Legal Interoperability Issues in the Framework of the OpenMinTeD Project: A Methodological Overview

Penny Labropoulou[1], Stelios Piperidis[1], Thomas Margoni[2]

[1]Institute for Language and Speech Processing/Athena Research Center
Athens, Greece
[2]Law School, University of Stirling
Stirling, United Kingdom
Email: penny@ilsp.gr, spip@ilsp.gr, thomas.margoni@stir.ac.uk

Abstract

This paper is a first analysis of the legal interoperability issues in the framework of the OpenMinTeD (OMTD) project (www.openminted.eu), which aims to create an open, service-oriented e-Infrastructure for Text and Data Mining (TDM) of scientific and scholarly content. The paper offers an overview into the methods for achieving such interoperability.

Keywords: Text and Data Mining, legal interoperability, methodology

1. Introduction

This paper is a first analysis of the legal interoperability issues in the framework of the OpenMinTeD (OMTD) project (www.openminted.eu) which aims to create an open, service-oriented e-Infrastructure for Text and Data Mining (TDM) of scientific and scholarly content. The paper discusses methods and tools for achieving such interoperability at the theoretical and practical levels.

In the following section we present our working material, i.e. the resources involved in TDM as envisaged in the project, and their legal status quo. Then, we take a closer look into the legal framework of TDM and Language Resources, focusing mainly on licensing issues (Section 3). Section 4 deals with the representation of legal elements in metadata descriptions for e-distribution and e-infrastructures. Section 5 discusses issues of interoperability. Finally, we conclude with considerations on future perspectives.

2. Types of Assets in OpenMinTeD

The elements involved in the TDM and relevant Language Processing processes in the framework of the project are distinguished into:

(a) **Content**, covering:

- *the textual content* that can be mined, such as documents, web pages, text corpora, or data input by the user; for the purposes of OMTD, we will be focusing on scientific and scholarly publications. This type of content is often protected by copyright, usually as a literary work, and depending on the circumstances by the Sui Generis Database Right (SGDR).
- *language/knowledge resources*, such as computational lexica, terminological resources, ontologies, authority lists and other reference vocabularies, language models, computational grammars, etc., that are used as reference and/or ancillary resources in the creation and/or operation of processing software. For instance, an OpenNLP powered web service is parameterisable as to the model it uses; or a term annotation service that looks up terms in different ontologies, is combined with the specific knowledge resources that address these tasks. This type of content is likely protected by the Sui Generis Database Right (SGDR) as far as it constitutes a non original database, but copyright may still be relevant both in relation to the structure or selection of the database and to the nature of the collected work.

(b) **Software**, which is usually made available as a downloadable tool, usually in executable form. Software as such is protected by copyright as a literary work. Other forms of protection that may be relevant in the case of software (e.g. patents) are not covered by this study.

(c) **Services**, mainly in the form of:

- *web services*.
- *workflows (pipelines of web services*.

Web services and workflows perform the desired task. The use of services is often regulated by specific Terms of Use or Terms of Service (ToS).

(d) **Derived assets**: Ultimately, of course, there is the final output of the process, which is the mined data or information. The processed data between components of the web service (or of web services, in the case of a workflow) are likewise by-products of the TDM process, and they are also potentially protectable as original or derivative works (or other subject matter) and consequently licensable.

To make things more complex, the web service (or workflow) can be made up of a mixture of software components (or services) and the input data can also be the aggregation of two or more datasets.

Users of the OMTD infrastructure who want to run a web service on a specific dataset, thus, have to check the entire set of the licences of these resources in order to be sure that the output they obtain at the end is legally consistent. If they wish to distribute this output in some form, they must also ensure that the licensing terms they will impose on the output do not violate any of the licensing terms of the ingredients of this process.

