This document is part of the Coordination and Support Action “Preparation and Launch of a Large-scale Action for Quality Translation Technology (QTLaunchPad)”. This project has received funding from the European Union’s Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 296347.

Deliverable D 4.5.1

Legal Framework

Author(s): Prodromos Tsiavos, Stelios Piperidis, Maria Gavrilidou, Penny Labropoulou, Tasos Patrikakos

Dissemination Level: Public

Date: 18.03.2014

Grant agreement no.	296347
Project acronym	QTLaunchPad
Project full title	Preparation and Launch of a Large-scale Action for Quality Translation Technology
Funding scheme	Coordination and Support Action
Coordinator	Prof. Hans Uszkoreit (DFKI)
Start date, duration	1 July 2012, 24 months
Distribution	Public
Contractual date of delivery	February 2014
Actual date of delivery	March 2014
Deliverable number	4.5.1
Deliverable title	Legal Framework
Type	Report
Status and version	
Number of pages	90
Contributing partners	DFKI, Subcontractors
WP leader	ILSP
Task leader	ILSP
Authors	Prodromos Tsiavos, Stelios Piperidis, Maria Gavrilidou, Penny Labropoulou, Tasos Patrikakos
EC project officer	Aleksandra Wesolowska
The partners in QTLaunchPad are:	Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany
	Dublin City University (DCU), Ireland
	Institute for Language and Speech Processing, R.C. “Athena” (ILSP/ATHENA RC), Greece
	The University of Sheffield (USFD), United Kingdom

For copies of reports, updates on project activities and other QTLaunchPad-related information, contact:

DFKI GmbH
QTLaunchPad
Dr. Aljoscha Burchardt
Alt-Moabit 91c
10559 Berlin, Germany

aljoscha.burchardt@dfki.de
Phone: +49 (30) 23895-1838
Fax: +49 (30) 23895-1810

Copies of reports and other material can also be accessed via http://www.qt21.eu/launchpad

Table of Contents

Abbreviations & Acronyms 6
1 Introduction 7
2 Methodological Approach 8
3 Legal Issues of MT & MP 13
3.1 Copyright and Related Rights 14
3.1.1 Key Copyright Concepts 14
3.1.2 Related (or Neighbouring) Rights 15
3.1.3 Sui Generis Rights 15
3.1.4 Technological Protection Rights 15
3.1.5 Differentiating Copyright from other Intellectual Property Rights (IPR) 15
3.1.6 Copyright Boundaries 16
3.1.7 Public Domain 16
3.1.8 Works excluded from Copyright 17
3.1.9 Systems of Copyright Exceptions 18
3.1.10 Fair Use vs. EU limitations and exceptions approach 19
3.1.11 The user-creator rights movement 20
3.2 Licensing 20
3.2.1 Types of licences 20
3.2.2 Creative Commons Licences 22
3.3 Copyright Analysis of a MT & MP scenarios based on crawling 27
3.3.1 Objectives 27
3.3.2 Methodological Approach 28
3.3.3 LR processing and LT development 29
3.3.4 The issue of Data Mining and Crawling 30
3.3.5 Grouping and understanding Cases of LR re-use 36
3.3.6 Conclusions 41
3.4 Data Protection 42
3.4.1 Key Data Protection Concepts 42
3.4.2 Key Data Protection Principles 43
3.4.3 Sensitive Data 43
3.4.4 Historical Evolution of Personal Data Protection and its implications for MT & MP 44
3.4.5 Standard Licensing Models for Personal Data 51
3.4.6 Legal Basis for the Use of Personal Data 56

