Legal aspects of linked data – The European framework

Víctor Rodríguez-Doncel [a,*], Cristiana Santos [b], Pompeu Casanovas [b,c], Asunción Gómez-Pérez [a]

[a] Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain
[b] Institute of Law and Technology (IDT-UAB), Universitat Autònoma de Barcelona, Barcelona, Spain
[c] Faculty of Business and Law, Data to Decisions Cooperative Research Centre, Deakin University, Geelong, Australia

ABSTRACT

This paper portrays a general overview of the existing European legal framework that applies to the publication and consumption of linked data resources in typical settings. The point of view of both data publishers and data consumers is considered, identifying their rights and obligations, with special attention to those derived from the copyright and data pro-tection laws. The goal of this analysis is to identify the practices that help to make the publication and consumption of linked data resources legally compliant processes. An insight on broader regulations, best practices and common situations is given.

1. Introduction

The World Wide Web was born 25 years ago and it has changed the way humans access information. The key of its success lay largely in the general adoption of common practices and their ability to create a network of linked documents or *hypertext*. These practices were formalised by the World Wide Web Consortium (W3C) as "W3C Recommendations", namely, public specifications discussed by a broad community. For example, the version 5 of HTML is a W3C Recommendation published in 2014.

The W3C Consortium, still led by its founder Tim Berners-Lee, has published a new set of W3C Recommendations in the last few years towards the implementation of a "semantic web" of data. Much like the linkability of documents in the World Wide Web, the new web's most distinct feature is the ability to reference chunks of data eventually published by others. This *web of data* (term opposed to the original *web of documents*) is intended to be accessed not only by humans, but also and mostly by machines. The ultimate practices recommended by the W3C to publish data on the web refer to it as "*Linked Data*". Linked data is configuring a global data space of high quality data, strongly interconnected and with peculiar features that make it different from a database from the legal point of view.

For example, machines understand particularly well the information offered as linked data, because data models are typically specified by computer ontologies which enable automated reasoning. Linked data favours connecting distributed pieces of information, relating apparently disparate datasets. Linked data uses crowdsourced vocabularies, where no authority can impose the use of a specific variant. Linked data can be ephemeral, and it is hard to prove that some data was online at a precise time. In sum, a collection of *linked data* is

* *Corresponding author.* Artificial Intelligence Department, Universidad Politécnica de Madrid, Campus de Montegancedo s/n, Boadilla del Monte, 28660 Madrid, Spain.
E-mail address: vrodriguez@fi.upm.es (V. Rodríguez-Doncel).

not merely a database and its unique characteristics require special attention.

Linked data is gaining momentum. Many public institutions are publishing their datasets as *linked data* in open data portals, where anyone can download and reuse the information for commercial or non-commercial purposes, unleashing valuable resources for businesses, citizens, and other public administrations.[1] In Europe, the EU Open Data Portal,[2] the cornerstone of the EU open data strategy, provides access to datasets of different institutions (from Eurostats to national weather agencies) with a high penetration of linked data as the publishing format.[3] The *CELLAR* repository, the central component of the information system set by the European Publication Office, uses *linked data* to provide semantic indexing, advanced search and data retrieval for multilingual resources.[4] *Linked data* has also attracted attention in academic publishing,[5] governance development[6] and in the legal domain[7] (including legislation,[8,9]case law[10] and legal education[11]). Industrial organisations are publishing and consuming *linked data* in the operation of their business (as non-open linked data, also called *linked enterprise data*) and new business models are starting to take off, supporting every step of the lifecycle of *linked data*.[12] Thus, standardisation trends have received also a renewed attention, as within the emerging scenarios, rules and principles are at stake with users' behaviour, which may consume, discuss, follow or ignore them.[13]

So far as we are aware there, is no evidence thus far of case-law nor out-of-court disputes regarding *linked data* resources.[14] However, both in EU and USA, there is a large set of case-based decisions already dealing with the unintended effects of knowledge aggregation and profiling – which are capabilities uplifted by semantic web technologies. Conflicts have arisen in different sectors, including financing marketplace, health, and the areas of Freedom, Security, and Justice (which is quite intense in balancing exceptions to liberty rights and security threats[15]). Also, more conflicts are expected to appear as *linked data* is playing a more important role.

The objective of this article is to describe the applicable legal framework in Europe for *linked data*, identifying the applicable rights and obligations in common scenarios and facilitating the *linked data* publication and consumption processes to be legally compliant. The ascribable discipline of rights and obligations in relation to *linked data* arise from legal instruments, such as copyright law, database law, trade secrecy law or data protection law which are briefly examined in this paper.

Section 2 presents linked data resources and features and its role within the Semantic Web. Section 3 analyses the legal framework of linked data in Europe, describing the applicable law with respect to copyright, data protection and other legal statutes, like competition law or trade secrecy law. Section 4 concludes with a review of the overall legal aspects and presents the forthcoming initiatives on the analysed legal framework which shall affect the linked data universe.

2. Linked data

2.1. *Description of linked data*

The definition of *linked data* is often given as the result of adopting *a set of best practices for publishing and connecting structured data on the Web*.[16] These practices can be summarised in:

(i) URIs[17] are used as identifiers of the resources. A resource can be a city, a person or anything there is data about. A URI can be for example http://en.wikipedia.org/wiki/London.

[1] Archer P. et al. (2014). Study on business models for Linked Open Government Data. Technical Report, BM4LOGD, https://joinup.ec.europa.eu/node/72473 (visited April 2016).

[2] The European Union Data Portal is available at https://open-data.europa.eu/. As of April 2016, it accounted 8546 different datasets. A notable precedent was similar intent was the portal http://PublicData.eu, developed by the Open Knowledge Foundation and still online.

[3] As of April 2016, the portal accounted that 46% of the datasets were already machine readable. http://www.europeandataportal.eu/mqa-service.

[4] Francesconi, E., Küster, M.W., Gratz, P., & Thelen, S. (2015). The Ontology-Based Approach of the Publications Office of the EU for Document Accessibility and Open Data Services. In Electronic Government and the Information Systems Perspective, pp. 29–39. Springer Int. Publishing.

[5] Peroni, S. Semantic Web Technologies and Legal Scholarly Publishing, vol. 15, LGT Series, Springer, Dordrecht (2015).

[6] Davies, T. and Edwards, D. (2012). Emerging Implications of Open and Linked Data for Knowledge Sharing in Development, IDS Bulletin 43 (5), 117–127.

[7] Casanovas, P., Palmirani, M., Peroni, S., van Engers, T., Vitali, F. (2016). Special Issue on the Semantic Web for the Legal Domain Guest Editors' Editorial: The Next Step, Semantic Web Journal, 7(2): 213–227.

[8] World Legal Information Institute, http://www.worldlii.org/ (visited April 2016).

[9] Casellas, N., Bruce, T.R., Frug, S.S., Bouwman, S., Dias, D., Lin, J. & Venkataraman, S. (2012). Linked legal data: improving access to regulations. In Proc. of the 13th Annual Int. Conf. on Digital Government Research, pp. 280–281, ACM.

[10] The project EU Cases (http://eucases.eu/, visited in April 2016) offers case law as linked data.

[11] Casanovas, P. (2012). Legal crowdsourcing and relational law: what the Semantic Web can do for legal education. In Journal of Australian Law Teachers Association, Vol. 5 (1 & 2) 159–176.

[12] Auer S. et al. (2012). Managing the Life-Cycle of Linked Data with the LOD2 Stack. International Semantic Web Conference (2), vol. 7650 of Lecture Notes in Computer Science, pp. 1–16. Springer.

[13] Polleres A. (2013) Agreement Technologies and the Semantic Web. In S. Ossowski, ed. Agreement Technologies, vol. 8 of LGT Series, pp. 57–68. Springer, Dordrecht.

[14] Search at Lexis Nexis and West Law legal databases made as of mid 2015.

[15] Boehm, F. (2012) Information Sharing and Data Protection in the Area of Freedom, Security and Justice, Springer, Dordrecht.

[16] Bizer, C., Heath, T., Berners-Lee, T. (2009). Linked data-the story so far. Int. J. on Semantic Web and Information Systems, vol. 5(3), pp. 1–22.

[17] A URIs is a string of characters used to identify a resource on the Internet (like mailto:john@example.com or http://example.com). The specification is defined in "RFC 3986 Uniform Resource Identifier (URI): Generic Syntax".

(ii) URIs resolve; namely, when typed in a web browser they return meaningful information about the resource. Technically this is done by establishing a HTTP connection; HTTP requests can specify in their headers whether information is to be consumed by a human (expecting then a well laid out HTML webpage as the response) or by a machine (expecting then data in a machine readable format like RDF or JSON).

(iii) Local resources (data published in one's internet domain) are connected to external resources (published by others in another domain); much like web pages have hyperlinks to other external web pages.

(iv) The unit of information is the RDF statement, namely, a triple made of a subject, a property and an object. The subject and the property are also resources (URIs), while the object can be either a resource or simple data (e.g. a number, a text string). This fourth recommendation is sometimes relaxed, and other forms like JSON (or JSON-LD[18]) are accepted.

