the TRIPS Agreement (in particular, it relates to the protectable subject matter and requirements for protection (Article 2), acts of unfair acquisition of information (Article 3), rights and remedies conferred (Article 5 et seq), etc.). In this regard it may also be queried how far the Draft Directive, if adopted, would improve the protection for data, the preparation of which consumed much effort, but which for one or another reason may not reach the level of protectable know-how. Here the key obstacles in applying know-how protection to the processed data in research relate to the need (in order to be protected) for such data to be secret, subject to the confidentiality measures and have economic value.

Thus in SoBigData, as noted, there is a large mass of research data, comprising multiple data types, formats, words, figures, numerical parameters, abbreviations, etc. In order to satisfy the requirements for know-how protection, this data must first be secret and treated as such and, second, it must possess economic value. To satisfy the criterion of secrecy, the information, sought to be protected, must not be available and/or easily accessible to the broader public. The use of such information must be subject to

⁹³ Footnote 10 to Article 39 TRIPS Agreement.

⁹⁴ Proposal for a Directive on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure, Brussels, 28.11.2013, COM(2013) 813 final, 2013/0402 (COD).

⁹⁵ Hogan Lovells International LLP, Report on Trade Secrets for the European Commission, Study on Trade Secrets and Parasitic Copying (Look-alikes), MARKT/2010/20/D.

confidentiality measures. The application of confidentiality measures means that the data must be stamped as “Confidential” and the sharing of such data must be made upon non-disclosure obligation. Disclosure of such datasets via Internet would by contrast destroy the regime of secrecy so that protection would be forfeited. As regards the requirement of economic value of know-how, this will be considered to be present if through publication, the research investment and competitive standing of the entity doing the work would be undermined⁹⁶.

In relation to the volumes of research data in SoBigData, this requirement, besides being at odds with the underlying culture of academic research, would create further workload. The data subject to the regime of confidentiality must first be strictly identified. The confidentiality mark would need to be attached to the data and any use and disclosure of such data to any third party (either RI-Partner or the user) must be made upon signing the non-disclosure agreement. This preservation of the confidentiality mark, conclusion of NDA and control over handling such data as confidential would present another challenge.

Against this background, know-how protection might, in principle, be considered as a possibility for protecting some defined amount of data, but hardly makes a feasible solution, when protection of large amounts of data, processed in SoBigData is sought. It also may operate against the principle of openness, if optimal use is to be made of the data by the research community, exploiting the full potential of available datasets.

2.4 OWNERSHIP OF IP RIGHTS

The intellectual property law protects the results of intellectual activity or investment, deployed in creating socially beneficial goods, be they writings, inventions, music compositions, etc. Generally speaking, intellectual property law aims to remunerate creators and other producers of intellectual goods and services with exclusive time-limited rights to control the use of those productions. The IP rights do not apply to the physical object in which the creation may be embodied but instead to the intellectual creation as such⁹⁷.

The distinction between ownership of IP rights in contrast to physical objects is well established in the law.

Starting with the International law first, the Universal Declaration of Human Rights⁹⁸ treats ownership (Article 17) separately from copyrights and the right to participate in the cultural life (Article 27 (2)):

⁹⁶ K. Lodigkeit, *Intellectual Property Rights in Computer Programs in the USA and Germany*, Peter Lang GmbH, 2006, pp. 98-101.

⁹⁷ WIPO, *The Concept of Intellectual Property*, available at: <http://www.wipo.int/export/sites/www/about-ip/en/iprm/pdf/ch1.pdf> 10.08.2016.

⁹⁸ UN Universal Declaration of Human Rights of 10.12.1948.

“Everyone has the right to the protection of the moral and material interests resulting from any scientific, literary or artistic production of which he is the author”. The UN International Covenant on Economic, Social and Cultural Rights⁹⁹ reflects this principle in Article 15 (1) (c): “The States Parties to the present Covenant recognize the right of everyone...to benefit from the protection of the moral and material interests resulting from any scientific, literary or artistic production of which he is the author.” In both cases, the protection of the moral and material interests in the works resulting from intellectual activity is linked to the authorship, i.e. production in the literary, scientific and artistic domain¹⁰⁰.

Proceeding to the European law, the notion of Intellectual Property appears for the first time in the EU Charter of Fundamental Rights¹⁰¹. In Article 17 (2), the Charter simply declares: “Intellectual property shall be protected.” By this, the Charter makes clear that the provisions of Article 17 (1) apply to Intellectual Property as well¹⁰². Article 17 (1) states: “Everyone has the right to own, use, dispose of and bequeath his or her lawfully acquired possessions.” In transposition to the Intellectual Property law, it means that everyone has the right to own, use, dispose of and bequeath his IP at his own discretion.

However, Intellectual Property constitutes an atypical type of ownership. Distinction between ownership over a manuscript, where opinions of the author are incorporated, and ownership over expressed opinions of the author was made by a number of prominent scientists in the 1780-ies¹⁰³. And it is the author of the opinions, and not the possessor of a manuscript, who owns a “natural, inborn and inalienable right”¹⁰⁴ over the expressed opinions. A work with the expressed opinions of the author as a product of his intellectual activity constitutes an object of IP protection; the ownership over these opinions constitutes the right, attributable to the creator of the work alone¹⁰⁵.

⁹⁹ UN International Covenant on Economic, Social and Cultural Rights, Adopted and opened for signature, ratification and accession by General Assembly resolution 2200A (XXI) of 16 December 1966, entry into force 3 January 1976, in accordance with article 27.

¹⁰⁰ Article 2 Berne Convention; Jakob Schneider, Menschenrechtlicher Schutz geistigen Eigentums, 2006; Haimo Schack, Zur Rechtfertigung des Urheberrechts als Ausschliesslichkeitsrecht, in Geistiges Eigentum: Schutzrecht oder Ausbeutungstitel? Springer, Volume 5, 2008.

¹⁰¹ CHARTER OF FUNDAMENTAL RIGHTS OF THE EUROPEAN UNION 2012/C 326/02, OJEU, C 326/391, 26.10.2012.

¹⁰² Hans D. Jarass, EU-Grundrechte, 2005, S. 485 ff; EuGHE 1998, I-1953 Rn. 21 = GRUR Int.

1998, 596 – Metronome Musik/Music Point Hokamp.

¹⁰³ Immanuel Kant, Von der Unrechtmäßigkeit des Büchernachdrucks, 1785, Nachgedruckt in UFITA 106 (1987) 137-144, mit einer Einführung von Hubmann 145-154; Schack, supra; *Le Chapeliers* in den Beratungen des ersten französischen Urhebergesetzes von 1791, See: *André Bertrand*, Le droit d’auteur, Paris 1991, S. 47; *Schack*, UrhR, Rn. 100.

¹⁰⁴ Fichte, Beweis der Unrechtmäßigkeit des Büchernachdrucks, 1793, nachgedruckt in UFITA 106 (1987) 155-172, 163; Schack, supra.

¹⁰⁵ Schack, supra.

An approach, taken by Herbert Zech, to divide the information into three layers: semantic, syntactic and structural¹⁰⁶ may support and clarify this distinction better. This division resides on differentiation of information into content layer, code layer and physical layer, represented by Benkler¹⁰⁷ and Lessig¹⁰⁸. The semantic layer of Zech corresponds to the contents, which the information bears (e.g. expressed opinions), syntactic layer is the level of symbols, in which such information is recorded (e.g. the text of the manuscript), the structural layer would be the tangible medium, where such information is embodied (e.g. manuscript per se). The primary goal here is to categorize the existing legal rights in relation to the various informational goods¹⁰⁹.

Building on this, a starting point for determining the holder of IP rights would be to establish whether some dataset qualifies as an object protected by IP rights, what type of IP applies and to define the holder of IP rights according to the origin of a work and the applicable law, respectively.

The relations of ownership with regard to works, protected by copyrights, are normally governed by the copyright laws. According to the general principle of copyright, original copyright passes to the creator. It means, the person, who created a copyrighted work, typically acquires copyrights in this work, unless this principle is overridden by the law or by the contract. For example, in case of computer software, created in employment relation, software copyrights pass to the employer¹¹⁰. Who owns copyrights in works, created in employment relation, depends on the principles of the national copyright law. For example, the German copyright law attributes copyrights in a work, created in employment relations, to the author. This principle may also be derogated from by the contract¹¹¹.

According to these general principles of copyright, in circumstances, where the user creates a copyrighted work using the SoBigData facilities, the copyright would normally pass to that user. It may be demonstrated by the example of writing a thesis using Microsoft word or taking a picture by Nikon camera. In both cases, copyrights in a thesis and image rights in a picture would pass to the user alone. Neither Microsoft nor Nikon may claim copyrights. What requirements the works must sustain in order to be protected by copyright, we considered in Section 2.1.2 above. The legal requirements for application of sui generis database rights and the beneficiaries of such protection we described in Section 2.2 above.

¹⁰⁶ Zech, Daten als Wirtschaftsgut – Überlegungen zu einem „Recht des Datenerzeugers“, CR 2015, 141; Hoeren/Völkel in Big Data und Recht, S. 20.

¹⁰⁷ Y. Benkler, From Consumers to Users: Shifting the Deeper Structures of Regulation.

¹⁰⁸ L. Lessig, The Future of Ideas, The Fate of the Commons in a Connected World, New York 2002.

¹⁰⁹ H. Zech, supra, p. 2.

¹¹⁰ Article 2 (3) Directive 2009/24/EC of 23 April 2009 on the legal protection of computer programs, OJEU L 111/16, 5.5.2009.

¹¹¹ Article 43 Act on Copyright and Related Rights (Copyright Act) of Germany.

3 RESPONSIBLE IP PRINCIPLES

As noted in Section 2 above, the SoBigData RI is supposed to aggregate a large amount of data from various sources. A number of datasets are already available. This data includes call graphs from mobile phone call data, networks crawled from many online social networks, including Facebook and Flickr, transaction micro-data from diverse retailers, query logs both from search engines and e-commerce, societywide mobile phone call data records, GPS tracks from personal navigation devices, survey data about customer satisfaction or market research, billions of tweets, and data from location-aware social networks. The dataset catalog is available at: <http://www.sobigdata.eu/private/ouhnr3inebchcerhncurhebcn8749892/datasets>.

The preliminary analysis conducted on the SoBigData datasets explored - to the extent possible within the given descriptions - the substance of the datasets, the potential application of IP rights (e.g. copyrights, image rights, related rights, database rights), the terms of use (license), whether special requirements apply to the sharing of such datasets, the basic rights, which are supposed to go to the user, etc. The results of this analysis, reflecting the data types, applicable IP rights and terms of use for each of 63 datasets are accumulated in table, attached as Appendix B to this deliverable.

For instance, a number of SoBigData datasets, like Aalto-Twitter, No 4, Archive Twitter Dataset, No 5, Italian Twitter Dataset, No 30, make use of the data of the social platform Twitter. Twitter content may be protected by copyright and data protection law and reside within proprietary rights of Twitter. Use of the Twitter content is subject to the Twitter Terms of Use¹¹². And when the Twitter content is integrated from Twitter via Twitter API, then the use of the Twitter content by a third party service (connecting to Twitter and retrieving Twitter content via Twitter API) is governed by the Twitter API license agreement (Twitter Developer Agreement)¹¹³. Twitter API license agreement makes a binding contract between Twitter, acting as the licensor, and a third party developer, acting as a licensee, and governs the access to and use by a developer of the Licensed Material from Twitter, which includes Twitter API and Twitter Content. Acceptance of the API license agreement by a developer is a precondition for obtaining access to and using Twitter API. A third party developer, who makes use of the Twitter API for connecting his service to Twitter for data sharing, agrees that he has read and agrees to "COMPLY WITH AND TO BE BOUND BY THE TERMS AND CONDITIONS OF THIS AGREEMENT AND ALL APPLICABLE LAWS AND REGULATIONS IN THEIR ENTIRETY WITHOUT LIMITATION OR QUALIFICATION."¹¹⁴ Against this background, a RI-Partner, who retrieves Twitter content via Twitter API, is bound by the Twitter API license agreement and agrees to use the Twitter content in accordance with the terms, stipulated by Twitter.