3.5 Public Sector Information 57
3.5.1 Works not granted copyright or exempted from copyright and similar rights protection 58
3.5.2 Expiration of the Copyright Term (Public Domain Works) 60
3.5.3 Limitations and Exceptions 60
3.5.4 Use of Marks and Notices 61
3.5.5 Licensing of Public Domain material released as PSI? 62
3.5.6 Concluding Remarks and Recommendations for PSBs licensing 63
3.6 Confidential Information and Other Rights/Agreements 67
3.7 Risk Mitigation Strategies and Measures 67
4 Concluding Remarks 70
5 Appendix: Use Cases 72
5.1 Public Sector Information (PSI) Use Cases 72
5.1.1 Case #1: Uploading-copying Public data (normally under PSI directive) to a repository 72
5.1.2 Case #2: Uploading-copying Public data (normally under PSI directive) to a repository 73
5.1.3 Case #3: Uploading-copying "Open" data to a repository 74
5.2 Open Data and Web crawling Use Cases 75
5.2.1 Case #4: Uploading-copying "Open"/"Public domain"/web crawled data to a repository 75
5.2.2 Case #5 Distributing web crawled data I 76
5.2.3 Case #6 Distributing web crawled data II 77
5.2.4 Case #7 Distributing web crawled data III 78
5.2.5 Case #8 Distributing web crawled data IV 79
5.2.6 Case #9 Distributing web crawled data V 80
5.2.7 Case #10 Distributing web crawled data VI 81
5.2.8 Case #11 Distributing web crawled data VII 82
5.3 Distribution of Translated Data Use Cases 83
5.3.1 Case #12 Distribution of Translated Data I 83
5.3.2 Case #13 Distribution of Translated Data II 84
5.3.3 Case #14 Distribution of Translated Data III 85
5.4 Data Anonymisation Use Cases 86
5.4.1 Case #15 Anonymising Data Sets 86
5.5 Lexicon Distribution 86
5.5.1 Case #16 Lexicon Distribution (based on other datasets) I 86
5.5.2 Case #17 Lexicon Distribution (based on other datasets) II 87

5.5.3 Case #18 Lexicon Distribution (based on other datasets) III 88
5.6 Annotation Cases 89
5.6.1 Case #19 Distribution of Annotations of a Data Set I 89

Abbreviations & Acronyms

CR: Copyright

DB: Database Right

IFLA: International Federation of Library Associations and Institutions

IPR: Intellectual Property Rights

LR: Language Resources

MP: Machine Processing

MT: Machine Translation

PSI: Public Sector Information

SGDBR: Sui Generis Database Right

1 Introduction

Objective of this report is to present the key legal issues with regard to Language Resources (LR) re-use, particularly in Machine Translation (MT) and Machine Processing (MP) settings, provide a simple report as to the permitted acts and make key policy suggestions as to amendments in the relevant bodies of legislation.

LR processing requires the use and re-use of information of various kinds, in a variety of ways, by different types of organisations and, as a result, it involves a wide range of legal regimes. It is mostly related to Intellectual Property Rights (IPR), but it may also involve Personal Data Protection, Public Sector Information and Geodata Regulation.

In order to assess the conditions under which LR re-use in MT and MP may lawfully take place, we need to appreciate the acts it involves, the degree to which they are regulated by different types of laws and the permissions that someone needs to obtain in order to perform such acts.

This report comprises the following sections:

Section 2: features the key methodological premises of this report. It explains how the flows of rights are to be treated in a paradigmatic MT & MP scenario and what the key emerging issues are.

Section 3: presents:

(a) core concepts from the main legal regimes that influence the LR-based MT & MP. It also revisits the methodology explained in Section 2 in order to make a more elaborate presentation of the key legal issues featuring in each of the different kinds of legal regimes.

(b) various legal issues related to LR-based MT & MP in more detail (See Figure I). More specifically it presents:

- IPR issues
- Data Protection Issues
- Public Sector Information Issues
- Confidentiality and Other types of rights Issues

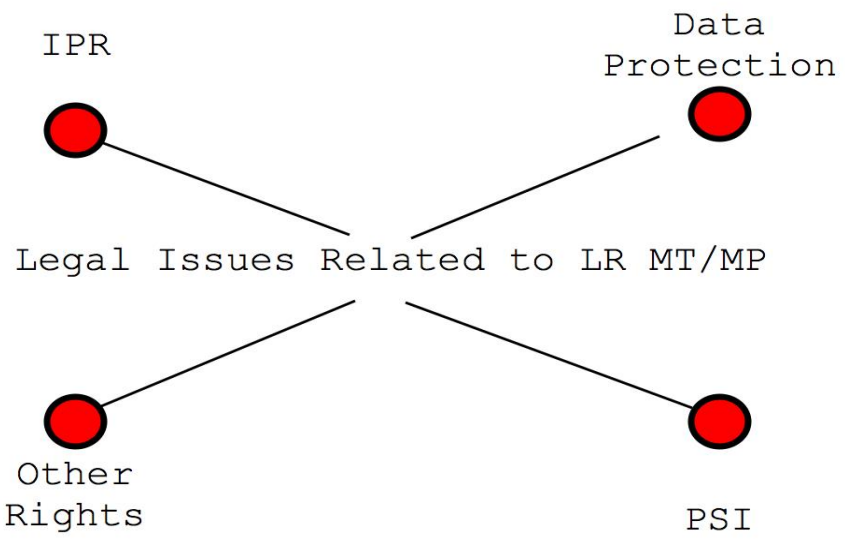

Figure I

Section 4: presents the core conclusions with regard to the way LR-based MT & MP are to be treated.