The atomic unit of information in linked data, the RDF triple, can express things like "*Alice is 6 feet tall*", provided that Alice is a resource identified by a URI (for example http://example.com/Alice), the height is also a property identified by a URI (for example http://example.com/hasHeight) and 6 is a constant (technically a "literal"). URIs can be shortened by using namespaces (e.g. "http://example.com/" can be reduced to "ex:"), and make easier the expression of the triple:

<ex:Alice, ex:hasHeight, "6">

Further RDF triples can be stated about the resource 'Alice' (like her friends, id or bank account number), creating a meaningful unit of information. Bigger collections of triples (for example information about the professors in a certain University) are known as *datasets*. Datasets are usually accessible per resource (querying Alice's URI), as a data file (an RDF dump with the whole set of professors) or through specific access protocols (like SPARQL[19]).

Some resources and properties are recurrently used by linked data adopters. For example, the URI "foaf:knows" is used to express friendship relationships. This property is formally defined as "*A person known by this person (indicating some level of reciprocated interaction between the parties)*".[20] These well-known properties and resources are usually published in collections covering a certain domain. They are referred to as *vocabularies* if information about them is simple (e.g. giving human readable definitions for each term) or as *computer ontologies*[21] if a complex model is specified. For example, the property "ex:hasHeight" used in the triple above might be further described in a computer ontology by declaring the unit of measurement (feet or centimetres) or by specifying that the property is *functional* (namely, at most one value can be given for one individual). Computer programs can process this information and react intelligently.

Some of the triples within a dataset relate resources in the local domain (under control by the linked data publisher) to resources in external domains. For example, the triple:

<ex:Alice, foaf:knows, ex2:Bob>

would express that Alice (resource identified within the domain ex) knows Bob (identified by a URI in another domain, like http://example2.com/Bob). Very often the declared relationship is an identity match: "*They URI identifying Alice within the faculty refers to the same person identified by another URI within her club*". A collection of these triples used to match resources is known as *mapping* or *linkset*.

Semantically-enabled applications use these resources to provide advanced functionalities, most notably improving search portals, information retrieval algorithms and in general almost every natural language processing task.

Along this paper, we will consider linked data resources to be in a wide sense: (i) *data models* (ontologies, vocabularies, thesauri, etc.); (ii) *linked data itself* (RDF dumps, data accessible via SPARQL endpoints, RDF mappings, etc.) and (iii) *semantic services and tools* (ontology-based applications, semantic services, etc.).

2.2. *Generation of linked data*

The super-set of connected datasets of linked data is known as the *Linked Open Data* cloud. Four billion triples have been estimated to be online.[22] Data in the web can be dumped by anybody, consequently its quality and trustfulness are variable, the availability not granted and its maintenance irregular – features all shared by documents in the world wide web. Yet, *linked data* is better described and more meaningful than any other kind of data in the big data revolution. Further, more refined guidelines are appearing and standard practices are being followed to generate linked data. The complete lifecycle of a linked dataset, as described by Villazón et al.,[23] is represented in Fig. 1.

[18] JSON-LD is a lightweight Linked Data format based on JSON (JavaScript Object Notation) a structured format similar in intention to XML.

[19] SPARQL is an RDF query language specified by W3C at https://www.w3.org/TR/sparql11-overview/.

[20] The definition is specified at http://xmlns.com/foaf/spec/.

[21] The most quoted definition of "ontology" is "*an explicit specification of a conceptualization*". Gruber, T.R. (1993). A translation approach to portable ontologies. In Knowledge Acquisition, 5(2), pp. 199–220. An ontology is an abstract, simplified view of the world of a system that has to be represented for some purpose, reached after consensus, and formalised with a machine readable language, typically OWL. Taxonomies and vocabularies are organised lists of terms, with sound textual definitions and a hierarchical structure.

[22] Käfer, T. and Harth A. (2014) Billion Triples Challenge data set, Study online at http://km.aifb.kit.edu/projects/btc-2014/ (visited in April 2016).

[23] Villazón-Terrazas, B, Vilches-Blázquez, L.M., Corcho, Ó. and Gómez-Pérez, A. (2011) Methodological Guidelines for Publishing Government Linked Data. Linking Government Data. D. Wood (Ed.) Part 1, pp.: 27–49. Springer.

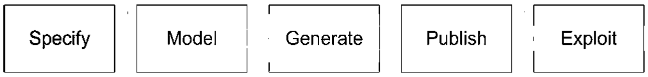

Fig. 1 – Linked data life-cycle.

After the specification of requirements, a data model is designed (typically with ontologies), and the dataset is accommodated to that model. The data is then published and exploited. Different options for the generation and publication of linked data are shown in Fig. 2. Linked data is generated either from existing structured data or from text analysed by an entity extractor. Transformations are made from previous sources like CSV, XML or relational databases to RDF. This RDF is stored in specific stores (triplestores), in RDF dumps or in other forms; and then it is offered in various forms: linked data interfaces, SPARQL endpoints, downloadable files, etc.

3. Legal rights in relation to linked data

Data has been said to be legally inert in itself, being "*more accurate to speak of the different kinds of rights that may arise* 'in relation to' *data, rather than simply of rights* 'in' *data*".[24]

The legal qualification of linked data resources (RDF datasets, ontologies, etc.) depends on their nature: data may consist in a literary work and abide copyright protection; may contain the attributes of a database to attract database right; may have the quality of confidentiality to enable enforcement from trade secrecy law or even convey personal information and recall data protection law. In a theoretical stance, the intersection of rights with the process of collecting, analysing and disseminating data might increase the ensuing recognition of rights: we aim to analyse this layered nature of rights and obligations in relation to linked data.

The most important supranational legal sources relevant for Europe considered for this analysis are listed in Table 1.

However, this list cannot be considered exhaustive, for many reasons.

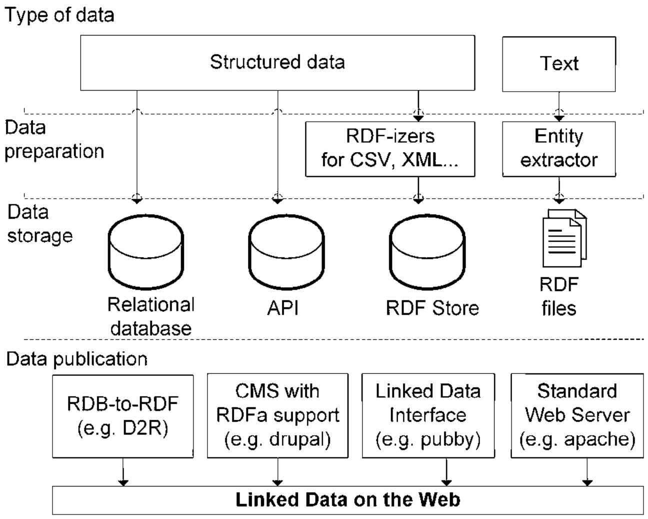

Fig. 2 – Linked data publishing options and workflows. Adapted from Heath and Bizer (Heath T., Bizer, C. Linked Data (2011). Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1–136. Morgan & Claypool).

First, because laws affecting linked data are changing or likely to change in Europe in a near future. The entire DP Directive is being replaced by the so-called General Data Protection Regulation package (GDPR) since 2012. The EU Parliament has very recently passed[25] the Regulation after four years of intense negotiations, and will be put in force in 2017.[26] The Copyright Directive is likely to suffer transformations driven by the EU Digital Single Market Strategy.[27] This strategy aims at creating a single market to boost Europe's digital economy, and it leans on three pillars (better online access to digital goods, an environment where digital networks can prosper, and 'digital' as a driver for growth), in one way or another related to linked data. The importance of the way in which data is offered has been explicitly acknowledged by EU legislators. This is quite clear since the PSI Directive brought about that public information had to be published in interoperable,[28] machine-readable[29] formats complying with open standards, and facilitating cross-linguistic search in data portals.[30] Formal open standards are not completely settled yet, and legal interoperability across EU countries constitutes an issue to be precised in the immediate future.[31] The PSI does not explicitly mention linked data but the reference is almost direct – especially when linked data is the champion technology to fa-

[24] Kemp, R. (2012) Big Data – Legal Rights and Obligations, Kemp Little Report (2012).

[25] The General Data Protection Regulation (GDPR) passed on 2016 April 14 in Strasbourg.

[26] See http://ec.europa.eu/justice/newsroom/data-protection/ (visited on April 2016).

[27] For example, in December 2015 a proposal for a regulation "on ensuring the cross-border portability of online content services in the internal market" was published (COM(2015) 627 final 2015/0284 (COD)).

[28] "To facilitate re-use, public sector bodies should, where possible and appropriate, make documents available through open and machine-readable formats and together with their metadata, at the best level of precision and granularity, in a format that ensures interoperability, e.g. by processing them in a way consistent with the principles governing the compatibility and usability requirements for spatial information under Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)". PSI, Recital 20.

[29] "Public sector bodies shall make their documents available in any pre-existing format or language, and, where possible and appropriate, in open and machine-readable format together with their metadata. Both the format and the metadata should, in so far as possible, comply with formal open standards." Ibid., Art. 5.