Provided the Twitter data is collected via Twitter API, then according to the Twitter Developer Agreement, Section I B, the developer, who integrates Twitter API for data sharing into his services, may: "2. Copy a

¹¹² <https://twitter.com/tos?lang=en> <https://dev.twitter.com/overview/terms/agreement>
https://dev.twitter.com/overview/terms/policy#6.Update_Be_a_Good_Partner_to_Twitter

¹¹³ Twitter developer agreement, available at: <https://dev.twitter.com/overview/terms/agreement>

¹¹⁴ Ibid.

reasonable amount of and display the Content¹¹⁵ on and through your Services¹¹⁶ to End Users¹¹⁷, as permitted by this Agreement;... 3. Modify Content only to format it for display on your Services;”

In addition to the Twitter API license agreement, Twitter provides some guidelines and rules for developers, “*who interact with Twitter’s ecosystem of applications, services, website, web pages and content (“Twitter Services”). Policy violations are also considered violations of the Developer Agreement.*”¹¹⁸ According to the Twitter Developer Policy, Section I.F “Be a good partner to Twitter”, a developer (referred by Twitter as “you”) agrees to the following:

“2.If you provide Content to third parties, including downloadable datasets of Content or an API that returns Content, you will only distribute or allow download of Tweet IDs and/or User IDs.

a.You may, however, provide export via non-automated means (e.g., download of spreadsheets or PDF files, or use of a “save as” button) of up to 50,000 public Tweets and/or User Objects per user of your Service, per day.

b.Any Content provided to third parties via non-automated file download remains subject to this Policy.”¹¹⁹

In this constellation, the user has a right to use the Twitter content as part of/in course of using SoBigData services. Any subsequent use of the Twitter data is subject to the terms of Twitter. For further details please see Appendix B. Appendix B provides an initial overview of the SoBigData datasets, applicable rights and terms of use. As of the current stage of the project development and use cases, analysed so far¹²⁰, the usages expected to be made of SoBigData datasets within the project, which are possibly in breach with the licence terms have not been identified.

As the analysis of the SoBigData catalog shows, most data resources have been collected from external sources (which in some cases may not be existing/be available anymore) and/or generated by the RI-Partners themselves. Against this background, these are the RI-Partners, who provide the data to the SoBigData RI, who possess the necessary access to the information about the source, from where the data has been collected, about the means, how such data was collected (e.g. crawling, API, copy and paste, procession by analytical tools, etc.) or generated (e.g. output of data procession, raw data, etc.), the legal basis of such data collection and/or generation and last but not least about the terms, under which such data has been collected/generated and may be provided into use.

¹¹⁵ Tweets, Tweet IDs, Twitter end user profile information, and any other data and information made available to you through the Twitter API or by any other means authorized by Twitter, and any copies and derivative works thereof.

¹¹⁶ Your websites, applications and other offerings that display Content or otherwise use the Licensed Material.

¹¹⁷ Users of your Services, who are bound to enforceable service terms with you.

¹¹⁸ Twitter Developer Policy, available at: https://dev.twitter.com/overview/terms/policy#6.Update_Be_a_Good_Partner_to_Twitter

¹¹⁹ Ibid.

¹²⁰ SoBigData, Exploratories & Stories, available at: <https://support.d4science.org/projects/sobigdata-eu/wiki>.

As of the stage of the preliminary legal analysis of SoBigData datasets and applicable legal terms, represented in this Deliverable, the RI-Partners, who keep/make the data available for the SoBigData project, have been instructed to investigate the terms, applicable to the datasets, which they provide, and define the usages, which are allowed to be carried out with their data in the SoBigData project. LUH, with the technical and organizational support of TUDelft, CNR, UNIPI, will follow the process of providing the SoBigData datasets with the applicable terms of use, and will provide the legal support in this process, as may be required. On its part and to the extent possible within the information available in SoBigData, LUH will conduct the legal analysis of SoBigData use cases against compliance with the terms, governing the data used in the use cases concerned, and will report the findings in the deliverables of Task 2.4.: D.2.4 IP principles and business models (M12, M24, M36 Type: R)¹²¹ to follow. If SoBigData is aimed to be provided as a transparent infrastructure, the IPR, applicable to the datasets and simple rights and terms of use for each dataset should be clear to the user. Compliance with the rights of third parties makes one of the most important IP principles, which need to be maintained within the SoBigData platform. As one of the approaches how to manage the IP issues in SoBigData it is suggested to integrate the necessary information regarding the IPR into the metadata of each dataset. The definition of appropriate description/metadata to the SoBigData datasets is to be carried out within WP10, T10.1 e-infrastructure interoperability. Task 10.1 provides for *“Definition of the e-infrastructure resource model: types of resources (e.g. datasets, services, algorithms) and metadata properties of the resources (e.g. classifications, access parameters, hardware parameters, IPR issues, input/output data types).”*¹²² These activities shall last over the project lifetime and interrelate with WP8 and WP9, *“starting from their current status in existing national infrastructures (see WP7) and different research areas in order to align the community towards common and agreed-upon best practices. Task leader: CNR. Participants: ALL PARTNERS”*. The integration will be carried out within Task 10.2. Integration to the e-infrastructure.¹²³

It is recommended that the IP metadata should reflect the type of IPR, which may be applicable to the dataset at issue, the source, where the dataset originates from, license (terms of use) and link to the license, basic user rights, requirements applicable for display and distribution, restrictions on use, if any, etc. The list of IP-metadata, recommended for integration into the data management in SoBigData is provided in Appendix A. The IP metadata should be completed by the partner, who provides the data to the SoBigData platform and has the necessary knowledge and access to the information about the data that he provides.

The legal analysis on IPR issues in the SoBigData project is ongoing and will be refined in subsequent deliverables, producible in WP 2, T. 2.4 Intellectual property and data ownership framework T 2.4 is responsible for the deliverables *“D.2.4 IP principles and business models (M12, M24, M36 Type: R). Report describing responsible IP principles and practices in SoBigData and listing of best practices that have evolved in and outside SoBigData so as to accommodate both legitimate IP claims in Data and privacy related rights.”*¹²⁴

¹²¹ SoBigData Proposal/DOW, footnote 1, supra, WP2, T.2.4, p. 44 et seq.

¹²² Ibid, WP10, p. 72 et seq.

¹²³ Ibid.

¹²⁴ Ibid, WP2, p. 44 et seq.

4 DATA OWNERSHIP ISSUES

In this section we elaborate on the issue of data ownership and focus on the questions of to whom belongs the data generated in the SoBigData research project and who owns the personal information.

4.1 A NOTION OF DATA OWNERSHIP

Data have become an important economic good. There are numerous business models that base on the systematic collection and processing of (personal) data for the purpose of selling the data. It is hoped through analysis of the data to discover new correlations which can be commercially exploited. There is a growing interest in an allocation of data in the sense of “ownership”.

The law typically understands (in simplified terms)¹²⁵ ownership as the societally recognized right of a person (legal or natural) to exert exclusive control over a thing¹²⁶. However, things typically do not emanate their state of ownership. Simply looking at an object will not tell you who the owner is. In order to facilitate legal transactions, both the common law¹²⁷ and the civil law¹²⁸ have variations of the idea that possession, for example physical control, is generally indicative of ownership.

However, possession of data is not as clearly definable as is possession of chattels. Indeed, the intangible nature of data or information means that traditional approaches to ownership cannot be easily applied to it. A tangible asset, for example a hammer, is fully rivalrous – it can only be used in a zero-sum fashion. This is in contrast to intangible assets such as data, which can be replicated without limit and are non-rivalrous, i.e. which can principally be used by more than one person simultaneously without their

¹²⁵ This simplification ignores that society does not recognize any absolute property right, as becomes clear for example during a criminal investigation, where the state can seize property. See also article 14 paras 1 and 2 of the German Basic Law (Grundgesetz): “(1) Property and the right of inheritance shall be guaranteed. Their content and limits shall be defined by the laws. (2) Property entails obligations. Its use shall also serve the public good.” See also Barbara J Evans, ‘Much ado about data ownership’, (2011) 25 Harv. J. L. & Tech. 79 et seqq, p. 69-130.

¹²⁶ See the following description of how people imagine property to be: “There is nothing which generally strikes the imagination, and engages the affections of mankind, as the right of property; or that sole and despotic dominion which one man claims and exercises over the external things of the world, in total exclusion of the right of any other individual in the universe.” William Blackstone, Blackstone’s Commentaries on the Laws of England (first printed in 1765), .P 2 The Rights of Things. Book II. Ch. i. Accessible at http://avalon.law.yale.edu/18th_century/blackstone_bk2ch1.asp.

¹²⁷ Carol M. Rose, ‘Possession as the Origin of Property’, (1985) Faculty Scholarship Series, Paper 1830, see http://digitalcommons.law.yale.edu/fss_papers/1830, p. 74 et seqq.

¹²⁸ E.g. section 1006 of the German Civil Code (Bürgerliches Gesetzbuch).

usefulness being affected¹²⁹. This can be particularly seen where data is stored in the cloud. A multitude of parties may have (simultaneous) access to data stored in the cloud.

Even though basically all jurisdictions have adopted¹³⁰ so-called intellectual property rights such as copyright, patents, and trademark, which grant a type of control over certain intangibles, the data per se, where the protected objects may be encoded (such as patent description) may not fall under the IPR regime¹³¹.

Another important aspect is to define the terms “information” and “data” more closely, as part of determining what the putative object of legal protection should be. This would serve as an important clarificatory step, as currently different areas of science and law operate with diverse terms. Thus, when it comes to personal data, the law on data protection operates with the term “data”¹³². The criminal law, dealing with information in the context of computer technologies, speaks about “computer data”, meaning “*any representation of facts, information or concepts in a form suitable for processing in a computer system, including a program suitable to cause a computer system to perform a function*”¹³³. By contrast, intellectual property law and the competition law, when protecting know how, use the term information¹³⁴. Dreier suggests understanding the “*information in the sense closer to “data” than to knowledge*”¹³⁵ and highlights the role of IP rights as “legal rules which [...] regulate the generation, flow, storage and use of information”.

4.2 GOVERNING USE OF DATA BY CONTRACTUAL RELATIONS

The control over data, which is available or processed in the SoBigData RI, needs to be considered in the complex of factors and legal relations, which surround the generation of data. Such factors include the materials, with and on which the research producing the data has been conducted, and the rights thereto,

¹²⁹ Barbara J Evans, ‘Much ado about data ownership’, (2011) 25 Harv. J. L. & Tech. 78, p. 69-130.

¹³⁰ See as a proxy for this statement the list of member states of the World Intellectual Property Organization, available at: <http://www.wipo.int/members/en/>.

¹³¹ Gilad Rosner, ‘Who owns your data?’ in Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp '14 Adjunct). ACM, New York, NY, USA (2014), 623-628, 625. DOI=10.1145/2638728.2641679, see <http://doi.acm.org/10.1145/2638728.2641679>.