3. Overview of the Legal Framework

3.1. Copyright and Licences

Copyright and the Sui Generis Database Right (SGDR) are the most relevant rights for TDM purposes (De Wolf & Partners, 2014; Guibault & Wiebe, 2013). Other rights or regulations such as personal data protection and Public Sector Information (PSI) may also play a role, sometimes an important one. (Keller et al, 2014). However, generally speaking these forms of legal regulation cannot be managed through a licensing approach, and will therefore be addressed only to the extent that they are relevant in relation to the interoperability considerations covered in this paper.

In accordance to the above, it is at the level of copyright licences for content and software and to the Terms of Use employed for services that we need to direct our analysis. It is important to bear in mind that the legal framework on which copyright licences rest is not always clear and coherent, but rather a complex mixture of broad rights and unharmonised exceptions. This situation often stifles the scientific activity of researchers instead of promoting it, thereby reinforcing even further the need of a clear and interoperable set of licences.

When a publication or a language resource meets the usually not very high thresholds for protection (of either originality or substantial investment), it will automatically be under an "all rights reserved" legal status (Guibault & Wiebe, 2013), i.e. the default legal framework is that these resources cannot be used unless a specific authorisation accompanies them. This specific authorisation is called a (copyright) licence.

This shows how crucial it is to properly license content and tools, because by omitting a rights statement, or by stating something approximative or wrong, the legal result is that the resource, content or software, cannot be rightfully used or reused.

It is conceptually important at this stage to note that there are exceptions to this "all rights reserved" rule. They are called "exceptions and limitations to copyright" in continental European countries and "fair dealing or fair use" in countries belonging to the common law tradition (UK, Ireland, USA, Australia, etc.). However, as explained in the relevant literature, especially for the European situation, the available exceptions are not a satisfactory solution (De Wolf, 2014; Guibault & Margoni, 2015).

Accordingly, for present purposes, the default legal status of resources is "all rights reserved" which makes it necessary to verify under which conditions the use and further distribution of the original and of the mined content is permitted.

These conditions are usually contained in licences or other documents intended to regulate the use of specific content, tools or services, also known as copyright licences, public licences, terms of use, acceptable user policies, service level agreements, etc. Unfortunately, in many instances the legal documents that regulate the use and reuse of publications, software and other resources appear as lengthy and complex ad hoc (i.e. not standardised) legal agreements that the researchers are not prepared or trained to understand. This is not only a question of possessing the proper legal skills, but also a matter of transaction costs: even in those situations where a specifically trained lawyer is available, the number of legal documents to be analysed and the lack of standardisation in the documents, clauses and conditions sharply contrast with the scientific and academic needs of clear, efficient and interoperable rules on use and reuse of sources.

An example can illustrate the situation. Even if some resources are stated to be in "Open Access", this term – although having received a rather clear definition – is nonetheless loosely employed in a variety of forms that not only may imply different legal requirements but even be in contrast with each other. More importantly, Open Access is a (fundamental) statement of principles that has to be properly translated into appropriate legal documents (licences): Merely stating that a resource is in Open Access only adds confusion and uncertainty in a field which is in deep need of the opposite. In other words, due to the inconsistent and inappropriate use of the term, it is often not possible to combine two resources released under the same "Open Access" label, regardless of the intention of the right holders. While it is clear that the reason for such an inefficient situation does not rest with the concept of Open Access itself but rather with an incorrect use of the term, the resulting situation is one where use and reuse of information is made more difficult instead of facilitated.

From an interoperability point of view, it is important to consider what happens when several resources with different licences are required to interact. Each licence may have different requirements and conditions regulating the use of the resulting content. A lack of licence standardisation and interoperability is a clear stumbling point to researchers who wish to adopt TDM in their research. Both deeper and clearer interoperability rules between these licences are essential for the swift adoption of TDM within and outside professional communities.