[30] "Member States shall make practical arrangements facilitating the search for documents available for re-use, such as asset lists of main documents with relevant metadata, accessible where possible and appropriate online and in machine-readable format, and portal sites that are linked to the asset lists. Where possible Member States shall facilitate the cross-linguistic search for documents". Ibid., Art. 9.

[31] See https://joinup.ec.europa.eu/ for recent initiatives (visited April 2016).

Table 1 – Legal sources of interest to assess the legal status of linked data in Europe.

Short	Document
Berne Convention (BC)	Berne Convention for the Protection of Literary and Artistic Works, of September 9, 1886, completed at Paris on May 4, 1896; revised at Berlin on November 13, 1908; completed at Berne on March 20, 1914; revised at Rome on June 2, 1928; at Brussels on June 26, 1948, at Stockholm on July 14, 1967, and at Paris on July 24, 1971, and amended on September 28, 1979
WCT	WIPO Copyright Treaty, 1996
TRIPS Agreement	Agreement on Trade-Related Aspects of Intellectual Property Rights, 1994
Copyright Directive (CP)	Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society
Computer Programs Directive	Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs
Database Directive	Directive 1996/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases
Data Protection (DP) Directive	Directive 1995/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data [repealed by the new General Data Protection Regulation on 14/04/2016]
Proposal for Police and Criminal Justice Data Protection Directive	Directive on the protection of individuals with regard to the processing of personal data by competent authorities for the purposes of prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and the free movement of such data and repealing Council Framework Decision 2008/977/JHA (05418/1/2016 – C8-0139/2016 – 2012/0010(COD)) – SEC(2012) 72 final [still pending]
General Data Protection Regulation (GDPR)	Regulation (EU) 2016/. . . of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Brussels, April 6th, 5419/16, approved on 14/04/2016.
Public Sector Information (PSI) Directive	Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information
Convention 108	Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, CETS No.108, 28/01/1981
Damages Directive	Directive 2014/104/EU of the European Parliament and of the Council of 26 November 2014 on certain rules governing actions for damages under national law for infringements of the competition law provisions of the Member States and of the European Union
Reuse of information Directive	Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information

cilitate interoperability. Finally, the delicate balance between freedom and security that the EU is experiencing right now may have an impact on the legislation which is relevant to the linked data sector.

The *second* reason lies in the fact that national legislative acts apply differently for each EU country (for instance, rights of access to information and data protection rights are not uniformly distributed). For example, classification of documents and disclosure of information affecting security, present different requirements for each country linked to public Legal Enforcement Agents (LEA) policies, the organisation of Government Agencies and Public Administrations, and the structure of the judiciary.

Finally, the *third* reason for a non-exhaustive account lies in that linked data technology is young, it is still developing and it may raise new challenges, especially if used in connection with other disciplines in the Artificial Intelligence field. We contend that linked data and semantic web technologies might play an important role in the manner that regulations themselves are issued and in the notions of e-governance, legal compliance, and participatory tools. From 2012 onwards, the EU *Better Regulations* (later on, *Smart Regulation*) initiative is covering the whole policy cycle – planning, adoption, design, implementation, application (including enforcement), evaluation and revision.[32] Smart Regulation "*ensures that policy is prepared, implemented and reviewed in an open, transparent manner, informed by the best available evidence and backed up by involving stakeholders*".[33] EU-supported research projects like Openlaws,[34] EUCases[35] or BO-ECLI are already publishing existing legislation, case law and other legal documents as linked data, and making use of the European Law Identifier (ELI) and the European Case Law Identifier (ECLI) – having stable URIs with official support are the prelude to producing linked data.

If drafts and intermediate documents discussed in the process of making new regulations used linked data technologies, they would be better accessible and more transparent. Documents represented in linked data can be better searched, better referenced, and the provenance of each modification better traced. These properties are much in line with the Smart Regulation objectives.

The Smart Regulations principles have been brought about to facilitate citizens' participation and collective consultation. Drafters had in mind social interventions. From this participatory perspective, the linked data principles could be viewed as another kind of P2P or Web Services use, in which the creator, administrator, and server of the database link with other stakeholders (consumers, companies, agencies, administrators) through the chosen licence or type of relational

[32] See http://ec.europa.eu/smart-regulation (visited April 2016).

[33] Ibid.

[34] http://www.openlaws.eu (visited April 2016).

[35] http://eucases.eu/ (visited April 2016).

contract. The same holds for end-users choosing a type of licence or a type of protection for their personal data. It is an individual decision in this case, connected to some effects through rules and the choice of rights. If we interpret this kind of regulatory behaviour through the Better Regulations lens, we would be able to frame it legally into the social ecosystems in which linked data is embedded. There is some work still to be carried out, as Principles of Common Better Regulation on wide participation throughout the policy cycle (subsidiarity, proportionality, coherence, expected impact, transparency, objectivity, and balanced intervention)[36] are blatantly similar to Alan Westin's FIPs (*Fair Information Practices*) and Ann Cavukian's PbD (*Privacy by Design*)[37] principles. A comparative table could be drawn,[38] but it would fall beyond the scope of this article. However, it is worth to be noted that under the new GDPR, *democratic governance principles, professional best practices, and global ethics will enrich what counts as EU law*.[39] This is particularly important for the linked data field, where technology is likely to move faster than specific legal provisions.

3.1. *Copyrights in linked data*

3.1.1. *Linked data resources as copyrightable works*

Patents protect technical inventions. In Europe, no linked data resource is patentable *per se*, and it might only be patented in case it is connected with a computer implemented invention solving a technical problem.[40] This is different under US laws, which have more relaxed criteria to accept software patents. Some of the patentable inventions in the realm of Big Data, and applicable to linked data, have been pointed out by Lehrer.[41]

Copyright is a legal entitlement that renders a form of protection to the authors of "original"[42] works. Following the traditional civil analysis, only those endeavours that are *intellectual creations* can be qualified as *works*.[43] This has been maintained since the original Berne Convention in 1896.[44]

Many linked data resources can be approached as intellectual creations tractable as copyrightable works. Copyright protects the *form* in which a creation is expressed – e.g. a computer program is expressed through its source code. In order to apprehend if linked data resources benefit from copyright protection, and thus assert the correspondent rights and obligations endorsed by its authors, it is necessary to account the legal norms and its requisites on the protectability of works.

The recognition of ontologies or other linked data resources as categories of literary and artistic work is a matter of interpretation of Art. 2 (1) of the BC (as given in the WIPO "Guide" to Copyright,[45] §2.19) and such a legal characterisation is delegated to national legislature and judicature. We contend that producing an *ontology*, a *mapping* or an *RDF dataset* might well entail an intellectual creation attributable to its author. In particular, creating an *ontology* is not always an automatable task, and there is usually a creator of original works with a creativity component remarkable enough as to invoke copyright law. Conversely, *typical content of data* (even with little originality) does not fall under the copyright protection law and, consequently, is not constrained to legal barriers to be exposed as open data. As an example of ontology, the Music Ontology is attributable to a person, original and with a large share of creativity, but the Linked Movie Database is a collection of automatically harvested metadata; the former is an ontology probably protected by copyright, but the latter is a dataset not protected by copyright. In practice, data models, such as *ontologies*, *vocabularies*, *thesauri*, and *RDF datasets* may conform to the requirements uttered above and hence bestowed with copyright law protection.

[36] Guidelines http://ec.europa.eu/smart-regulation/guidelines/tool_1_en.htm (visited April 2016).

[37] The "Privacy by design" principle has not been immune from criticism. cf. Koops, B-J, Leenes, R. Privacy regulation cannot be hardcoded. A critical comment on the 'privacy by design' provision in data-protection law, International Review of Law, Computers & Technology, Vol. 28 (2) Issue 2, pp. 159–171 (2014).

[38] Casanovas, P., Zeleznikow, J. Online Dispute Resolution and Models of Relational Law and Justice: A Table of Ethical Principles. In P. Casanovas et al. (eds.), AI Approaches to the Complexity of Legal Systems IV. Social Intelligence, Models and Applications for Law and Justice Systems in the Semantic Web and Legal Reasoning, LNAI 8929, Springer, pp. 55–69 (2014). Also in P. Casanovas (2015), Semantic Web Regulatory Models: Why Ethics Matter. Philosophy & Technology 28, Issue 1, pp. 33–55.

[39] See art. 23 of the GDPR on restrictions on specific principles and on the rights of information, access, rectification and erasure, the right to data portability, the right to object, decisions based on profiling, as well as on the communication of a personal data breach to a data subject and on certain related obligations of the controllers: "*the prevention, investigation, detection and prosecution of breaches of ethics for regulated professions*" (art. 23g). For example, Recital 4 states that "*the processing of personal data should be designed to serve mankind*".

[40] '*the following in particular shall not be regarded as inventions [. . .] schemes, rules and methods for performing mental acts, playing games or doing business, and programs for computers*'. Art. 52.2, European Patent Convention.

[41] Lehrer, J.E. "Patenting Big Data", Mondaq Business Briefing (Aug. 20, 2012) http://www.mondaq.com/unitedstates/x/192640/Patent/Patenting+Big+Data (accessed April 2016).

[42] The postulate of "*originality*" imputes an individual intellectual creation originated from the author, "*reflecting the personality of the author*" (Guide, §2.8), irrespective of the concept of novelty/innovation.