¹³² Article 2 Data Protection Directive.

¹³³ Council of Europe, CONVENTION ON CYBERCRIME, Budapest, 23.XI.2001.

¹³⁴ Article 39 TRIPS Agreement; Commission Regulation (EU) No 316/2014 of 21 March 2014 on the application of Article 101(3) of the Treaty on the Functioning of the European Union to categories of technology transfer agreements Text with EEA relevance OJ L 93, 28.3.2014, p. 17–23.

¹³⁵ T. Dreier, ‘Regulating information: Some thoughts on a perhaps not quite so new way of looking at intellectual property’

the legal and contractual relations, under which the work, generating the data has been performed, the contractual regulations over the results of such work, etc.

“Ownership” may not be considered as an absolute right in relation to data. The legal relations in respect of data outputs and research results are therefore regulated by contracts between the parties. The term “ownership”, used in the literature or some legal documents, does not create an absolute right for the party of the contract referred to as the “owner”.

In the context of research projects, the materials, which serve the purposes of research (e.g. instruments, computers, etc.) and which have been acquired from the budget costs of the universities normally pass into the ownership of the state (or federal state or province) as a carrier of the university. However, the ownership over the research materials does not automatically lead to the ownership of research results¹³⁶. Ownership may also be acquired via processing.

The research projects are normally conducted by educational or research institutions and the research is typically done by the research associates, who are bound by the employment relations with the institution. The attribution of ownership over the research results occurs via the employment relations. Usually, the research associates conduct the research on the materials of the institution and achievement of scientific results in dependent position belongs to their employment obligations. In such circumstances, the researcher receives remuneration for the work he does, the institution acquires the ownership and also the exploitation right over the achieved results, provided otherwise is not stipulated by the contract¹³⁷.

Provided the research results achieve the IP protection by copyrights or database rights, the ownership over such rights needs to be considered in the context of IP law, as we described in Section 2.4 above.

Students or PhD students, who produce some research results under a membership relation to the university, do not have an obligation to create some type of ownership and are not obliged to pass the ownership to the supervisor. In this constellation, the respective student owns the results of his work. In contrast, the PhD students, who are bound by the employment relations with the university and do the research by order of the university, fall under the regulation of ownership relations in employment, considered above. Thus, the ownership over research results, achieved by the PhD student in the employee position, would normally pass to the institution.¹³⁸

In third party funded projects, the relations of ownership over the research results are typically governed by the contract. The sponsor is typically interested to exploit the project results and funding is typically granted upon condition that the sponsor acquires the ownership and exploitation rights over the research

¹³⁶ H.-D. Lippert, Wem gehören Daten, die im Rahmen von Forschungsprojekten gewonnen werden? in Geistiges Eigentum: Schutzrecht oder Ausbeutungstitel? Springer, Volume 5, 2008.

¹³⁷ Ibid, pp. 364-365.

¹³⁸ H.-D. Lippert, supra.

results. This model does not cause problems in the practice, because the acquisition of ownership and exploitation rights is typically foreseen by the contract. The participating institutions are bound by these contractual relations and it is their obligation to procure the ownership over the research results from the persons, whom they engage into the project. This principle may be derogated from by the contract. In case of Horizon 2020 funded projects, the ownership of results passes to the party or parties, who conduct the work. The ownership of results is typically regulated by the Consortium Agreement. In Section 8, the template Consortium Agreement, based on the DESCAs Model 1.2., provides that *“Results are owned by the Party that generates them.”*¹³⁹

Against these considerations, no hard rule with respect to the data ownership in SoBigData may be inferred. Who owns the data and who owns what rights in respect of the data, collected by the SoBigData platform, needs to be assessed on the case by case basis depending on the dataset at issue, the substance of such dataset, the legal relations surrounding its creation and provisions of the national law, applicable to it.

4.3 OWNERSHIP OF PERSONAL INFORMATION

The personal data in the European Union is subject to a data protection framework that grants certain rights to its use to a number of actors and creates a rights and control regime under which the use of the data is regulated. There are opinions in the scientific literature that assume that in the case of personal data the data subject is given by the data protection regulations sufficient rights of disposal regarding his or her personal data which is why a “quasi-ownership” would be justified¹⁴⁰. However, the data subject does not have unrestricted power of control over the data and he cannot transfer his legal position¹⁴¹.

Although there are increasingly voices calling for ownership of data – be it personal or non-personal- it is also questionable whether it is desirable to establish such a right. On the one hand it is argued that legal insecurities are harmful for the development of data based markets¹⁴². On the other hand it is argued that

¹³⁹ DESCAs - Horizon 2020 Model Consortium Agreement (www.DESCA-2020.eu) Version 1.2, March 2016:

http://www.desca-2020.eu/fileadmin/content/Desca_2020_1.2/DESCA2020_v1.2_March_2016_with_elucidations.pdf.

¹⁴⁰ Kilian, Strukturwandel der Privatheit, in Wovon – für wen – wozu. Systemdenken wider die Diktatur der Daten – Wilhelm Steinmüller zum Gedächtnis, S. 210.

¹⁴¹ Zech, Daten als Wirtschaftsgut – Überlegungen zu einem „Recht des Datenerzeugers“, CR 2015, 141; Hoeren/Völkel in Big Data und Recht, S. 20.

¹⁴² Sahl, Brauchen wir ein Datenschutzrecht für Maschinendaten?, <https://berliner-datenschutzrunde.de/node/162>, 15.04.2015.

we live in an information society and exclusive rights on data should be an exception¹⁴³. The European Commission is planning to set up a group of experts to assess the need for guidance on specific issues of data ownership, in particular for data gathered through IoT technology¹⁴⁴.

¹⁴³ Härtig, Acht Thesen zum „Dateneigentum“, <http://www.cr-online.de/blog/2016/02/17/acht-thesen-zum-dateneigentum/>, 17.02.2016.

¹⁴⁴ European commission, Communication from the Commission to the European Parliament, the council, the European Economic and Social Committee and the Committee of the Regions, 2 July 2014, http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=6210.

5 CONCLUSIONS

This deliverable was produced in the result of preliminary analysis of IPR issues in the SoBigData project and reflects the results reached so far. The three main aspects, which have been considered here, concern the protectability of SoBigData datasets by IPR (Section 2), elaboration on the responsible IP principles (Section 3) and data ownership issues (Section 4).

The analysis of the SoBigData datasets revealed that some datasets have been produced by enriching, post-processing, analysing the already existing datasets, which are available in the public domain. The key legal issues, raised in this regard are: (a) whether the datasets produced for the SoBigData by some digital means of data processing may render such datasets works protected by copyright; (b) whether the datasets, taken from the public domain, in particular social networks, may be protected by copyrights, meaning that certain types of processing of such datasets in the SoBigData RI may constitute copyright relevant actions and require consent of the right holders.

Answering the first question, it must be noted that traditional law of copyright protects original intellectual creation of the author, expressed in a work. Considering against the technical background that most of the SoBigData datasets have been generated in the result of digital data procession (i.a. crawling, data mining, analysing by software tools, copying, extracting, post-processing, etc.) lacking any expression of original creativity, the requirement for the protection of SoBigData datasets may not be considered as fulfilled and copyright protection is most likely to be denied.

Apart from the protection of intellectual creation, the law of copyright also affords protection to the economic investment, deployed in producing some type of works by related rights. The list of objects protected by related rights is strictly limited and covers phonograms, fixations of films, fixations of broadcasts and performances, audiovisual fixations. The datasets produced for the SoBigData project do not constitute such fixations. The protection, provided by related rights, may not be considered as applicable to protect the investment of digital data procession, spent on producing the SoBigData datasets. Thus, protectability of SoBigData datasets by copyright or related rights may hardly be an option.

Another possibility might be the protection of SoBigData datasets by sui generis database rights. However, since the SoBigData RI is not supposed to integrate a central repository and the datasets are stored and managed on premises by each individual party, the assessment whether each party repository may qualify as a database and be covered by sui generis rights is for the individual parties to ascertain. The legal requirements and guidelines for sui generis database protection are provided in detail in Section 2.2 above. A characteristic feature of sui generis protection is that sui generis rights apply to the repository contents as a whole or in substantial part and protect the repository contents from copying or extraction as a whole or in substantial part. A party, having sui generis rights, may assert legal claims to the user or another RI-Partner against copying or extracting the contents of its repository in whole or in substantial part without its consent. However, it may not claim such rights to individual data items collected in the repository. Individual datasets may be protected in their own right (such as copyrights or knowhow) and be covered in parallel by sui generis rights as part of the repository contents as well.

A third type of protection, which may be considered for the SoBigData datasets, is the legal regime of know how. This legal regime may well be applicable to the data items, which by one or another reason,

must be kept and handled as confidential and which has not been revealed to the broader public before. The datasets, subject to the knowhow protection, must be stamped as “confidential” and may only be disclosed upon signing the NDA. The main problems in this type of protection may be additional workload, the lack of control over how such datasets are handled upon release and enforcement of NDA.

Continuing the exploration of IPR issues in SoBigData project, it must be observed that copyrights and related rights may well be applicable to the datasets, from or on top of which the SoBigData datasets have been generated. A number of SoBigData datasets make use of the data from the social networks like Twitter, Flickr and Facebook. The content, uploaded by the users to the network services, may reach the necessary level of creativity and be protected by one or another copyright (e.g. image rights, related rights, copyrights in writings, drawings, etc.). Any use of such content would require permission of the right holder. Also, the aspect that such content items are hosted and taken from the platform services may have a result that the use of such contents may be subject to the platform terms of use and/or be governed by the API license agreement, provided the content items have been collected via API. Protectability of the initial datasets by one or another copyright and the terms of use shall be taken into account while using and providing such data on the SoBigData services.

Compliance with the rights of third parties and terms of use of external datasets makes one of the essential IP principles for the SoBigData RI. The legal analysis revealed that a major number of datasets may be protected by copyrights and the use of such data is subject to the terms of use of the right holder or initial data provider. The overview of data types in SoBigData, applicable rights and terms of use is presented in Appendix B. It is advised that the information on IP aspects, surrounding each particular dataset, should be integrated into the dataset metadata and be completed by the RI-Partner, providing the data. The recommended list of parameters of IP relevant metadata (i.e. type of IP, source, license/terms of use, basic rights, etc.) is provided in Appendix A.

The analysis of data ownership issues revealed that the relations of ownership over the data are to be assessed in the complex of factors. The materials, with and on which the research producing the data has been conducted, and the rights thereto, the legal and contractual relations, under which the work, generating the data has been performed, the contractual regulations over the results of such work, may play a role. The generation of data by a researcher in the employee position of the institution would normally render the ownership of research results to the institution. Production of data in the context of third party projects may be affected by the contractual framework of the project. The ownership over research results, which reach the level of IP protection, needs to be assessed from the perspective of IP law, in particular copyright. According to the general principles of copyright, original copyright passes to the creator, unless otherwise is provided by the law or agreed by the contract. The ownership of personal information is to be considered from the perspective of data protection law.