3.2. Types of Licences and the Socio-legal Framework

The creation, use and distribution of Language/Knowledge resources is rooted in the Corpus Linguistics tradition, which was at the very beginning mainly research oriented and driven by individuals and organisations that had the dual role of resource creator and resource consumer. Thus, licensing was not so important at first; consequently, a lot of these resources have been and may still be licensed with loose unofficial agreements on a case-by-case basis, or general statements such as "for research only". It is only more recently, with the increasing request for data consumption by other users besides their creators and the realisation that data brokerage can be a profitable business, that licensing has attracted attention. This also brought to an increasing use of more standardised licences through institutional sites, dedicated agencies (e.g. ELRA www.elra.info, LDC www.ldc.unipenn.edu) and infrastructures (e.g. META-SHARE www.meta-share.org, CLARIN www.clarin.eu). In this ecosystem, we find mixed together open licences (e.g. CC, META-SHARE), licences with terms for specific communities, various proprietary licences and terms of use with similar licensing conditions but still not standardized, free text statements/legal notices (e.g. for research use, open access) etc.

Software licences, on the other hand, are more standardised. Next to the proprietary licences of companies for specific market products, Free and Open Source Software licences (FOSS) are extensively used for software mainly in the form of downloadable and installable versions. As a matter of fact, FOSS licences are used even for data resources, which shows how much data providers are unfamiliar with legal notions.

As for web services and workflows, we witness the use of FOSS but also, in increasing amounts, terms of services usually with specific restrictions (e.g. time of processing or size of content to be processed).

3.3. The Importance of a Licence Multi-layer Approach

In the field of TDM it is important to properly address the licence compatibility issue by employing a "multi-layer licence approach". The starting point is of course to focus on just one "layer", e.g. content licences or software licences or terms of use, and try to resolve compatibility issues "within" the same type of licences. This means to verify the compatibility of the same kind of licences in order to determine whether two or more content licences can be combined, or two or more software licences can be combined.

A multi-layer approach applies the same compatibility principle across the 3 categories identified (content licences, tools or software licences, and service agreements). In this way, it will be possible to develop an interoperability model or matrix that is not limited to content, tools or services individually considered, but that, by taking a holistic approach, is able to offer a more complete analysis of the licence compatibility issues faced by TDM researchers. In other words, this formulation, instead of taking a theoretical legal approach, puts at its centre the needs and the skills of TDM researchers, who usually are not legally trained.

4. Legal Metadata

The term "legal metadata" refers to the elements that describe in a formalised way all parameters related to the legal status of an asset, such as its usage terms and conditions, the copyright holders etc.

Attaching a licence to an asset (content, software tools or services) is the first step towards achieving legal interoperability in the ecosystem we are discussing; the clear indication of this licence in the description of the asset, e.g. by explicitly linking it to its licence, through the licence name, a url or a free text field with the legal text, is the next one, since it gives the user direct access to the licensing terms (Piperidis, 2012); the promotion of standard licences further increases legal interoperability, as the combination of widely used licences with known licensing terms becomes more manageable.

However, if we look at various distribution sites, we see that content and data providers tend to be agnostic or seemingly indifferent to stating access rights and rights of use. In addition, where providers do state rights, the serious lack of use of standardised frameworks makes interoperability a very difficult goal. For instance, the use of classification badges/categories such as embargo, closed/open access, restricted (from OpenAire), rights reserved – free / paid access (from Europeana) may be sufficient for the original purposes for which a particular infrastructure has been built, or when the user intends only to read or view a resource for his/her personal use, but it doesn't satisfy any other needs. Can these resources be safely used for TDM and, if yes, can the outcome of the process be used for commercial applications?

Finally, an important instrument for achieving legal interoperability is the encoding of licensing terms (a la CC primitives) in the form of conventional metadata rather than free text statements. This, however, can only be fully accomplished if the semantics of the licensing terms are properly defined thus allowing for valid mappings between concepts of different licences. Rights Expression Languages (REL), such as ODRL, with their non-flat structure, support a better modelling of the licensing terms and conditions; they are also extensible and can, therefore, represent new licensing terms should the case arise (Rodriguez-Doncel and Labropoulou, 2015; ODRL Version 2.0 Core Model, 2012).

5. Interoperability Problems

The OMTD project is confronted with various legal interoperability issues in order to cater for automatic processing.

At the theoretical level, we need first to clarify copyright, related rights and SGDR and how these influence the use of assets, as discussed in Section 3.