[43] Art. 2 of the BC defines and lists (non-exhaustively) the types of literary and artistic works for which protection is endowed with: "*[A]ll intellectual creations irrespective of whether they may be regarded as belonging to the literary domain, to the artistic domain or to both at the same time*" [Guide03§2.15]. The term 'literary' has to be understood in a very wide sense as "*information-oriented productions expressed in letters, numbers or any other similar symbols, irrespective of whether they are legible for everyone or are coded (and thus available only to those who know and may use the code, or through the use of appropriate equipment)*". Ficsor, M.: Guide to the Copyright and Related Rights Treaties Administered by WIPO and Glossary of Copyright and Related Rights Terms, Geneva WIPO (2003).

[44] The Berne Convention has been internally amended many times, the last one in 1978. The *WIPO Copyright Treaty* amended the *Berne Convention* to include the copyright protection of computer programs as literary works (Art. 2 of the *Berne Convention*) in 1996. The TRIPS Agreement did not entail a formal amendment of BC, but supplemented it in 1994.

[45] Guide to the Copyright and Related Rights Treaties Administered by WIPO and Glossary of Copyright and Related Terms, WIPO Publication No. 891(E) (2003).

Linked *datasets* can be designated as "databases", according to the Database Directive's terminology – in Art. 1 (2) of the Directive, a definition of a database is provided in a broad and technology-neutral way, 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*". We assume that given the broadness of the database definition in the Directive, the term dataset falls under this definition and will be used interchangeably with the term database.

The copyright protection of *databases* is refracted in the BC, in Art. 5 of the WCT, in the TRIPS Agreement (Art. 10.2) and in the Art. 3 of the Database Directive, which provides that databases, by reason of the selection or arrangement of their contents, constitute the author's own intellectual original[46] creation and are protected by copyright. Thus, *individual facts* and *individual data* (e.g. single RDF statements) cannot be object of copyright protection. On the contrary, *RDF datasets* are, in principle, object of copyright protection. Nevertheless some *RDF datasets* and *mappings* have been merely created very often by transforming data from already existing databases. For example, a dataset with the temperatures acquired from a weather station, or the list with the iTunes top-selling songs should not be categorised as literary or artistic works. Therefore most of the *RDF datasets* and *mappings* are not eligible for copyright protection. Also, *RDF statements* that are inferred from the axioms of an ontology will not afford from copyright protection.

On occasion, rights related to the authors of works may possibly be bundled within the linked data resource. For example, the creator of a collection of literary work codified as RDF would warrant rights on the collection, but each of the included works in the collection would be still independently protected.

Community-based databases (for example Dbpedia, the linked data version of Wikipedia) are well analysed by Kierkegaard and Adrian,[47] and typically imply the acceptance of the site maintainer's conditions, hence transferring or waiving the exploitation rights; they also benefit from copyright law.

3.1.2. *Protection of linked data resources under copyright law*

The author of a copyrightable work (*creator of any work whose intellectual creative activity brings such works into existence* (Guide §2.60)) is endowed with *moral* and *exploitation rights*. *Moral rights* allow the author to take certain actions to preserve the personal link between himself or herself and the work. Its recognition is mandated by BC (Art. 6 bis) and it comprises the rights of:

(i) *attribution* or *paternity*: the right of the author to "*claim authorship*" of the work. On this basis, the author has the right to be identified as the author of a work, the right of not being falsely attributed as the author of another work, as well as the right for the author to remain anonymous;

(ii) *integrity* or *respect*: the right "*to object to any distortion, mutilation or other modification of, or other derogatory action in relation to, the said work, which would be prejudicial to his honor or reputation*".

The author of a work may pursuit his or her *exploitation rights*[48] (which allow the rights owner to derive financial reward from the use of his or her works by others), comprising:

(i) *communication* and *making available to the public*: right of the owner of the work to authorise or prohibit any communication to the public, irrespective of the tool used. Uploading a linked data resource (an ontology, a dataset) into a webpage and opening its access would be an example of the exercise of this right;

(ii) *adaptation*: right to authorise or prohibit adaptations, arrangements and other alterations of the works (Guide §12). It refers to a combination of the pre-existing elements of the works with some new ones, as a result of which normally a new work emerges. The new work is protected as a derivative work, as provided for under Art. 2 (3) of the Convention. Converting a format, extending an ontology, transforming a computer program into another computer language, or from source code into object code, configure examples of the right of adaption. In order to publish an adapted work, authorisation must be obtained both from the owner of the copyright in the original work and from the owner of copyright in the adapted work;

(iii) *distribution*: the right to authorise or prohibit any form of distribution of their works by sale or otherwise. For example, this would be the case of selling a compact disc with a protected ontology;

(iv) *reproduction*: the right to authorise or prohibit direct or indirectly, temporary or permanently copies of works. The definition of this right was made in a time when no digital technologies existed, and reproducing content was a costly process capable of generating profit: distributors had to buy physical copies. Nowadays, this right is a vestige with hardly any meaning for digital works. A current interpretation of the reproduction right in the digital environment asserts that every use of a work entails a reproduction thereof,[49] and thus reproduction should no longer implicate infringement. A conceptual shift from a system based on the concept of reproduction (copy-right) to a system based on the reutilisation of works (reuse right) is inquired through the prism of linked data.[50]

[46] Jurisprudence of the ECJ has been filling up the criteria for the assessment of originality in databases, as to the "*imprint of the personality of the author*" (ECJ case C-145/10) rendered in the composition, design, and tools elicited for the database.

[47] Kierkegaard, P. Adrian, A.: Wikitopia: Balancing intellectual property rights within open source research databases, Computer Law & Security Review 26 (5) pp. 502–519 (2010).

[48] Each country defines exploitation rights in a slightly different manner, but the exposed ones, extracted from the Berne Convention and from the Copyright Directive are almost universal.

[49] De Filippi P.: The Concept of a Digital Copy. Proc. of the 15th International Legal Informatics Symposium (IRIS): Transformation of Legal Languages, Salzburg, Austria (2014).

[50] De Filippi, P. and Gracz, K.: Resolving the Crisis of Copyright Law in the Digital Environment: Reforming the 'Copy-Right' into a 'Reuse-Right'. In Proc. of the 7th Int. Conf on the Interaction of Knowledge Rights, Data Protection and Communication (2013).

(v) *translation*: exclusive right of making and of authorising the translation of their works throughout the term of protection of their rights in the original works. A common case in the semantic web would be translating the definitions in an ontology or the literals of an RDF dataset.

Considered as rights *in rem*,[51] the listed rights are enforceable against the whole world and entail, with infringement remedies, from temporary to permanent injunctions.

Works are protected under copyright for a limited time frame and after the term of protection has expired a work falls into the public domain and can be *freely* used by anyone without requiring copyright compliance (in accordance with the applicable national rules on moral rights). However, the current terms of copyright protection are still quite long and not appropriated to the digital environment: only few ontologies or linked datasets may have actual value after a few years.

3.1.3. *Ontologies as computer programs*

Computer programs, whose source code is considered a sort of intellectual creation, receive special attention in copyright law. *Ontologies* might be understood as computer programs, especially if they contain *rules* for modelling a domain, if they are machine-enforceable, or have declarative constraints on their well-formed use. These features enable the matching between the concept of ontology with the computer program as defined by WIPO: "*a set of instructions expressed in words, codes, schemes or in any other form, which is capable, when incorporated in a machine-readable medium, of causing a 'computer' – an electronic or similar device having information-processing capabilities – to perform or to achieve a particular task or result*".[52] Computer programs are considered as a sub-category of "literary works", whatever may be the mode or form of their expression, as depicted in the TRIPS Agreement (Art. 10.1), in the WCT (Art. 4) and in the Computer Programs Directive (Art. 1).

3.1.4. *Database right*

The Database Directive entitles the author of a database to a *sui generis right* for non-creative works, regarding databases in respect of which there has been qualitatively and/or quantitatively substantial investment in either the obtaining, verification or presentation of its contents (Art. 7 (1)). Thus, to most *RDF datasets* and *mappings* in Europe, the Database Directive applies. However, no directly equivalent right is found in the United States, Japan, China or Australia.[53] The Database Directive confers on qualifying databases the rights of extraction and re-utilisation:

(i) *extraction*: the transfer of data to another medium as a whole or a substantial part thereof (e.g. downloading a dataset);

(ii) *re-utilisation*: any form of making available to the public the database. For example, even if a person is authorised to use a *RDF dataset*, this cannot be communicated to the public or to a third party.

A database right is infringed if a person, without the owner's consent, "extracts or re-utilises all or a substantial part of the contents of the database", or carries out "repeated and systematic extraction or re-utilisation of insubstantial parts of the contents of a database". In our case, the extraction of *individual RDF triples* is allowed, unless they convey substantial information (in terms of quantity or quality, or a combination of both).

Different term-lengths are applied in the different Member States. According to the database duration, the first generation of a database right is of 15 years duration from the end of the year when the database was completed, as depicted in Art. 10.