APPENDIX A. IP RELEVANT METADATA

Metadata	Description	Example
IP, if applicable	Whether dataset is covered by any rights: copyright, related rights, database right, know how, proprietary, etc.	Copyright: image, publication, writing, articles, comments Related rights: broadcasts, phonograms, Database: ontologies Know how: datasets with confidentiality mark
Source/Origin	From where/whom the dataset originates	Twitter
Link to the source	www.	www.twitter.com
Right holder, if identified and different from the source	Who holds rights in the content	Source: Twitter Rightholder: Twitter users Source: Flickr Rightholder: Flickr users
License, if applicable	Terms under which dataset can be used	Creative Commons (CC), Twitter license, etc.
Link to the license	www.	www.license.com
Field/Scope of use	In what field the dataset may be used: Any use Non-commercial	Twitter Developer Agreement, IB: use of the content “on and through your Services to End Users” For the user it means use in course of using SoBigData Services only. Any subsequent use subject to the Twitter terms

	<p>only</p> <p>Research only</p> <p>Non-commercial research only</p> <p>Private use</p> <p>Use for developing and providing a service</p>	
Basic rights	<p>What the user can do with the dataset:</p> <p>Search and browse</p> <p>Temporary download of a single copy only</p> <p>Download</p> <p>Copying</p> <p>Distribution</p> <p>Modification</p> <p>Communication</p> <p>Making available to the public</p> <p>Other rights</p>	<p>Twitter Developer Agreement, IB:</p> <p>“- Copy a reasonable amount of and display the Content on and through your Services to End Users, as permitted by this Agreement;</p> <p>- Modify Content only to format it for display on your Services”</p> <p>https://dev.twitter.com/overview/terms/agreement-and-policy</p>
Restrictions on use	<p>Any restrictions on how where the dataset may be used</p>	<p>Twitter Developer Agreement:</p> <p>Geographic Data. Your license to use Content in this Agreement does not allow you to (and you will not allow others to) aggregate, cache, or store location data and other geographic information contained in the Content, except in conjunction with a Tweet to which it is attached. Your license only allows you to use such location data and geographic information to identify the location tagged by the Tweet. Any use of location data or geographic information on a standalone basis or beyond</p>

		<p>the license granted herein is a breach of this Agreement.</p> <p>https://dev.twitter.com/overview/terms/agreement-and-policy</p>
Prohibited actions	What the user cannot do with the dataset	<p>Twitter Developer Policy, I2:</p> <p>“Do not modify, translate or delete a portion of the Content</p> <p>Do not remove or alter any proprietary notices or marks on Content or the Twitter API.</p> <p>Do not (and do not allow others to) aggregate, cache, or store location data and other geographic information contained in the Content, except as part of a Tweet. Any use of location data or geographic information on a standalone basis is prohibited.”</p> <p>https://dev.twitter.com/overview/terms/agreement-and-policy</p>
Sublicense rights	Whether the user can pass the rights/content to any third parties	<p>Twitter Developer Agreement, IB:</p> <p>“non-transferable, non-sublicensable, revocable license”</p> <p>https://dev.twitter.com/overview/terms/agreement-and-policy</p>
Attribution requirements	Whether the user needs to acknowledge the source when using/distributing data/developing service	<p>Census Bureau API Terms of Service</p> <p>All services, which utilize or access the API, should display the following notice prominently within the application: "This product uses the Census Bureau Data API but is not endorsed or certified by the Census Bureau."</p> <p>https://www.census.gov/data/developers/about/terms-of-service.html</p>
Display requirements	Whether the user, when displaying the dataset in any media or form, must follow certain display requirements, e.g. attach copyright notice	<p>Twitter Developer Policy, I2f:</p> <p>“Follow the Display Requirements and Twitter Rules.”</p> <p>https://dev.twitter.com/overview/terms/agreement-and-policy</p>

Distribution requirements	Whether the user, when distributing the dataset, if allowed, must follow certain requirements	CC BY: give appropriate credit, provide a link to the license, and indicate if changes were made https://creativecommons.org/licenses/by/4.0/
Territory of use	In what territory dataset may be used	Worldwide Germany EU
License term	Period of time during which the dataset may be used	Perpetual dd.mm.yyyy. As long as you use the SoBigData services
Requirement of non-disclosure (confidentiality mark)	Whether the dataset bears confidentiality mark/may be used and shared subject to the obligation of non-disclosure	
Other		

APPENDIX B. SOBIGDATA DATASETS, APPLICABLE RIGHTS AND TERMS OF USE

	Dataset	Providing lab	Data type and rights	License/Terms of Use	Rights
1	.ee Web archive	ATI @ UT	<p>Web page snapshots</p> <p>Web content may fall under copyright and data protection law. Use of the content may be subject to the terms of right holders/webpage terms/terms of initial data provider.</p>	<p>.ee Web archive data license agreement</p> <p><i>The .ee Web archive datasets are distributed by University of Tartu for research purposes only. The dataset may be obtained from University of Tartu for free by signing a data license agreement with the same terms and conditions as the ClueWeb12 dataset:</i></p> <p>http://lemurproject.org/clueweb12/organization_agreement.clueweb12.worder.Sep15-15.pdf</p>	<p>Permitted Uses</p> <p>Same as for the ClueWeb12 dataset.</p> <p><i>The Information may only be used for research and development of natural-language processing, information-retrieval or document-understanding systems by the specific Group identified above.</i></p> <p><i>2. Summaries, analyses and interpretations of the linguistic properties of the Information may be derived and published, provided it is not possible to reconstruct the Information from these summaries.</i></p> <p><i>3. Small excerpts of the Information may be displayed to others or published in a scientific or technical context, solely for the purpose of describing the research and development carried out and related issues.</i></p> <p><i>“use only by persons working within Organization’s specific Group”</i></p> <p>Attribution requirements are the same as for Clue12 project. The attribution should be made to University of Tartu:</p> <p>http://lemurproject.org/clueweb12/orga</p>

					nization_agreement.clueweb12.worder.Sep15-15.pdf
2	20lines	KDD Lab ISTI CNR	Social network data/collaborative writing Social content may fall under copyright and data protection law. Use of the content may be subject to the terms of right holders/platform terms/terms of initial data provider.	Social Network Terms on use of the content.	To be identified by the data provider
3	Aalto-Foursquare	Aalto Data Mining Group	Tweets Twitter content may be protected by copyright and data protection law. Use of the content subject to the terms of right holders/Twitter Terms of Use.	Foursquare's Developer Policies https://developer.foursquare.com/overview/community Twitter Terms: https://twitter.com/tos?lang=en https://dev.twitter.com/overview/terms/agreement https://dev.twitter.com/overview/terms/policy#6.Update_Be_a_Good_Partner_to_Twitter	Twitter Developer Agreement, I B: "2. Copy a reasonable amount of and display the Content ¹⁴⁵ on and through your Services ¹⁴⁶ to End Users ¹⁴⁷ , as permitted by this Agreement; 3. Modify Content only to format it for display on your Services;" The user has a right to use Twitter content as part of/in course of using SoBlgData Services. Any subsequent use – subject to the terms of Twitter. Foursquare License: „Herunterladen und zur Nutzung der

¹⁴⁵ Tweets, Tweet IDs, Twitter end user profile information, and any other data and information made available to you through the Twitter API or by any other means authorized by Twitter, and any copies and derivative works thereof.

¹⁴⁶ Your websites, applications and other offerings that display Content or otherwise use the Licensed Material.

¹⁴⁷ Users of your Services, who are bound to enforceable service terms with you.

					<i>Foursquare-Materialien zum Zweck der (i) Entwicklung, Einbindung und Integration in seinen Service und (ii) Nutzung, Vervielfältigung, Verbreitung, Übermittlung und Darstellung der Foursquare-Materialien als Teil seines Service“</i>
4	Aalto-Twitter	Aalto Data Mining Group	Tweets Twitter content may be protected by copyright and data protection law. Use of the content may be subject to the terms of the right holders/Twitter Terms of Use.	Twitter Terms https://twitter.com/tos?lang=en https://dev.twitter.com/overview/terms/agreement https://dev.twitter.com/overview/terms/policy#6.Update_Be_a_Go od_Partner_to_Twitter	Twitter Developer Agreement, I B: “2. Copy a reasonable amount of and display the Content ¹⁴⁸ on and through your Services ¹⁴⁹ to End Users ¹⁵⁰ , as permitted by this Agreement; 3. Modify Content only to format it for display on your Services;” The user has a right to use Twitter content as part of/in course of using SoBlgData Services. Any subsequent use – subject to the terms of Twitter.
5	Archive Twitter Dataset	ETH Zurich	Tweets Streaming Twitter content may be protected by copyright and data protection law. Use of the content may be subject to the terms of the right holders/Twitter	Twitter Terms https://twitter.com/tos?lang=en https://dev.twitter.com/overview/terms/agreement https://dev.twitter.com/overview/terms/policy#6.Update_Be_a_Go od_Partner_to_Twitter	Twitter Developer Agreement, I B: “2. Copy a reasonable amount of and display the Content ¹⁵¹ on and through your Services ¹⁵² to End Users ¹⁵³ , as permitted by this Agreement; 3. Modify Content only to format it for display on your Services;” The user has a right to use Twitter

¹⁴⁸ Tweets, Tweet IDs, Twitter end user profile information, and any other data and information made available to you through the Twitter API or by any other means authorized by Twitter, and any copies and derivative works thereof.

¹⁴⁹ Your websites, applications and other offerings that display Content or otherwise use the Licensed Material.

¹⁵⁰ Users of your Services, who are bound to enforceable service terms with you.

¹⁵¹ Tweets, Tweet IDs, Twitter end user profile information, and any other data and information made available to you through the Twitter API or by any other means authorized by Twitter, and any copies and derivative works thereof.

¹⁵² Your websites, applications and other offerings that display Content or otherwise use the Licensed Material.

¹⁵³ Users of your Services, who are bound to enforceable service terms with you.

			Terms of Use.	od_Partner_to_Twitter	content as part of/in course of using SoBigData Services. Any subsequent use – subject to the terms of Twitter.
6	Articles and comments of four major Estonian news portals	ATI @ UT	<p>Articles/Comments</p> <p>Content may fall under copyright and data protection law, collection of data may be protected by database rights. Use of the data may be subject to the terms of right holders/initial data provider.</p>	<p>Same as ClueWeb12 data license agreement</p> <p><i>The Estonian news portals articles and comments are distributed by Register OÜ for research purposes only. The dataset may be obtained from Register OÜ for free by signing a data license agreement and paying a fee covering distribution costs, under the terms and conditions as the ClueWeb12 dataset:</i></p> <p>http://lemurproject.org/clueweb12/organization_agreement.clueweb12.worder.Sep15-15.pdf</p>	<p>Permitted Uses</p> <p>Same as for the ClueWeb12 dataset.</p> <p><i>The Information may only be used for research and development of natural-language processing, information-retrieval or document-understanding systems by the specific Group identified above.</i></p> <p>2. <i>Summaries, analyses and interpretations of the linguistic properties of the Information may be derived and published, provided it is not possible to reconstruct the Information from these summaries.</i></p> <p>3. <i>Small excerpts of the Information may be displayed to others or published in a scientific or technical context, solely for the purpose of describing the research and development carried out and related issues.</i></p> <p><i>“use only by persons working within Organization’s specific Group”</i></p> <p>Attribution requirements are the same as for Clue12 project. The attribution should be made to Register OÜ:</p> <p>http://lemurproject.org/clueweb12/organization_agreement.clueweb12.worder.Sep15-15.pdf</p>