Given that OMTD (and likewise any other digital infrastructure) operates at a supra-national level, we must also look at how national law and national licences can operate at a cross-border setting: how assets created and copyrighted in one country circulate in countries with different legal provisions?

Multiple licensing of an asset can also hinder interoperability as it is not always clearly used: multiple licences are used for accumulative cases (e.g. for a corpus accessed via an interface, where each of these components is licensed with a different licence and the user must conform to the licensing terms of both of them), or for different uses in different contexts (e.g. a resource distributed free of charge for research and through an interface but for-a-fee in a downloadable form for commercial uses).

Finally, combinations of content and tools licences, service agreements, and similar agreements in the case of creating workflows from different web services (or web services from different components), or combining input data from different sources.

At the more practical level of legal metadata, we encounter problems stemming mainly from the unclear semantics or poorly defined licensing elements (or differently defined across different licences). For instance, terms such as "adapted", "derived", "modified version” are not clear to non-legal experts, and their use in different licences creates confusion. Or the term "attribution" as defined in the CCPL ("You must give appropriate credit, provide a link to the license, and indicate if changes were made") includes the element of link to the licence, whereas OKFN includes this in the "notice" term ("The **license** *may* require retention of copyright notices and identification of the license”).

We will need to build a licence interoperability matrix that includes standard licences and their possible combinations showing which ones result to legitimate uses in the OMTD perspective; moreover, this should be implemented and included in the OMTD processes, so that only assets licensed under acceptable combinations are allowed to be selected. For this, we will need to identify the elements that are important for ensuring legal use vs. violation of rights, see how these interact across licences and formally encode them in the metadata. The display of the filtered aggregation of licences must also be user-friendly (Cieri & DiPersio, 2015). Accommodating properties of the user performing a mining operation, as these can be made available by authentication and authorization modules of the OMTD infrastructure, and correlating them with licensing metadata constitutes an additional level of regulating access to assets of the infrastructure.

For OMTD purposes, a calculus that computes the licence values of the mined output based on the licences of the input data and the components that participated in the operation could prove beneficial; the automatic generation of new metadata derived from the original metadata for legal elements is also in the same line.

6. Future Work

In the framework of OMTD, we will take initiatives to help clarify as far as possible the legal framework and overcome interoperability issues. Standardizing licences and promoting their use, as well as enforcing their encoding with metadata will be the first step. The standardization of the metadata and adoption of a common legal vocabulary will be promoted. And, of course, training users in understanding licences will be a key action.

7. Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme (H2020-EINFRA-2014-2) under grant agreement No. 654021. It reflects only the author's views and the EU is not liable for any use that may be made of the information contained therein.

8. Bibliographical References

Cieri, C., DiPersio, D. (2015). A License Scheme for a Global Federated Language Service Infrastructure. WLSI 2015: *The Second International Workshop on Worldwide Language Service Infrastructure*, Kyoto, January 22-23 (PDF).

De Wolf & Partners, (2014). *Study on the legal framework of text and data mining* (TDM).

Guibault, L. and Wiebe, A. (Eds.) (2013). *Safe to be open. Study on the protection of research data and recommendations for access and usage*. Göttingen: Universitätsverlag Göttingen.

Guibault, L. andMargoni, T. (2015). Legal Aspects of Open Access to Publicly Funded Research. In OECD*, Enquiries into Intellectual Property's Economic Impact*, pp. 373-414.

Keller, P., Margoni, T., Rybicka, K., and Tarkowski, A. (2014). *Re-Use of Public Sector Information in Cultural Heritage Institutions*, IFOSS Law Review.

ODRL Version 2.0 Core Model, Final Specification: 24 April 2012.

Piperidis, S. (2012). The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC’12), 23-25 May,European Language Resources Association (ELRA).

Rodriguez-Doncel, V. and Labropoulou, P. (2015). RDF Representation of Licenses for Language Resources, *4th Workshop on Linked Data in Linguistics: Resources and Applications*, ACL-IJCNLP 2015, Beijing, China.