3.1.5. *Limitations and exceptions*

In order to maintain a balance[54] between freedom of expression and information in the digital environment (incentive and access trade-offs), copyright-protected works, due to its non-absolute nature, may in some cases be used without the authorisation of the authors or rights holders and with or without compensation. A *three-step test*[55] operates as the criteria to constrain the kinds of copyright exceptions and limitations which each State enacts. An exception or limitation to copyright is permissible only if:

(i) it covers only special cases (copies made for scientific research and teaching purposes; uses in educational institutions or quotation);

(ii) it does not conflict with the normal exploitation of the work and;

(iii) it does not unreasonably prejudice the legitimate interests of the author.

The precise definition of exceptions and limitations reside within the national legislature and thus vary substantially in number and scope (due to particular social, economic and historical conditions); this differentiation originates legal fragmentation and uncertainty, and has compounded effects on the functioning of the digital single market, considering the development of cross-border activities, for instance, regarding the question of fair compensation of right holders. In some instances, Member States are obliged to compensate right holders for the damage inflicted by a limitation or exception to their rights, whereas in other jurisdictions they are not obliged to do so, but may decide to provide for such compensation.

[51] Each right – copyright, database right, confidentiality and trademarks – is enforceable as an intellectual property right against the world (as a right *in rem*). See also Kemp, R., Hinton, P., Garland, P.: Legal rights in data. Computer Law & Security Review, Vol. 27(2), Apr. 2011, pp. 139–151 (2011).

[52] The foregoing is taken from the definition included in the WIPO "Model Provisions on the Protection of *Computer Software*" adopted in 1978, 6–19.

[53] Wilks, J., Christie, A. (2013) Rights in Data Handbook, Tech. rep., DLA Piper Technical Report.

[54] Acknowledged in the preamble of WCT or in recital 31 of the Copyright Directive.

[55] See Art. 10 of WCT, Art. 9(2) of the BC, Art. 6(3) of the Computer Programs Directive, Art. 6(3) of the Database Directive and Art 5(5) of the Copyright Directive.

For our purposes, we mention some exceptions related to the communication rights, as we consider more allied to linked data: (a) teaching and non-commercial scientific research; (b) use by disabled persons; (c) news reporting; (d) criticism or review; and (e) caricature, parody or pastiche, as consigned in the Art. 5(5) of the Copyright Directive.

3.2. *Data protection and linked data*

In this subsection, we investigate the current landscape of linked data with respect to the DP Directive. We should bear in mind here what we said above in regard to the enactment of the General Data Protection Reform package (2012–2016), which will substitute the still valid Directive 95/46/EC. It has been a long ongoing process until completing the new Regulation draft.[56] Rights, soft law (Privacy Impact Assessments) and more sanctions will be put in place, new concepts – such “genetic data”, “biometric data”, and “data concerning health” – are added, but the main guidelines remain.[57] The wording of Draft Regulation defines *personal data* as “any information relating to an identified or identifiable natural person ‘data subject.’ ”[58]

Linked data is not necessarily published data; there is linked data not publicly available but still benefitting from its principles (i.e. a strongly linked piece of information in RDF possibly accessible from a restricted domain). This linked data may contain personally identifiable information, in which case data protection laws apply. There is no evidence of enterprises currently using linked data at massive scale to internally store personal data, but this is a situation important enough as to consider. The main question of the 2015 Linked Data Awareness Barometer (“*Is Linked Data as mature enough to be used on a large scale in enterprises?*”) is still pending.[59] We advance with domain definitions, such as personal and sensitive information, its corresponding obligations towards its protection, and potential risks scenarios fuelled with the use of *linked data*.

3.2.1. *General aspects*

Personal data are defined (both in the Directive and in the Convention 108) as *information relating to an identified or identifiable natural person*.[60] Both types of information (identified and identifiable information) are protected in the same manner. Identification requires elements which describe a person in such a way that he or she is distinguishable from all other persons and recognisable as an individual. Possible identifiers may be: “a name, a photo, an email address, bank details, posts on social networking websites, medical information, or a computer’s IP address”.[61]

A special category of personal data is *sensitive data* (as depicted in Art. 8 of the Directive, and Art. 6 of the Convention 108) and according to this definition, it is illegal to process personal data revealing special features: racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and data concerning health or sex life.

A significant part of the Linked Open Data cloud can be now considered personal data, although sensitive data is less frequent. User generated content and information from the social networks (*Facebook*, *Twitter*, *YouTube* or *LinkedIn*) represents roughly the 50% of the Linked Open Data cloud[62] – large parts of it constituting personal data records. Semantic Web enthusiasts publish their personal FOAF profiles as RDF, this data being publicly accessible. FOAF declare simple personal information, like this excerpt extracted from Tim Berners-Lee profile[63]:

```
<http://www.w3.org/People/Berners-Lee/card.rdf> rdf:type
owl:NamedIndividual,
foaf:PersonalProfileDocument;
dc:title “Tim Berners-Lee’s FOAF file”;
cc:license <http://creativecommons.org/licenses/by-nc/3.0/>;
foaf:primaryTopic <https://www.w3.org/People/Berners-Lee/card#i>.
<https://www.w3.org/People/Berners-Lee/card#i> rdf:type
con:Male;
foaf:mbox <mailto:timbl@w3.org>;
foaf:phone <tel:+1-(617)-253-5702>;
```

With respect to a personal data file, the DP Directive identifies the roles of *data subject* (the person about which data has been collected), *data controllers* and *data processors*. *Data Controllers* (Art. 2 (d)) are those who, either alone or with others (joint controllership), control the contents and use of personal data (purpose and means of processing). They can be

[56] See Kuan Hon’s infographics on EU data protection legislative process at http://blog.kuan0.com/2015/01/data-protection-directive-vs-draft.html. Visited on April 2016.

[57] See for the main differences: Gutwirth, S.; Leenes, R.; de Hert, P. (Eds.) Reforming European Data Protection Law, Springer, Dordrecht (2015).

[58] “[. . .] an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that person.”

[59] See Linked Data Awareness Barometer 2015, http://www.meetup.com/Vienna-Semantic-Web-Meetup/messages/76529763/. Visited on April 2016.

[60] A person is considered *identifiable* if a piece of information contains elements of identification through which the person can be identified, directly or indirectly. As depicted in Recital 26 of the DP Directive, the standard is whether it is likely that reasonable means for identification (e.g. deanonymisation) will be available and administered by the foreseeable users of the information (either by the controller or by any other person to identify the said person, which includes third-party recipients). Recursively, if *anyone* can identify someone from the data, it is personal data.

[61] European Commission’s press release announcing the proposed comprehensive reform of data protection rules, http://europa.eu/rapid/press-release_IP-12-46_en.htm?locale=en.

[62] Schmachtenberg, M., Bizer, C., Paulheim, H., Adoption of the Linked Data Best Practices in Different Topical Domains. The Semantic Web – ISWC 2014 vol. 8796 LNCS pp 245–260 (2014).

[63] The resource http://www.w3.org/People/Berners-Lee/card.rdf is online since 2000.

either legal entities such as companies, Government departments or voluntary organisations, or they can be individuals. *Data processors* (Art. 2 (e)) are a legally separate entity that processes personal data on behalf of a controller.

The definition of "processing of personal data" (contained in Art. 2 (c) of Convention 108; and Art. 2 (b) and Art. 3 (1) of the DP Directive,) covers "*any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction*". In practice, this definition covers any possible usage of linked data.

Personal data are processed under the liability of the data controller, who must implement all measures, both technical and organisational, to ensure that no harm is caused to the data subject by the processing. It is intricate to define the *users* of personal data in the complex environment of linked data, where many scenarios can be foreseen involving controllers and processors, alone or jointly, with different degrees of autonomy and responsibility. This role qualification may render legal responsibility[64] for complying with the respective obligations under data protection law.

3.2.2. *Obligations about personal linked data*

Four major categories of obligations are required on those processing personal linked data:

(i) *lawfully process the personal data* (Art. 6 (1) (a) and (b) of the DP Directive; Art. 5 (a) and (b) of Convention 108). Those managing a personal linked dataset will have to demonstrate a legitimate legal basis for processing the personal data file. Storing and processing a personal linked data is possible if certain conditions hold.

(ii) gather explicit consent[65] from the data subjects. A valid consent shall fulfil these requirements: *free*, *informed* and *concrete*. Open consent, namely, unconditionally disclosing personal information,[66] cannot be sought within linked data. We may ask if the risks of hampering data protection are the same. If the risks are different, then perhaps there may exist grounds for a reconsideration of consent requirements and for a definition of open consent, which would stray further from the traditional model of specific informed consent (e.g. for research uses 'in general').
Other conditions to store and process personal linked data exist. For example, if there is a *vital interest* of the data subject, if it is done in compliance with a *legal obligation* by the controller, if it is executed as a necessary step for the performance of a *contract*, and finally if there is a *legitimate interest* pursued by the controller.

(iii) *to have a legitimate purpose*: personal data must be collected for specified, explicit and legitimate purposes that are established by the data controllers, and not for other purposes which exceed the stated ones (according to the principle of purpose limitation, Art. 6 (b)). The compatibility with these purposes can be assessed in relation to the specific and definable purpose: publishing and consuming linked data in a legally compliant way. However, if personal data no longer serve their initial purpose, but are to be kept in a personalised form for the purpose of historical, statistical or scientific use, the DP Directive and Convention 108 allows this possibility on condition that appropriate safeguards against misuse are applied (Art. 6 (1) (e); and Convention 108, Art. 5 (e)).