7	BrightKite	KDD Lab ISTI CNR	Public dataset of checkin, collected from Brightkite website Content may fall under copyright and data protection law. Use of the data may be subject to the Brightkite terms of use.	Brightkite terms, no longer available	Brightkite terms, no longer available Permitted uses to be identified by the data provider.
8	ClueWeb 2009	HPC Lab ISTI CNR	Webpage notes Content may fall under copyright law. Use of the content may be subject to the terms of right holders/webpage terms.	ClueWeb09 dataset license http://www.lemurproject.org/clueweb09/index.php#Using <i>A ClueWeb09 dataset license is required before you can begin using a hosted copy of the dataset. There is no cost for a dataset license; it is free.</i> <i>Each individual who will use or have access to the dataset must sign an Individual Agreement.</i>	ClueWeb09 dataset license <i>“use only by persons working within Organization’s specific Group”</i> Permitted Uses <i>The Information may only be used for research and development of natural-languageprocessing, information-retrieval or document-understanding systems by the specific Group identified above.</i> <i>2. Summaries, analyses and interpretations of the linguistic properties of the Information may be derived and published, provided it is not possible to reconstruct the Information from these summaries.</i> <i>3. Small excerpts of the Information may be displayed to others or published in a scientific or technical context, solely for the purpose of describing the research and development</i>

					<p>carried out and related issues.</p> <p>http://lemurproject.org/clueweb09/organization_agreement.clueweb09.worder.Sep15-15.pdf</p>
9	ClueWeb 2012	HPC Lab ISTI CNR	<p>Webpage notes</p> <p>Content may fall under copyright law. Use of the content may be subject to the terms of right holders/webpage terms.</p>	<p>ClueWeb12 data license agreement</p> <p><i>The ClueWeb12 datasets are distributed by Carnegie Mellon University for research purposes only. A dataset may be obtained from Carnegie Mellon by signing a data license agreement with Carnegie Mellon University, and paying a fee that covers the cost of distributing the dataset.</i></p> <p>http://lemurproject.org/clueweb12/index.php#Obtaining</p>	<p>Permitted Uses</p> <p><i>The Information may only be used for research and development of natural-languageprocessing, information-retrieval or document-understanding systems by the specific Group identified above.</i></p> <p>2. <i>Summaries, analyses and interpretations of the linguistic properties of the Information may be derived and published, provided it is not possible to reconstruct the Information from these summaries.</i></p> <p>3. <i>Small excerpts of the Information may be displayed to others or published in a scientific or technical context, solely for the purpose of describing the research and development carried out and related issues.</i></p> <p><i>“use only by persons working within Organization’s specific Group”</i></p> <p>Attribution requirement</p> <p>http://lemurproject.org/clueweb12/organization_agreement.clueweb12.worder.Sep15-15.pdf</p>
1	Collection of texts for data	A^3 lab	Miscellaneous		

0	compression	UNIPI	Content may fall under copyright and data protection law, database and proprietary rights. Use of the content may be subject to the terms of right holders/ terms of initial data provider.		
			Census: U.S. demographic statistics in tabular format (type: database).	<p>Census Bureau Application Programming Interface Terms of Service Agreement</p> <p>https://www.census.gov/data/developers/about/terms-of-service.html</p> <p><i>All of the content, documentation, code and related materials made available to you through the API are subject to these terms. Access to or use of the API or its content constitutes acceptance to this Agreement.</i></p>	<p><i>You may use the Census Bureau API to develop a service or service to search, display, analyze, retrieve, view and otherwise "get" information from Census Bureau data.</i></p> <p>Attribution requirement in services using API</p> <p><i>"use only by persons working within Organization's specific Group"</i></p>
			Dna: collection of families of genomes (type: highly repetitive biological data).	<p>Pizza&Chili Corpus Compressed Indexes and their Testbeds</p> <p>API Terms of Use</p> <p>http://pizzachili.dcc.uh.edu/cl/api.html</p> <p>the API interface</p>	To be identified by the data provider

				offered by all the compressed indexes available for <u>downloading</u> . Please read carefully the COPYRIGHT information that comes with each of them.	
			Mingw: archive containing the whole mingw software distribution (type: mix of source codes and binaries).	Matt Mahoney terms (tbc)	To be identified by the data provider
			Wikipedia: dump of English Wikipedia (type: natural language).	<u>Creative Commons Attribution-ShareAlike 3.0 License (CC-BY-SA)</u> and the <u>GNU Free Documentation License (GFDL)</u> .	<u>Creative Commons Attribution-ShareAlike 3.0 License</u> Rights: to Share—to copy, distribute and transmit the work, and to Remix—to adapt the work Requirements: Attribution—You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work.) Share Alike—If you alter, transform, or build upon this work, you may distribute the resulting work only under the same, similar or a compatible license.
1 1	Coop	KDD Lab ISTI CNR	Purchases of Unicoop Tirreno customers, description and information of the shops. Use of the content may be subject to the terms	Terms of the data owner/provider	To be identified by the data provider

			of initial data provider.		
1 2	CoPhIR	HPC Lab ISTI CNR	<p>Content-based Photo Image Retrieval)</p> <p>Based on images from Flickr. The data collected consist of 106 million processed images.</p> <p>Content may fall under copyright and data protection law, image rights. Use of the content may be subject to the terms of right holders/social network (Flickr)/data provider.</p>	<p>Flickr API Terms of Use</p> <p>https://www.flickr.com/services/api/tos/</p> <p>Comply with any requirements or restrictions imposed on usage of the photos by their respective owners. Remember, Flickr doesn't own the images - Flickr users do. Although the Flickr APIs can be used to provide you with access to Flickr user photos, neither Flickr's provision of the Flickr APIs to you nor your use of the Flickr APIs override the photo owners' requirements and restrictions, which may include "all rights reserved" notices (attached to each photo by default when uploaded to Flickr), Creative Commons licenses or other terms and conditions that may be agreed upon between you and the owners. In ALL cases, you are solely responsible for making use of Flickr photos in compliance with the photo owners' requirements or restrictions. If you use Flickr photos for a commercial purpose,</p>	Rights attached to each single image

				the photos must be marked with a Creative Commons license that allows for such use, unless otherwise agreed upon between you and the owner. You can read more about this here: www.creativecommons.org or www.flickr.com/creativecommons .	
1 3	Corpora of web crawls in English, French, Italian and German	USFD	Webcrawls Content may fall under copyright and data protection law. Use of the content may be subject to the terms of right holders/ webpage terms.	Terms of use of the webcontent.	To be identified by the data provider
1 4	DBLP computer science bibliography	KDD Lab ISTI CNR	Dump of the DBPL archive (computer science bibliography) Content may fall under copyright law. Use of the content subject to the terms of data owner/initial data provider.	Terms of use of the DBPL archive.	To be identified by the data provider
1 5	DE webarchive	LUH	Content from the .de top level domain as crawled by the Internet Archive. Content may fall under copyright and data protection law. Use of the content subject to the terms	Memorandum of Understanding between Internet Archive and Leibniz Universität Hannover	No replicas may be provided. SoBigData rights: 1. Creation of a topical subset of the .de domain from the .de Web Data collection provided by IA. 2. The subset will be uploaded into the Gate Cloud. Gate Cloud (https://gatecloud.net/) is a text

			of right holders/initial data provider.		<p>processing infrastructure provided by the SoBigData partner University of Sheffield. Only the Gate Cloud administrator and developers involved in the SoBigData project have access to the data for maintenance and development purposes.</p> <p>3. Gate Cloud will process the data subset to extract entities, topics, sentiments and keywords.</p> <p>4. The processing results will be transferred to other processing services provided by SoBigData consortium partners (Appendix 1).</p> <p>5. The Web Data will be removed from the Gate Cloud platform as soon as the processing was successful.</p>
1 6	e-mid	SNS	<p>Interbank market transactions in the Italian market.</p> <p>Use of the data subject to the terms of initial data provider.</p>	<p>Terms set by the data owner:</p> <p>Data owned by SNS as part of the QuantLab laboratory jointly with LIST Spa (tbd).</p>	To be identified by the data provider
1 7	EMID Data	IMT	<p>Interbank market transactions in the Italian market. Use of the data subject to the terms of initial data provider.</p>	<p>Data is bought from EMID company and can be used for publication only by buyers</p>	Can be used for publication only by buyers
1 8	English documents annotated for IE tasks	USFD	<p>Documents annotated for IE tasks.</p> <p>Content may fall under copyright law. Use of the data subject to the terms of initial data</p>	Limited by terms and conditions	To be identified by the data provider

			provider.		
19	English Newswire	USFD	Text and Social Media Mining. Content may fall under copyright and data protection law. Use of the data may be subject to the terms of right holders/platform terms.	Limited by terms and conditions	To be identified by the data provider
20	Estonian public sector Web services	ATI @ UT	Descriptions of all Estonian public sector Web services (and their “consumers”), which can be complemented with E-government service usage stats (monthly). See https://opendata.riik.ee/dataset/riha-linkandmed . Content may fall under copyright law and be subject to the terms of initial data provider.	Creative Commons BY 3.0 https://opendata.riik.ee/dataset/riha-linkandmed	You are free to: Share — copy and redistribute the material in any medium or format Adapt — remix, transform, and build upon the material for any purpose, even commercially. The licensor cannot revoke these freedoms as long as you follow the license terms. Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use. No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.
21	Facebook - EuroSys '09	UI IIT CNR	Social and interaction graphs. Content may fall under copyright and data protection law. Use of the content may be subject to the terms of Facebook/initial data	Terms of use of the data owners: ravenben at cs dot ucsb dot edu. Facebook Terms, if applicable: https://developers.facebook.com/policy/#data .	Permission to access the data must be obtained directly from the owner (ravenben at cs dot ucsb dot edu). Data cannot be distributed beyond the immediate research group of the person who obtained the permission.

			provider.		
2 2	Facebook - WOSN '09	UI IIT CNR	Social and interaction graphs, wall posts Content may fall under copyright and data protection law. Use of the content may be subject to the terms of Facebook/initial data provider.	Facebook Terms, if applicable: https://developers.facebook.com/policy/#data .	Facebook Platform Policy, Section 3 Protect data: 9. Don't sell, license, or purchase any data obtained from us or our services. 10. Don't transfer any data that you receive from us (including anonymous, aggregate, or derived data) to any ad network, data broker or other advertising or monetization-related service. 11. Don't put Facebook data in a search engine or directory, or include web search functionality on Facebook. 12. If you are acquired by or merge with a third party, you can continue to use our data only within your app.
2 3	Flickr and Wikipedia Turism Trajectories	HPC Lab ISTI CNR	Public user-generated content Content may fall under copyright and data protection law. Use of the content may be subject to the terms of platforms.	Creator/Flickr/Wikipedia (CC/GDFL) terms	To be identified by the data provider
2 4	Formal network of Estonian companies and board members	ATI @ UT	Database accessible via API Content may be protected by copyright, company and data protection law, sui generis database rights. Use of the content subject to the terms of initial data	Creator/provider terms Owned by Register OU. https://developers.iree/graph-api Users can query the API to the extent permitted by the trial account.	A trial license can be requested allowing a limited amount of API calls for data retrieval. Data obtained via the API can be stored and processed by the recipient but not further distributed. https://developers.iree/graph-api