(iv) *to notify the supervisory authority* (NDA) "before carrying out any wholly or partly automatic processing operation". Such notification must contain specific details including, between others, "the purpose or purposes of the processing" (Art. 18 and 19 (1) (b) of DP Directive).

(v) *to respect data subject's rights*: the right of access to data, the right of rectification, erasure or blocking data when their processing does not comply with the provisions of the DP Directive (Art. 12), and the right to object to personal data processing (Art. 14).

The *principles* pursuant to the DP Directive when processing personal data applied to linked data are subject of our analysis.

(i) the *principle of proportionality* imposes that personal data processing "*must be adequate, relevant and not excessive for the purposes for which they are collected and/or further processed*" (Art. 6 (c) DP Directive). Nevertheless, we may question how to respect this principle when reuse of personal data is allowed. We further contend how to respect the data minimisation principle with linked open data.

(ii) the *principles of purpose specification and use limitation*[67] (Art. 6 (1) (b) DP Directive; Art. 5 (b) Convention 108,) impose that the purposes for which personal data are collected should be specified not later than at the time of data collection, and the subsequent use shall be limited to the fulfilment of those purposes.

(iii) the *principle of accountability* (DP Directive, Art. 6 (2)) postulates the need of recording and making accessible the events related to the personal data. Semantic technologies particularly favour keeping a transparent account of the data transactions and chains of provenance are naturally represented as RDF graphs.

(iv) the *data quality principles, such as relevancy* (Art. 6 (1) (c) DP Directive *and accuracy* of data (Art. 6 (1) (d) DP Directive). The limited retention of data (Art. 6 (1) (e) DP Directive; Art. 5 (e) of the Convention 108) warrants particular thought. It posits that data should be kept up to date when necessary, within a "time of conservation" (re-

[64] Art. 29 Working Party (2010), Opinion 1/2010 on the concepts of "controller" and "processor", WP 169, Brussels, 16 February 2010.

[65] Concerning the validity of individual consent (Art. 2 (h) and Art. 7 of the Directive); see Opinion 15/2011 on the definition of consent, Art. 29 Data Protection Working Party.

[66] Lunshof, J.E., Chadwick, R., Vorhaus D., and Church, G. (2008). From genetic privacy to open consent. Nature Reviews Genetic.

[67] An influential set of OECD guidelines (non-binding, but which have been incorporated into a number of binding statutes and conventions over the years) published as long ago as 1980 (OECD 1980) set out the Purpose Specification Principle and the Use Limitation Principle.

tention period) that permits identification of data subjects for no longer than it is necessary for the collected purposes. Consequently, data would have to be anonymised if a controller wanted to store them after they were outdated and no longer served their initial purpose. Databases holders could also include a technical system that would help anonymising personal data after the storing time of their first processing in order to automatically allow reuse of these data after this anonymisation.

It is observable that these lockdown principles, especially the Purpose Specification and Use Limitation Principles, do not concord evenly with the serendipitous reuse of linked open data, which comprises obviously the Use Maximization Principle.[68] If data constitutes personal data, then the Use Limitation Principle may prevail towards its use maximisation. Nevertheless, when it is unclear whether an RDF dataset contains personal data or not, we may wonder if transparency demands of open data is dismissed and bows to data protection requirements.

Finally, processing for scientific research and statistics is allowed if the Member State provides appropriate safeguards (Art. 13 of the DP Directive).

3.2.3. *Risks*

As in the FOAF profile of Tim Berners-Lee example, many accept that information is published again (see the Creative Commons licence in the RDF excerpt). This is a lawful situation, though in practice it poses a practical problem: the information can be used for *spamming* or other bad purposes.[69]

Alleged concerns rely on the fact that integrating data from distinct sources of available linked data encompassing personal information, even with apparently innocuous or anonymised linked resources, may enhance a jigsaw of indirect *identification* and *re-identification*.[70] Envisaging an *RDF dataset* containing personal information, interlinked with another without it, this latter resource may inherit data protection compliance requirements; this scenario could escalate if there is access to rich information resources via the web. Thereby, identifiable personal information set through re-identification intrinsically abides to legal requirements. If this is the case, we may question the role of linked open data in the realm of current data protection settings.

Another commented risk fuelled when processing Linked Data is the *profiling* of individuals.[71] The use of big data analytics, machine learning, natural language processing and data mining techniques may enhance the integration of personal data to create and use profiles: consumer, movement, user and social profiles. Profiling vests companies, public authorities to determine, analyse or predict people's personality, behaviour, interests and habits without their cognition. Such processes may and are likely to epitomise privacy invasiveness or even waiving the data subjects' control upon their data.

When data are *anonymised* (Art. 2(a) and recital 26 of the Directive) all identifying elements have been eliminated from a set of personal data and cannot leave space to re-identify the person(s) concerned. Where data have been successfully anonymised, it is deemed to be no longer personal data and data publishers may be able to release, sell or publish the data without data protection requirements.

De-anonymisation strategy in data mining entails that anonymous data is cross-referenced with other sources to re-identify the anonymous data. Thus, the processing of datasets rendered anonymous may never be ensured.

When personal information is *pseudonymised*, the identifiers (name, date of birth, etc.) are replaced by one pseudonym. Pseudonymisation is achieved, for instance, by encryption of the identifiers in personal data. The Explanatory Report to Convention 108 states, in its Art. 42, that "[t]*he requirement [. . .] concerning the time-limits for the storage of data in their name-linked form does not mean that data should after some time be irrevocably separated from the name of the person to whom they relate, but only that it should not be possible to link readily the data and the identifiers*". This is an effect which can be achieved by pseudonymising data. For whom not possessing the decryption key, pseudonymised data can only be identifiable with difficulty. For those who are entitled to use the decryption key re-identification is easily possible. Therefore, the use of encryption keys by unauthorised persons must be particularly guarded against.[72]

We may argue if the responsible candidate of a breach would ultimately be the publisher of the final piece of the dataset that enabled the identification to take place, which configures an onerous duty on anyone releasing anonymised data into the public domain, without powers to control access (duty to ensure that no-one can be identified from the data),[73] especially when it is not possible to predict all the circumstances in which published data will be valuable, used or reused in any way possible by unaccountable data users.

In this line of reasoning, recognition towards empirical evidence of the presented assumptions are necessary to confirm if these mere theoretical risks do translate into a real-world threat within linked data, but in any way are negligible risks, as computational power is growing. Moreover, it could be a constraint if utility, from the monetisable value of data, and transparency concerns from the Open Data strategy have to

[68] O'Hara, K. (2011) Transparent government, not transparent citizens: a report on privacy and transparency for the Cabinet Office, 9.

[69] Nasirifard, P., Hausenblas, M., Decker, S. (2009). Privacy Concerns of FOAF-Based Linked Data. Proc. of the Trust and Privacy on the Social and Semantic Web at ESWC.

[70] This risk is described as *re-identification* from the aggregation of anonymised datasets. See further the Art. 29 Working Party's Opinions: "Opinion 6/2003 on the re-use of public sector information (WP29 2003)", "Opinion 3/2013 on purpose limitation (WP29 2013a)" and the "Opinion 6/2013 on open data and public sector information (PSI) reuse" (WP29 2013b). The Art. 29 Working Party on the Protection of Individuals with regard to the Processing of Personal Data is an independent advisory body on data protection and privacy. It is composed of representatives from the national data protection authorities of the EU Member States, the European Data Protection Supervisor and the European Commission.

[71] Weber, R.H., The digital future – A challenge for privacy? Computer Law & Security Review, Vol. 31(2) Apr. 2015, pp. 234–242 (2015).

[72] EU Agency for Fundamental Rights and Council of Europe (2014). Handbook on European data protection law, Publications Office of the European Union.

[73] See footnote 61.

be curtailed to preserve data protection due to the threat of identification of data subjects.[74]

3.3. *Other possible legal frameworks: competition law and trade secrecy law and the licensing and contractual cases*

If the market of linked data keeps on developing, it will constitute a sector where conflicts will undoubtedly appear and competition law may apply. In particular, it can be foreseen the competition among the linked data providers, among the data publishing platforms or among semantic service providers. In any of these cases, the market players will have to resort on the rules of competition law to prevent damage to their business models on which they have spent considerable time and investment.

The locus for regulatory intervention at EU level is the Treaty on the Functioning of the European Union (TFEU), further refined by the Damages Directive.[75] In particular, the TFEU Art. 101 outlaws the *non-exempt agreements and concerted practices that restrict competition and affect trade between EU member states* and TFEU Art. 101 *precludes an organization that hold a dominant position on a relevant market from abusing that position*. There is a certain risk that these situations will happen, but the distributed nature of linked data makes in principle the risk to be controlled.

Trade secrecy laws may be relevant for this type of data, as linked data may convey trade secrets, confidential documents or information whose disclosure may compromise national security. In November 2013, the EU Commission proposed a draft directive[76] that will align existing laws against the misappropriation of trade secrets across the EU. The draft Directive deals with unlawful conduct by which someone acquires or discloses, without authorisation and through illicit means, information with commercial value that companies treat as confidential in order to keep a competitive advantage over their competitors. Trade secrecy is acknowledged in Art. 39 in the TRIPS agreements. In general, the requirements for data to be legally considered as confidential information include: (1) the data not being generally known to the public; (2) the protected content must be secret not available to the public; (3) the data conferring some kind of economic benefit to its holder; and (4) the data being the result of a reasonable effort.

In any case, a requirement for linked data to be a trade secrecy implies its non-publication, a predicate contrary to the linked data nature. Thus, an additional prohibition may exist on linked data not published in the web: *disclosure* (as the communication of the information to third parties, only by obtainment of the consent of the trade secret holder will allow disclosure).

The exclusive copyrights are capable of assignment (legal transfer) and hence licenced by restricted acts, by time, by geography, etc., to a variety of third parties, either in whole or in part, under particular terms and conditions that stipulate precisely the manner and the extent to which the work can be legitimately exploited (recital 30 of the Copyright Directive), which copyrights are waived and under which circumstances.

Linked data publishers may limit their responsibility by publishing a *disclaimer* where no warranties are given on the data quality (reliance on the data's accuracy in their terms of use).

Creative Commons[77] initiative proposes a system of alternative licences for authors willing to waive the rights granted by default under the law. Nevertheless, legal uncertainty concerns can be foreseen regarding open content licences. The same licence may vest a legal status in different jurisdictions.[78] Any provision which may extend beyond the scope of the copyright regime cannot be enforced under copyright law, but only on the basis of contract law, hence, "the extent to which the provisions of a license can be enforced against third parties will ultimately depend upon whether or not the license can be regarded as an actual contractual agreement – as opposed to a license".[79]

Transactions of linked data in contractual relations are starting to appear. The *linked data market* has been defined as "*a specific type of marketplace, which is built on top of the Linked Data Web holding on to the linked data principles*"[80] and is starting to grow.[81] Transactions eventually required for contract negotiations may follow these major steps: (i) *Data transaction*. A data transaction is initiated when a data consumer issues a request to the data market, which is subsequently forwarded to one or more data providers who can potentially service the request; (ii) *Compose and offer contract*. The data provider generates a machine readable contract (known as an offer). The autogenerated contract is subsequently offered to the data consumer and (iii) *Accept contract*.

If the data consumer agrees to the terms of the contract, an agreement between the data consumer and the data publisher is generated and persisted for accountability and compliance purposes. These data markets can be regulated by digital rights expressions (or policies) enabling more complex scenarios including automated negotiation and consump-

[74] Cavoukian, A., El Emam, K. (2011) Dispelling the Myths Surrounding Deidentification: Anonymization Remains a Strong Tool for Protecting Privacy, Ontario: Office of the Privacy and Information Commissioner.

[75] http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=uriserv:OJ.L_.2014.349.01.0001.01.ENG.

[76] http://ec.europa.eu/growth/industry/intellectual-property/trade-secrets/index_en.htm.

[77] See http://creativecommons.org/ (visited April 2016).

[78] De Filippi, P.: Copyright Law in the Digital Environment: Private Ordering and the regulation of digital works. LAP LAMBERT Academic Publishing GmbH & Co. KG, pp.116. (2012).

[79] Ibid.

[80] Stahl, F., Schomm, F., Vossen, G. The Data Marketplace Survey Revisited. Technical report, Working Papers, ERCIS-European Research Center for Information Systems (2014).

[81] The European Research Center for Information Systems (ERCIS), registered a slight decrease in the number of service providers offering access to "raw data" and an increase in the provision of "high quality processed data". The latter means data that is represented in a manner which supports data integration and analytics, i.e. data represented in manner which is interoperable, flexible and extensible, characteristics which conform as cornerstones of the RDF data model.

Table 2 – Legal instruments applied to typical linked data resources.

	URI	RDF statement	RDF dataset and mappings	Ontology
Copyright Directive	No	No	Seldom	Often
Database Right Directive	No	No	Often	No
Data Protection Directive	No	No	Yes	No
Trade Secrecy Law	No	Yes	Yes	Yes
Computer Program Directive	No	No	No	Seldom
Patent Law	No	No	No	No

tion by machines.[82] The role of electronic contracts within this market is and is likely to draw interest in the next years. The notion of *Meta-rule of law* has been proposed to set the democratic conditions both of policies through digital or rights expression languages, and data (and meta-data) markets.[83]

3.4. *Summary*

As analysed in this Section 3, in order to grant the lawfulness in using or publishing a linked data resource, the nature of the resource must be determined. As a rule-of-thumb, Table 2 depicts the applicable law in the most common cases, assuming that the following assumptions hold: (i) linking data, i.e. publishing facts about other's *URIs*, is the central activity of the whole linked data paradigm; (ii) individual *RDF statements* have no IP protection; (iii) *datasets* and *mappings* are protected by database law (with the exception when triples are inferred); (iv) *ontologies* are regarded as copyright works and only occasionally as computer programs.

In order to ease the lawful consumption of linked data, resources should be accompanied by a description of their legal status (e.g. whether it has copyright, contains personal data, etc.), as postulated by Rodríguez et al.[84]

4. Conclusions

4.1. *Review of legal aspects of linked data*

We have assessed that some cautions are necessary for consuming and publishing linked data in a lawful manner, and hence the existing legal framework has to be taken into account.

As any specific legal characterisation of linked data resources has not yet been construed, the assertions of this paper stemmed from a legal analysis of the European legal framework, resorting to official sources and further discussion with experts on linked data. It was not intended as a thorough analysis of copyright or data protection domains, but merely to the extent that its substantive normative provisions could be subsumed into each of the plausible qualifications of the linked data resources (linked data itself and related ontologies and services). The potential threats of re-identification of individuals in anonymised datasets are boosted if these datasets are linked data – EU citizens, already concerned with security and privacy,[85] may not realise the power of these new technical developments.

It is our contention that rights and duties in relation to linked data resources configure a complex framework in a cross-border setting. Several, fragmented and differing legal instruments may be resorted among a mix of international, European and national rules: some resources can be protected by copyright and database laws, and their access may be limited by data protection, trade secrecy law, among others – each discipline conveying their own technical rules (in scope, enforcement and extent) within any single country.

Although it was meant to harmonise and adapt copyright to the digital age, EU copyright rules are considered to be maladapted, outdated and fragmented towards the increase of cross-border information interconnectedness and exchange facilitated by the Internet, and utters a burden on online activities as to whether they are compliant with the law.[86]

Differences remain and the geographical scope of rights is limited to the territory of the Member State granting them. Any activity could be legal in one country but illegal in another due to the Member States discretion concerning exceptions and limitations, causing cross-border friction and legal uncertainty. Moreover, these legal instruments ("national silos") may be applicable in a concurrently and overlapping way to the same data source. Rights and duties also may apply to the same source of data (dataset) in a stratified way. For example, the dissemination of copyright-protected content on the Internet – e.g. uploading an existing ontology, requires, in principle, an authorisation for each national territory in which the content is communicated to the public.

Copyright law attempted to replicate the rules of the physical world into the digital world, by applying inadequate patterns to an entirely new context. In consequence, the traditional stability and the territoriality of copyright regime disrupted in the

[82] Steyskal, S., Kirrane, S. (2015) If You Can't Enforce It, Contract It: Enforceability in Policy-Driven (Linked) Data Markets. in: Joint Proc. of the Posters and Demos Track of 11th Int. Conf. on Semantic Systems – SEMANTiCS 2015, A. Filipowska et al. (eds.), CEUR Workshop Proceedings, Vol. 1481, pp. 63–66.

[83] Casanovas, P. (2015). Conceptualisation of Rights and Meta-rule of Law for the Web of Data, Democracia Digital e Governo Eletrônico, vol.12, pp. 18–41; repr. Journal of Governance and Regulation, vol. 4, num. 4, pp. 118–129.

[84] V. Rodríguez-Doncel, A. Gómez-Pérez, N. Mihindukulasooriya. (2013) Rights declaration in Linked Data, in Proc. of the 3rd Int. W. on Consuming Linked Data. O. Hartig et al. (Eds) CEUR vol. 1034.

[85] In a recent Eurobarometer on Data Protection six out of ten respondents said that they did not trust online businesses (63%), or phone companies and internet service providers (62%). See the Data Protection Eurobarometer of June 2014.

[86] Dobusch, L. and Quack, S. (2012). Transnational Copyright: Misalignments between Regulation, Business Models and User Practice. Osgoode CLPE Research Paper No. 13/2012.

digital environment. In the instance of the knowledge-based society, and facing the Digital Single Market vision, copyright rules need to be framed within a technology-neutral and future-proof way.

4.2. *Outlook of future developments*

In a prospective thinking, further measures (legislative or non-legislative, including market-led solutions) need to be taken at EU level in the medium and long term ensuring an adequate level of protection for right holders and in the data protection domain alike.