			provider.		
2 5	German Academic Web	LUH	Archive of German academic institutions. Content may fall under copyright and data protection law. Use of the content subject to the terms of right holders/initial data provider.	Terms of use of German Academic Web (tbd)	To be identified by the data provider
2 6	Google+	KDD Lab ISTI CNR	Social Network with attributes on nodes. Content may fall under copyright and data protection law. Use of the content may be subject to the terms of right holders/ terms of Google.	Google Terms https://www.google.com/intl/en/policies/terms/regional.html	You may not use content from our Services unless you obtain permission from its owner or are otherwise permitted by law.
2 7	IMDb	KDD Lab ISTI CNR	Movie, TV and celebrity content. Content may be protected by copyrights and related rights, data protection law. Use of the content may be subject to the terms of right holders/initial data provider.	IMDb Conditions of Use http://www.imdb.com/conditions	Subject to your compliance with these Conditions of Use and your payment of any applicable fees, IMDb or its content providers grants you a limited, non-exclusive, non-transferable, non-sublicenseable license to access and make personal and non-commercial use of the IMDb Services, including digital content available through the IMDb Services, and not to download (other than page caching) or modify this site, or any portion of it, except with express written consent of IMDb. Additional license terms may be found in the Terms. The IMDb Services or any portion of such services may not be reproduced, duplicated, copied, sold, resold, visited, or otherwise exploited for any

					commercial purpose without express written consent of IMDb. This license does not include any resale or commercial use of any IMDb Service or its contents or any derivative use of this site or its contents. All licenses are non-exclusive and all rights not expressly granted to you in these Conditions of Use or any applicable Terms are reserved and retained by IMDb or its licensors, suppliers, publishers, rightsholders, or other content providers. You will use all IMDb Services in compliance with all applicable laws.
28	Infocamere	KDD Lab ISTI CNR	A sample of Italian Companies' data provided by Infocamere. Content may fall copyright, company and data protection law. Use of the content subject to Infocamere terms of use.	Infocamere terms of service/use (tbd) Infocamere privacy terms http://www.infocamere.it/note-legali	To be identified by the data provider
29	ISTAT Census Data Tuscany	KDD Lab ISTI CNR	Geometry of census sector and limited demographic information. Use of the content may be subject to the terms of initial data provider.	Istat legal notice http://www.istat.it/en/legal-notice Unless otherwise stated, content on this website is licensed under a Creative Commons License – Attribution – 3.0. http://creativecommons.org/licenses/by/3.0/	Data and analysis from the Italian National Institute of Statistics can be copied, distributed, transmitted and freely adapted, even for commercial purposes, provided that the source is acknowledged. No permission is necessary to hyperlink to pages on this website. Images, logos (including Istat logo), trademarks and other content owned by third parties belong to their respective owners and cannot be reproduced without their consent.
30	Italian Twitter Dataset	A^3 lab UNIPI	Tweets Streaming Twitter content may	Twitter Terms https://twitter.com/to	Twitter Developer Agreement, I B: "2. <i>Copy a reasonable amount of</i>

			be protected by copyright and data protection law. Use of the content subject to Twitter Terms of Use.	s?lang=en https://dev.twitter.com/overview/terms/agreement https://dev.twitter.com/overview/terms/policy#6.Update_Be_a_Good_Partner_to_Twitter	<i>and display the Content¹⁵⁴ on and through your Services¹⁵⁵ to End Users¹⁵⁶, as permitted by this Agreement;</i> <i>3. Modify Content only to format it for display on your Services;”</i> The user has a right to use Twitter content as part of/in course of using SoBigData Services. Any subsequent use – subject to the terms of Twitter.
3 1	LastFM	KDD Lab ISTI CNR	The dataset collected from Last.fm, a famous UK music website, contains the social network of the users and the weekly listenership for a period of 2 years. Content may fall under copyrights. Use of the content subject to LastFM terms of use.	Terms of use of last.fm website.	To be identified by the data provider
3 2	Linguistically annotated corpora of IRC chats, reviews, Q&A, email, blogs and comments, newsgroups	USFD	Web text - reviews, Q&A, email, blogs and comments, newsgroups annotated in a lot of ways (token, pos, syntax). Content may fall under copyright and data protection law. Use of the content subject to	LDC User Agreement for Non-Members https://catalog.ldc.upenn.edu/LDC2012T13	Non commercial linguistic education, research and technology development Publication, redistribution, communication to the public, etc – subject to LDC permission

¹⁵⁴ Tweets, Tweet IDs, Twitter end user profile information, and any other data and information made available to you through the Twitter API or by any other means authorized by Twitter, and any copies and derivative works thereof.

¹⁵⁵ Your websites, applications and other offerings that display Content or otherwise use the Licensed Material.

¹⁵⁶ Users of your Services, who are bound to enforceable service terms with you.

			the terms of initial data provider.		
3 3	Mobile Miner App Data	Dept of Digital Humanities Kings College London	Data collected by the MobileMiner Android smartphone app. [http://kingsbsd.github.io/MobileMiner/]. This includes time series of mobile cell tower connections, records of when apps opened network sockets by polling the /proc/net/ directory and using the network traffic API, a log of when apps send notifications, and also mobile and wireless network connections. This data is saved to an SQLite database that can be exported and analysed elsewhere.	SQLite database terms of use: http://kingsbsd.github.io/MobileMiner/	SQLite database can be exported and analysed elsewhere. Press "My Data" then "Export" to copy the current SQLite database to where your computer will be able to see it when you mount your device. The data will be in the "Download" directory of the device's flash memory, or SD card if it has one. The easiest way to explore the data is with an SQLite browser extension, there's SQLite Manager for Firefox, and codev for Chrome. If you want to play with your data with some code, SQLite is supported by Python's standard library .
3 4	OctoCalabria 2012 10 e 07	KDD Lab ISTI CNR	GPS dataset of private vehicles in Calabria region (Italy). Use of the content subject to the terms of initial data provider.	Restricted Access. Terms of use of Octo Telematics.	To be identified by the data provider
3 5	OctoMestre 2010 08	KDD Lab ISTI CNR	GPS dataset of private vehicles in the area of Mestre. Use of the content subject to the terms of initial data	Restricted Access. Terms of use of Octo Telematics.	To be identified by the data provider

			provider.		
36	OctoMilano 2007 04	KDD Lab ISTI CNR	GPS dataset of private vehicles in the area of Milano. Use of the content subject to the terms of initial data provider.	Restricted Access. Terms of use of Octo Telematics.	To be identified by the data provider
37	OctoPisa	KDD Lab ISTI CNR	GPS dataset of private vehicles in the area of Pisa. Use of the content subject to the terms of initial data provider.	Restricted Access. Terms of use of Octo Telematics.	To be identified by the data provider
38	Octoscana 2011 05	KDD Lab ISTI CNR	GPS dataset of private vehicles crossing Tuscany. Use of the content subject to the terms of initial data provider.	Restricted Access. Terms of use of Octo Telematics.	To be identified by the data provider
39	Orange D4D	KDD Lab ISTI CNR	Sample of CDRs collected in Ivory Coast and released by Orange for the D4D Challenge in 2014 Use of the content subject to the terms of initial data provider/Orange.	Terms of use of Orange.	To be identified by the data provider
40	Pisa Airport statistics	KDD Lab ISTI CNR	3 months of aggregated data over arrivals and departures to/from Pisa Airport: Jul, Nov	Pisa Airport terms/open access.	To be identified by the data provider

			and Dic. 2010 Use of the content subject to the terms of initial data provider/Pisa Airport.		
4 1	Pisa Hotel statistics	KDD Lab ISTI CNR	Bookings and reservations at Pisa hotels and accomodation structures. Use of the content subject to the terms of initial data provider.	Terms of data provider - Provincia di Pisa.	To be identified by the data provider
4 2	Pizza&Chili	A^3 lab UNIFI	The Pizza&Chili site (http://pizzachili.di.unipi.it/) offers a rich set of texts which cover a representative set of application areas, where the problem of full-text indexing might be relevant, and are freely available over the web. Use of the content subject to the terms of initial data provider.	Pizza&Chili API Terms (tbd) http://pizzachili.dcc.unipi.it/api.html	the compressed indexes available for <u>downloading</u> . Please read carefully the COPYRIGHT information that comes with each of them.
4 3	Product Reviews Rating Datasets	NeMIS Lab ISTI CNR	Hotel reviews from TripAdvisor. Content may fall under copyright and data protection law. Use of the content subject to the terms of TripAdvisor.	Hotel reviews: TripAdvisor terms TripAdvisor Website terms: https://www.tripadvisor.co.uk/pages/terms.html TripAdvisor API terms:	If reviews are copied from the TripAdvisor webpage, the following terms apply: Copying, transmission, reproduction, replication, posting or redistribution of the Website Content or any portion thereof is strictly prohibited without the prior written permission of TripAdvisor. To request permission, you may contact

				<p>https://developer-tripadvisor.com/content-api/terms-and-conditions/</p>	<p>TripAdvisor as follows:</p> <p>Director, Partnerships and Business Development TripAdvisor LLC 400 1st Avenue Needham, MA 02494, USA</p> <p>If reviews are gathered via TripAdvisor API, the following terms apply:</p> <p>to (i) use, perform and display (publicly or otherwise) the Licensed Content through Your Site, and (ii) allow End Users to access and use the Licensed Content through Your Site</p> <p><u>TripAdvisor API Display requirements:</u> https://developer-tripadvisor.com/content-api/display-requirements/</p> <p>Restrictions:</p> <p>to (i) Record, cache or else store any portion of the Licensed Content, (ii) alter the Licensed Content, or use it to create or update your own database of business listing information, (iii) create or disclose metrics about, or perform any statistical analysis of, the API, Licensed Content, (iv) use the API on behalf of any third party other than a group or affiliate company, (v) copy, rent, lease, sell, transfer, assign, sublicense, dissemble, reverse engineer or decompile (except to the limited extent expressly authorized under applicable statutory law), modify or alter any part of the API, (vi) display Licensed Content or TripAdvisor Marks in a manner that could reasonably imply an endorsement, relationship or affiliation with or sponsorship between you or a third party and</p>
--	--	--	--	--	---

					<p>TripAdvisor, other than your permitted use of the API under the terms of the Agreement, (vii) display the Licensed Content on any site that disparages TripAdvisor or its products or services, or infringes any TripAdvisor intellectual property or other rights, (viii) use the API in a manner that impacts the stability of TripAdvisor's servers or impacts the behavior of other applications using the API, (ix) use the API or Licensed Content in any manner or for any purpose that may violate any law or regulation, or any right of any person including, but not limited to, intellectual property rights, rights of privacy and/or rights of personality, or which otherwise may be harmful (in TripAdvisor 's sole discretion) to TripAdvisor, its providers, its suppliers, end users of this website, or your end users, (x) use the API, Licensed Content or TripAdvisor in a manner that could reasonably be interpreted to suggest that TripAdvisor is the author or entity that is responsible, in whole or in part, for the creation or development of any Licensed Content or that such Licensed Content represents the views of TripAdvisor. YOU MAY NOT DIRECTLY OR INDIRECTLY ATTEMPT TO RE-SELL, REDISTRIBUTE, SYNDICATE OTHERWISE COMMERCIALIZE THE TRIPADVISOR CONTENT IN ANY WAY WITHOUT TRIPADVISOR'S PERMISSION.</p> <p>https://developer-tripadvisor.com/content-api/terms-and-conditions/</p>
4 3	Product Reviews Rating	NeMIS Lab ISTI	Product reviews from Amazon.co.	Product reviews: Amazon terms	For reviews copied from Amazon website:

	Datasets	CNR	Content may fall under copyright and data protection law. Use of the content may be subject to the terms of Amazon.	<p>Amazon terms: http://www.amazon.com/gp/help/customer/display.html/ref=footer_cou?ie=UTF8&nodeId=508088</p> <p>Amazon EU Product Advertising API Licence Agreement (use of API for advertising purposes only)</p> <p>https://partnet.amazon.de/gp/advertising/api/detail/agreement.html</p> <p>Please check</p>	a limited, non-exclusive, non-transferable, non-sublicensable license to access and make personal and non-commercial use of the Amazon Services. This license does not include any resale or commercial use of any Amazon Service, or its contents; any collection and use of any product listings, descriptions, or prices; any derivative use of any Amazon Service or its contents; any downloading, copying, or other use of account information for the benefit of any third party; or any use of data mining, robots, or similar data gathering and extraction tools.
4 4	Query Log MSN RFP 2006	HPC Lab ISTI CNR	<p>MSN Search query log.</p> <p>Content may fall under copyright law. Use of the content subject to the terms of Microsoft.</p>	Microsoft terms.	To be identified by the data provider
4 5	Query-Based Twitter Dataset	ETH Zurich	Content may fall under copyright law. Twitter content may be protected by copyright and data protection law. Use	<p>Twitter terms:</p> <p>https://twitter.com/tos?lang=en</p> <p>https://dev.twitter.com/overview/terms/agg</p>	Twitter Developer Agreement, I B: "2. Copy a reasonable amount of and display the Content ¹⁵⁷ on and through your Services ¹⁵⁸ to End Users ¹⁵⁹ , as permitted by this Agreement;

¹⁵⁷ Tweets, Tweet IDs, Twitter end user profile information, and any other data and information made available to you through the Twitter API or by any other means authorized by Twitter, and any copies and derivative works thereof.