The appointed limitations have been already acknowledged by the European institutions, and the need for a reform was advanced in the Digital Agenda[87] of the European Commission, one of the seven pillars of the Europe 2020 Strategy, and within the Intellectual Property Strategy[88] in relation to the Digital Single Market.[89] A "*Public Consultation on the review of the EU copyright rules*" took place in 2014 with important conclusions.[90] Additionally, other market-led developments occurred, like the work in the Linked Content Coalition.[91]

The required balance between exceptions and limitations in copyright could also be designed in the digital environment without any unequal treatment compared to those granted in the analogue world. Therefore, future compatibility of exceptions and limitations by taking due account of the effects of media convergence should also be considered. According to the European Parliament's proposal,[92] there is a call for the adoption of an open norm, introducing flexibility in the interpretation of exceptions and limitations in certain special cases that do not conflict with the normal exploitation of the work and do not unreasonably prejudice the legitimate interests of the author or right holder.

As already stated, European data protection law has been under review for a long time.[93] This review eventually resulted in the recently approved General Data Protection Regulation (April 14th, 2016).[94] The Regulation will enter into force twenty days after the date of publication in the Official Journal of the European Union (OJEU), and will be fully applied two years after this date. It repeals Directive 95/46/EC, providing a new background to the relationship between linked data and Data Protection. Hopefully, the GDPR will put an end to the "privacy patchwork" of the 28 EU Member States, as national Courts must apply its content without further ratification by national Parliaments. Yet, the construction of a Single Digital Market, protections enacted by DGDR, and provisions of the EU Agenda against terrorism and organised crime – especially in the aftermath of December 2015 and the recent attacks in Brussels – should be able to find a common ground in the immediate future. On March 11th, 2016, the Council agreed its negotiating position with EU Parliament on the proposal for a Directive on combatting terrorism.[95] Very likely a proactive position (rather than merely reactive) will be held at the institutional level. It shall be harmonised with the basic positions and developments of GDPR.

Four main innovations will be put in place: (i) the new Regulation establishes a single, pan-European law for data protection; (ii) it sets a single regulatory framework for businesses as well, as companies – micro, small, medium-sized or corporations – will be treated differently but have to deal with one only Supervisory Authority (SA); (iii) Data Protection SA will be able to fine companies who do not comply with EU rules with up maximum fines of €20 million or 4% of a business's total worldwide annual turnover; (iv) thus, service providers and other personal data processors on behalf of other businesses will be liable for data breaches, as they will have some direct obligations, such as taking adequate measures to protect personal data transfers. Therefore, along with the upcoming Directive on personal data processing in criminal matters, GDPR shapes a new general framework for the protection and exercise of rights. This foundational public character cannot be ignored. The text eventually approved consists of 99 articles, with 179 Recitals offering open statements and detailed explanations about the character, scope and nature of its implementation and intended effects.

The new Regulation pays a special attention to the rights of access, rectification and erasure of personal information and it expands protections and rights to security issues.[96] An adequate level of data protection rights should be kept for the processing and exchange of information in criminal matters, as stated even by the controversial (and confidential) "umbrella agreement" between US and EU on data

[87] Digital Agenda for Europe, COM(2010) 245 final/2.

[88] A Single Market for Intellectual Property Rights, COM (2011) 287 final.

[89] Communication on content in Digital Single Market, COM (2012) 789 final.

[90] http://ec.europa.eu/internal_market/consultations/2013/copyright-rules/index_en.htm (visited on April 2016).

[91] http://www.linkedcontentcoalition.org/ (visited on April 2016).

[92] Draft Report on the implementation of Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society (2014/2256(INI)), 14/01/2015.

[93] Cf. S. Gutwirth, R. Leenes, P. De Hert (Eds.). Data Protection on the Move. Current Developments in ICT and Privacy/Data Protection. LGT Series vol. 25, Springer Verlag, Dordrecht (2016).

[94] Cfr. Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) COM/2012/011 final – 2012/0011 (COD). After nearly 4.000 amendments, the final draft of April 6th was approved by the EU Parliament on April 14th, 2014. See http://data.consilium.europa.eu/doc/document/ST-5419-2016-INIT/en/pdf.

[95] http://www.consilium.europa.eu/en/press/press-releases/2016/03/11-directive-on-combatting-terrorism/.

[96] Recital 19 states that GDPR does not apply to "the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, including the safeguarding against and the prevention of threats to public security and the free movement of such data". It refers straight to the content of the next Directive on the protection of individuals, still pending, specifying that Member States may entrust competent authorities within the meaning of Directive (EU) 2016/. . . with tasks which are not necessarily carried out for these purposes.

transfers (October 9th, 2015).[97] Other protective firewalls have been built, after the EU Court of Justice ruling against Commission's US Safe Harbour Decision declaring secure enough the transfer of personal data from EU to USA. Thus, there is a complex political issue behind the tension between security and data protection principles.[98] It cannot be confused with a technical legal issue.[99]

GDPR will be joined soon by the new *Directive on the processing of personal data by competent authorities for the purposes of prevention, investigation, detection or prosecution of criminal offences.*[100] The draft is now at the stage of second reading by the EU Parliament, and it intends to go deeper in the same direction. Thus, the same matrix of protecting principles and restrictions will apply to prevent mass surveillance, indiscriminate recording, and unlimited retention.[101] Perhaps linked data will have to be explicitly annotated whenever it contains personally identifiable information so that erasure processes can be automated. Further, encrypting technologies for linked data will probably need additional development.

What else will it mean for linked data? By using the linked data fabric, divergent interests and fundamental rights emerge: economic growth, freedom of information, access and openness, reuse arguments vs. privacy and data protection, confidentiality, transparency, accountability arguments and values of transparency. We still query how principles and provisions of data protection and privacy are compatible with publishing or consuming linked data, and with linked data use for security purposes. It is clear that the law will continue to respond, through legislative and judicial developments, to questions raised by new practices, such as the publication of linked data by public or private entities.

Challenges are more difficult to outline in the future to come. Consumers of information in the future web of data will be naturally machines rather than humans. These machines will eventually have purchasing power, with the intelligence and autonomy to decide and operate machine to machine data transactions, posing thereby new ethical and legal challenges.

Acknowledgements

This research is supported by the Ministry of Economy and Competitiveness (project TIN2013-46238-C4-2-R), DER2012-39492-C02-01 CROWDSOURCING, RTC-2014-2946-7 Crowds eAssessment, RTC-2014-2666-7 Info_COR, and a 520250-1-2011-1IT-ERASMUNDUS-EMJD grant.

[97] *"The United States must [. . .] guarantee that all EU citizens have the right to enforce data protection rights in U.S. courts, whether or not they reside on U.S. soil. Removing such discrimination will be essential for restoring trust in transatlantic relations"* Agreement between the United States of America and the European Union on the Protection of Personal Information relating to the Prevention, Investigation, Detection, and Prosecution of Criminal Offenses, http://europa.eu/rapid/press-release_IP-15-5812_en.htm. For criticisms, see http://free-group.eu/2015/10/14/eu-us-umbrella-data-protection-agreement-detailed-analysis-by-douwe-korff/.

[98] European Court of Justice. Judgment in Case C-362/14 *Maximillian Schrems v Data Protection Commissioner*, October 6th, 2015. The Irish High Court of Justice, first, and the EU Court of Justice later on, overruled the Decision of 26 July 20002, in which the Commission considered that under the "safe harbour" scheme the United States ensured an adequate level of protection of personal data. See *Court of Justice of the European Union, Press release n. 117/15*, Luxembourg, 6 October 2015. Cf. the Judgement of the Court (Grand Chamber).

[99] Cfr. the recent Article 29 Data Protection Working Party, *Opinion 01/2016 on the EU – U.S. Privacy Shield draft adequacy decision, adopted on 13 April 2016*: "The check and controls of the adequacy requirements must be strictly performed, taking into account the fundamental rights to privacy and data protection and the number of individuals potentially affected by transfers. The Privacy Shield needs to be viewed in the current international context, such as the emergence of big data and the growing security needs".

[100] *Directive on the protection of individuals with regard to the processing of personal data by competent authorities for the purposes of prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and the free movement of such data and repealing Council Framework Decision 2008/977/JHA (05418/1/2016 – C8-0139/2016 – 2012/0010(COD))*{SEC(2012) 72 final}. See the text of the draft adopted on March 14th, 2014 at the first reading at http://www.europarl.europa.eu/sides/getDoc.do?type=TA&language=EN&reference=P8-TA-2016-0126.

[101] Cfr. Marju Lauristin, Rapporteur, quoted on this point: "What are some of the main innovations for the police? We do not currently have a common framework for data protection or data processing in 28 Member States. We had a general directive that we replaced by regulation, but in the police field we did not have anything like that. This is the first regulatory act which will provide police in all Member States with their common rules. In this framework, *the very important thing is that the general principles of proportionality, legitimacy and purpose-limitation are included in police work. That means that no form of mass surveillance is possible. The collection of data is not possible. Retention for an unlimited or unclear period is not possible. Another important point is that we foresee the inclusion of data protection professionals in the police institutional setting: specifically, in police work* [Our emphasis]." Debate on the protection of individuals with regard to the processing of personal data for the purposes of crime prevention, Strasbourg, Wednesday, April 13th, 2016.