¹⁵⁸ Your websites, applications and other offerings that display Content or otherwise use the Licensed Material.

¹⁵⁹ Users of your Services, who are bound to enforceable service terms with you.

			of the content subject to Twitter Terms of Use.	<p>https://dev.twitter.com/overview/terms/policy#6.Update_Be_a_Go</p> <p>od Partner to Twitter</p>	<p>3. <i>Modify Content only to format it for display on your Services;</i>"</p> <p>The user has a right to use Twitter content as part of/in course of using SoBigData Services. Any subsequent use – subject to the terms of Twitter.</p> <p>Foursquare License:</p> <p><i>„Herunterladen und zur Nutzung der Foursquare-Materialien zum Zweck der (i) Entwicklung, Einbindung und Integration in seinen Service und (ii) Nutzung, Vervielfältigung, Verbreitung, Übermittlung und Darstellung der Foursquare-Materialien als Teil seines Service“</i></p>
46	Russell 3000	SNS	Data on the price and volume of the 3000 stocks belonging to the Russell 3000 Index, roughly corresponding to the 3000 more capitalized stocks. Traded volume and high, low, opening, and closing value of the price for each minute. Use of the content subject to the terms of Russel.	Data purchased by SNS.	Use possible either in preprocessed format or in original. Publications possible only if one of the authors is from SNS.
47	SentiWordNet	NeMIS Lab ISTI CNR	SentiWordNet is a lexical resource for opinion mining. SentiWordNet assigns to each synset of WordNet three sentiment scores: positivity,	<p>Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0) license</p> <p>http://creativecommons.org/licenses/by-sa/3.0/</p>	<p>Share — copy and redistribute the material in any medium or format</p> <p>Adapt — remix, transform, and build upon the material for any purpose, even commercially.</p> <p>Requirements:</p>

			<p>negativity, objectivity.</p> <p>Content may fall under copyright and data protection law. Use of the content subject to the terms of initial data provider.</p>		<p>Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.</p> <p>ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.</p>
48	SMAPH Annotated query dataset	A ³ lab UNIFI	<p>This is a dataset of annotated search-engine queries. Mentions of entities in search-engine queries are tagged with the entity they refer to. Content may fall under copyright law. Use of the content subject to the terms of initial data provider.</p>	Creator/Provider terms of use (tbd)	To be identified by the data provider
49	Social Banks Tracker Twitter Dataset	A ³ lab UNIFI	<p>Content may fall under copyright and data protection law.</p> <p>Twitter content may be protected by</p>	<p>Twitter Terms:</p> <p>https://twitter.com/tos?lang=en</p> <p>https://dev.twitter.com/overview/terms/agg</p>	<p>Twitter Developer Agreement, I B:</p> <p>“2. Copy a reasonable amount of and display the Content¹⁶⁰ on and through your Services¹⁶¹ to End Users¹⁶², as permitted by this Agreement;</p>

¹⁶⁰ Tweets, Tweet IDs, Twitter end user profile information, and any other data and information made available to you through the Twitter API or by any other means authorized by Twitter, and any copies and derivative works thereof.

¹⁶¹ Your websites, applications and other offerings that display Content or otherwise use the Licensed Material.

			copyright and data protection law. Use of the content subject to Twitter Terms of Use.	ement https://dev.twitter.com/overview/terms/policy#6.Update_Be_a_Good_Partner_to_Twitter	3. <i>Modify Content only to format it for display on your Services;</i> The user has a right to use Twitter content as part of/in course of using SoBigData Services. Any subsequent use – subject to the terms of Twitter.
50	Strava	KDD Lab ISTI CNR	Data of cyclists' performances (nr. of training, nr. of travelled km, heart rate, GPS tracks). Nr. of users = about 30.000 Content may fall under copyright and data protection law. Use of the content subject to the terms of initial data provider.	Terms of use of Strava.	To be identified by the data provider
51	TAGME Datasets	A ³ lab UNIPI	Short text fragments drawn from Wikipedia snapshot. Content may fall under copyright law. Use of the content may be subject to the Wikipedia terms.	Wikipedia licenses (tbd) <u>Creative Commons Attribution-ShareAlike 3.0 License (CC-BY-SA)</u> and the <u>GNU Free Documentation License (GFDL)</u> .	You are free: <ul style="list-style-type: none"> • to Share—to copy, distribute and transmit the work, and • to Remix—to <u>adapt</u> the work Under the following conditions: <ul style="list-style-type: none"> • Attribution—You must <u>attribute</u> the work in the manner specified by the author or <u>licensor</u> (but not in any way that suggests that they endorse you or your use of the work.) • Share Alike—If you alter, transform, or build upon this work, you may distribute the resulting work only under the

¹⁶² Users of your Services, who are bound to enforceable service terms with you.

					same, similar or a compatible license.
5 2	TagMyDay (dayTag)	KDD Lab ISTI CNR	GPS trajectories collected by volunteers, with annotations about means of transport and activities. Content may fall under data protection law and be subject to the terms of data providers.	Creator/Provider terms (tbd)	To be identified by the data provider
5 3	Thomson Reuters	SNS	Reuters proprietary.	Data purchased by SNS. Use possible either in preprocessed format or in original. Publications possible only if one of the authors is from SNS.	Data purchased by SNS. Use possible either in preprocessed format or in original. Publications possible only if one of the authors is from SNS.
5 4	Trenitalia	KDD Lab ISTI CNR	The dataset contains the arrival times of the Trenitalia's regional trains at the Italian railway stations. Content may fall under copyright law. Use of the data may be subject to Trenitalia terms.	Trenitalia terms/Open access (tbd)	To be identified by the data provider
5 5	TripAdvisor Annotated Dataset	NeMIS Lab ISTI CNR	Content may fall under copyright and data protection law. Use of the content subject to the terms of TripAdvisor.	TripAdvisor reviews: TripAdvisor terms: TripAdvisor Website terms: https://www.tripadvisor.co.uk/pages/terms .	TripAdvisor reviews: TripAdvisor terms: TripAdvisor Website terms: https://www.tripadvisor.co.uk/pages/terms.html

				<p><u>html</u></p> <p>If collected via TripAdvisor API:</p> <p>Tripadvisor API terms: https://developer-tripadvisor.com/content-api/terms-and-conditions/</p> <p>Annotations:</p> <p>Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0) license</p>	<p>If collected via TripAdvisor API:</p> <p>Tripadvisor API terms: https://developer-tripadvisor.com/content-api/terms-and-conditions/</p> <p>Annotations:</p> <p>Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0) license</p>
5 6	Twitter Gardenhose @ USFD	USFD	Twitter content may be protected by copyright and data protection law and reside within proprietary rights of Twitter. Use of the content subject to Twitter Terms of Use.	<p>Twitter Terms:</p> <p>https://twitter.com/tos?lang=en</p> <p>https://dev.twitter.com/overview/terms/agreement</p> <p>https://dev.twitter.com/overview/terms/policy#6.Update_Be_a_Good_Partner_to_Twitter</p>	<p>Twitter Developer Agreement, I B:</p> <p>“2. Copy a reasonable amount of and display the Content¹⁶³ on and through your Services¹⁶⁴ to End Users¹⁶⁵, as permitted by this Agreement;</p> <p>3. Modify Content only to format it for display on your Services;”</p> <p>The user has a right to use Twitter content as part of/in course of using SoBlgData Services. Any subsequent use – subject to the terms of Twitter.</p>

¹⁶³ Tweets, Tweet IDs, Twitter end user profile information, and any other data and information made available to you through the Twitter API or by any other means authorized by Twitter, and any copies and derivative works thereof.

¹⁶⁴ Your websites, applications and other offerings that display Content or otherwise use the Licensed Material.

¹⁶⁵ Users of your Services, who are bound to enforceable service terms with you.

<p>5 7</p>	<p>Twitter Stream - Gardenhose Daily Access</p>	<p>HPC Lab ISTI CNR</p>	<p>Twitter content may be protected by copyright and data protection law and reside within proprietary rights of Twitter. Use of the content subject to Twitter Terms of Use.</p>	<p>Twitter Terms: https://twitter.com/tos?lang=en https://dev.twitter.com/overview/terms/agreement https://dev.twitter.com/overview/terms/policy#6.Update_Be_a_Good_Partner_to_Twitter</p>	<p>Twitter Developer Agreement, I B: “2. Copy a reasonable amount of and display the Content¹⁶⁶ on and through your Services¹⁶⁷ to End Users¹⁶⁸, as permitted by this Agreement; 3. Modify Content only to format it for display on your Services;” The user has a right to use Twitter content as part of/in course of using SoBlgData Services. Any subsequent use – subject to the terms of Twitter.</p>
<p>5 8</p>	<p>US Banks balance sheets</p>	<p>SNS</p>	<p>Public data preprocessed by SNS Quantitative finance group. Use of the content subject to the terms of initial data provider.</p>	<p>Terms of Use (tbd) website of the Federal Reserve Bank of New York https://www.newyorkfed.org/termsfuse.html !</p>	<p>Except as limited by these Terms of Use, you may reproduce any information contained on the Bank's website if the reproduced information is not used for private gain. This does not apply to other websites that may be linked to or from the Bank's website. They are governed by the individual policies of the relevant providers. Some of the images and multimedia content on the website have been provided by third parties. You may not download or print any such images for any use other than your personal use and you may not republish, retransmit or reproduce any such images. The Bank's website incorporates some materials licensed by third parties, which may be subject to separate copyright and license conditions. All third party materials are provided “as is” and all express or implied warranties and all liabilities, damages, indemnities and obligations of any kind with respect</p>

¹⁶⁶ Tweets, Tweet IDs, Twitter end user profile information, and any other data and information made available to you through the Twitter API or by any other means authorized by Twitter, and any copies and derivative works thereof.

¹⁶⁷ Your websites, applications and other offerings that display Content or otherwise use the Licensed Material.

¹⁶⁸ Users of your Services, who are bound to enforceable service terms with you.

					to such materials are disclaimed. Any terms that differ from the applicable license for such materials is offered by the Bank alone and shall not bind any provider or publisher of such materials. To the extent any license for such materials grants a licensee rights to use, copy or modify such materials that are broader than the rights granted herein, then such rights shall take precedence over these Terms of Use.
59	Web 1T 5-gram, English + 10 EU languages	USFD	Web1T n-gram corpus: English, Swedish, Romanian, Portuguese, Polish, Dutch, Italian, French, German, Czech. Content may fall under copyright law. Use of the content subject to the terms of initial data provider.	Restricted by Terms and Conditions.	To be identified by the data provider
60	Wikipedia Graph	HPC Lab ISTI CNR	Contains English/Italian Wikipedia graph, for each article is possible to retrieve the nodes linked and the nodes that link to link to it. REST API have been developed to provide convenient access. Wikipedia content may fall under copyright law and be subject to the	Wikipedia licenses (tbd) <u>Creative Commons Attribution-ShareAlike 3.0 License (CC-BY-SA)</u> and the <u>GNU Free Documentation License (GFDL)</u> .	You are free: <ul style="list-style-type: none"> • to Share—to copy, distribute and transmit the work, and • to Remix—to <u>adapt</u> the work Under the following conditions: <ul style="list-style-type: none"> • Attribution—You must <u>attribute</u> the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work.) • Share Alike—If you alter, transform, or build upon this work, you may distribute the resulting work only under the same, similar or a compatible

			Wikipedia terms.		license.
6 1	Wind 2012 02	KDD Lab ISTI CNR	<p>The dataset contains mobile phone records collected in the province of Pisa in February 2012. It contains about 8 mln of Call Data Records (CDRs), and the antennas' coverage. Nr. of CDRs = about 8 mln; Nr. of users = about 230.000</p> <p>Content may fall under copyright and data protection law. Use of the content subject to the terms of initial data provider.</p>	Restricted access. Terms of use of Wind S.p.A.	To be identified by the data provider
6 2	Wind 2013 10	KDD Lab ISTI CNR	<p>The dataset contains mobile phone records collected in the provinces of Pisa, Lucca, Livorno and Firenze in October 2013. It contains about 60 mln of Call Data Records (CDRs), and the antennas' coverage.</p> <p>Content may fall under copyright and data protection law. Use of the content subject to the terms of initial data provider.</p>	Restricted access. Terms of use of Wind S.p.A.	To be identified by the data provider
6 3	Wind 2014 03	KDD Lab ISTI CNR	The dataset contains mobile phone records collected in	Restricted access. Terms of use of Wind	To be identified by the data provider

			<p>the provinces of Pisa, Lucca, Livorno and Firenze in March 2014. It contains about 50 mln of Call Data Records (CDRs), and the antennas' coverage. Nr. of CDRs = about 50 mln; Nr. of users = about 860.000; Nr. of antennas = about 450; Content may fall under copyright and data protection law. Use of the content subject to the terms of initial data provider.</p>	S.p.A.	
--	--	--	--	--------	--

REFERENCES

SoBigData

SoBigData, Data Catalog, available at:

<http://www.sobigdata.eu/private/ouhnr3inebchcerhncurhebcnhn8749892/datasets>

Horizon 2020 Call: H2020-INFRAIA-2014-2015 Topic: INFRAIA-1-2014-2015 Type of action: RIA Proposal number: 654024 Proposal acronym: SoBigData.

SoBigData, Exploratories & Stories, available at: <https://support.d4science.org/projects/sobigdata-eu/wiki>

Legal acts:

International law:

Berne Convention for the Protection of Literary and Artistic Works, 9 September 1886

International Convention for the Protection of Performers, Producers of Phonograms and Broadcasting Organizations, Rome, 26 October 1961

TRIPS Agreement, Appendix 1C of the Marrakesh Agreement Establishing the World Trade Organization, Marrakesh, Morocco, 15 April 1994.

WIPO Copyright Treaty (WCT), Geneva, 20 December 1996

UN Universal Declaration of Human Rights, 10.12.1948

UN International Covenant on Economic, Social and Cultural Rights, Adopted and opened for signature, ratification and accession by General Assembly resolution 2200A (XXI), 16 December 1966, entry into force 3 January 1976, in accordance with article 27

European law:

Council of Europe, Convention on Cybercrime, Budapest, 23 November 2001

Charter of Fundamental Rights of the European Union 2012/C 326/02, OJEU, C 326/391, 26 October 2012

Directive 2001/29/EC on the harmonisation of certain aspects of copyright and related rights in the information society, OJEC, L 167/10, 22 June 2001

Directive 2004/48/EC on the enforcement of intellectual property rights, OJEU, L 195/16, 2 June 2004

Directive 2006/116/EC on the term of protection of copyright and certain related rights (codified version), OJEU, L 372/12, 27 December 2006

Directive 96/9/EC on the legal protection of databases, OJEU L 077, 27 March 1996

Proposal for a Directive on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure, Brussels, 28 November 2013, COM(2013) 813 final, 2013/0402 (COD)

Directive 2009/24/EC on the legal protection of computer programs, OJEU L 111/16, 5 May 2009

Commission Regulation (EU) No 316/2014 on the application of Article 101(3) of the Treaty on the Functioning of the European Union to categories of technology transfer agreements Text with EEA relevance OJ L 93, 28 March 2014

National law:

French Intellectual Property Code (Code de la propriété intellectuelle (CPI))

German Copyright Act (Gesetz über das Urheberrecht und verwandte Schutzrechte)

German Basic Law (Grundgesetz)

German Civil Code (Bürgerliches Gesetzbuch)

Case law:

CJEU, Case C-200/96 *Metronome Musik v Music Point Hokamp*

CJEU, Case C-444/02 *Fixtures Marketing Ltd v Organismos prognostikon agonon podosfairou AE*

CJEU, Case C-203/02 *The British Horseracing Board Ltd and Others v William Hill Organization Ltd.*

CJEU, Case C-338/02 *Fixtures Marketing Ltd v Svenska Spel AB*

CJEU, Case C-304/07 *Directmedia Publishing GmbH v Albert-Ludwigs-Universität Freiburg*

CJEU, Case C 5/08 *Infopaq International A/S v Danske Dagblades Forening*

CJEU, Case C 145/10 *REC, Eva-Maria Painer v. Standard VerlagsGmbH, Axel Springer AG, Süddeutsche Zeitung GmbH, Spiegel-Verlag Rudolf Augstein GmbH & Co. KG, Verlag M. DuMont Schauberg Expedition der Kölnischen Zeitung GmbH & Co. KG*

CJEU, Case C 406/10, *SAS Institute Inc v World Programming Ltd*, ECLI:EU:C:2012:259.

Chancery Division, *University of London Press Ltd v University Tutorial Press Ltd* ([1916] 2 Ch 60)

Literature:

David I. Bainbridge, "Introduction to Information Technology Law", 6th ed., Harlow: Pearson Longman, 2008

Yochai Benkler, "From Consumers to Users: Shifting the Deeper Structures of Regulation Toward Sustainable Commons and User Access", 52 Fed. Comm. L. J. 561 (2000)

William Blackstone, “Blackstone’s Commentaries on the Laws of England” (first printed in 1765), P 2 The Rights of Things. Book II. Ch. i.

Jasper A. Bovenberg, “Property Rights in Blood, Genes & Data: Naturally Yours?”, Martinus Nijhof Publishers, 2006

Mark J. Davison, “The Legal Protection of Databases”, Cambridge Studies in Intellectual Property Rights, Cambridge: Cambridge Univ. Press, 2003

Estelle Derclaye, “The Legal Protection of Databases: A Comparative Analysis”, Edward Elgar Publishing, 2008

DESCA “Horizon 2020 Model Consortium Agreement Version 1.2”, March 2016

T. Dreier, “Regulating information: Some thoughts on a perhaps not quite so new way of looking at intellectual property” in: Technology and Competition - Technologie et concurrence, Contributions in Honor of Hanns Ullrich, Brüssel, 2009, p. 35 – 53

European Commission, “Communication from the Commission to the European Parliament, the council, the European Economic and Social Committee and the Committee of the Regions”, 2 July 2014

European Commission, “DG Internal Market and Services Working Paper – First evaluation of Directive 96/9/EC on the legal protection of databases”, 2005

Barbara J. Evans, “Much ado about data ownership”, (2011) 25 Harv. J. L. & Tech. 79 et seqq

Johann Gottlieb Fichte, “Beweis der Unrechtmäßigkeit des Büchernachdrucks“, 1793, nachgedruckt in UFITA 106 (1987) 155-172

Paul Goldstein, P. Bernt Hugenholtz, “International Copyright: Principles, Law, and Practice,” Oxford: Oxford University Press, 2012

Härtig, “Acht Thesen zum ‚Dateneigentum‘“, <http://www.cr-online.de/blog/2016/02/17/acht-thesen-zum-dateneigentum/>, 17.02.2016

Thomas Hoeren, Jonas Völkel: “Eigentum an Daten”, in: Big Data und Recht, München 2014, S. 11-37

Hogan Lovells International LLP, “Report on Trade Secrets for the European Commission, Study on Trade Secrets and Parasitic Copying (Look-alikes)”, MARKT/2010/20/D.

Hans D. Jarass, “Die EU-Grundrechte”, München 2005

Immanuel Kant, “Von der Unrechtmäßigkeit des Büchernachdrucks“, 1785, Nachgedruckt in UFITA 106 (1987) 137-144, mit einer Einführung von Hubmann

Wolfgang Kilian, “Strukturwandel der Privatheit“, in: Wovon – für wen – wozu. Systemdenken wider die Diktatur der Daten – Wilhelm Steinmüller zum Gedächtnis, Berlin 2014, S. 195-224

Le Chapeliers in den Beratungen des ersten französischen Urhebergesetzes von 1791: André Bertrand, "Le droit d'auteur", Paris 1991

Lawrence Lessig, "The Future of Ideas, The Fate of the Commons in a Connected World", New York 2002

Hans-Dieter Lippert, "Wem gehören Daten, die im Rahmen von Forschungsprojekten gewonnen werden?" in: Geistiges Eigentum: Schutzrecht oder Ausbeutungstitel? Springer, Volume 5, 2008

Klaus Lodigkeit, "Intellectual Property Rights in Computer Programs in the USA and Germany", Peter Lang GmbH, 2006

Carol M. Rose, "Possession as the Origin of Property", (1985) Faculty Scholarship Series, Paper 1830

Gilad Rosner, "Who owns your data?" in Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp '14 Adjunct). ACM, New York, NY, USA (2014), 623-628, 625. DOI=10.1145/2638728.2641679,

Diane Rowland, Uta Kohl, Andrew Charlesworth, "Information Technology Law", 4th ed, 2012

S. Stokes, "Digital Copyright, Law and Practice," 4. Edition, Hart Publishing, Oxford and Portland, Oregon, 2014

Jan Christian Sahl, "Brauchen wir ein Datenschutzrecht für Maschinendaten?", <https://berliner-datenschutzrunde.de/node/162>, 15.04.2015

Haimo Schack, "Zur Rechtfertigung des Urheberrechts als Ausschliesslichkeitsrecht", in Geistiges Eigentum: Schutzrecht oder Ausbeutungstitel? Springer, Volume 5, 2008

Jakob Schneider, "Menschenrechtlicher Schutz geistigen Eigentums", 2006

WIPO, "The Concept of Intellectual Property" in: WIPO Intellectual Property Handbook: Policy, Law and Use, Chapter 1, 2nd Edition, Geneva 2004

Herbert Zech, "Daten als Wirtschaftsgut – Überlegungen zu einem ‚Recht des Datenerzeugers‘", CR 2015, 137-146

Herbert Zech, "Information als Schutzgegenstand", Jus Privatum 166, 2012