Section 5: presents core use cases that may be used as scenarios illustrating different aspects of the use of LR for MP/MT purposes.

2 Methodological Approach

LR-based MT & MP requires the collection of large amounts of information, the creation of compilations and databases, simple or annotated, in the form of *corpora*, their processing and the provision of services on the basis of such information. In addition, the nature of MT services requires that the texts used for MP are representative of a particular domain or activity. For instance, legislation, court decisions and other regulatory texts are necessary primary material for the provision of MT services in the area of law, whereas engineering manuals or other technical descriptions may be useful for the provision of MT services in an engineering sector and literature or news may be useful for literature or generic MT service provision. As a result, for any MT service - or indeed any LR-based service - to be provided, it is necessary to have access to the largest possible quantity of material that could be machine processed. This practically means that such raw material will almost invariably constitute some form of work that will be covered, either by a property right, as is the case with copyrighted documents, or by other types of rights, as is the case with the transcription of phone conversations, which are very likely to contain personal data.

It is also expected that a single set of data/information/works may constitute the subject matter of different types of rights. For instance, in the previous example, the same set of transcriptions of phone conversations constitutes a literary work protected by copyright, personal data that have to be processed under the data protection rules and, finally, may contain confidential information that should not be divulged beyond the persons having the conversations. Similarly, almost all Public Sector Information will be copyrighted subject matter, whereas it is very likely they contain personal data or other forms of confidential information.

It is, hence, necessary to trace the flows of rights in a paradigmatic case of MT services and set out the key questions that have to be asked irrespective of the specific legal regime that is involved. This is a task that is carried out in this section. The next step, then, is to further explore these issues with regard to specific types of legislation, as the ones identified above, namely: IPR, Data Protection, Personal Data Protection, Confidentiality and Public Sector Information Regulation.

Another important point to make at this stage is that different types of content are, in legal terms, differently treated and, consequently, the terminology used to describe such content varies. For instance, a document may constitute a work in terms of copyright law, but may also contain personal data or be confidential information and be part of a broader range of public sector information. We will use the term “content” here, as it is one that is more generic and appears legal regime neutral.

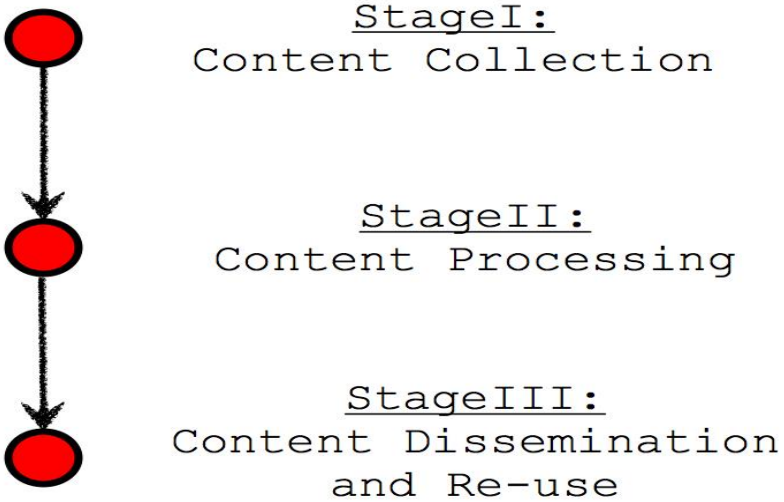

Figure II

We may identify the following three stages (see Figure II) in the life cycle of the content that can be machine processed in a case of Machine Translation service provision:

Stage I: Content Collection

At this stage, the content is collected from various sources. This is a rather important moment, as different sources come with different sets of rights. In all cases, the aggregator/collector of the content has to specify the following (Figure III):

(a) Whether the content is provided by the content provider under a licence, set of terms and conditions or any other prescription as to how the content is to be used.

(b) Whether there are any special rules that allow the content aggregator/collector to get access to that content without obtaining permission. In these cases, it is necessary to be very clear as regards the conditions set by the law that actually allow the obtaining of the content without the consent of the content provider. The content aggregator/collector may assess the degree to which such acts may be performed on the basis of what will take place during stages II and III, i.e. on the basis of how the content is to be processed and then further disseminated.

(c) If there is an open licence in place, then the LR-based MP/MT service provider only needs to check whether the processing of the LR is in accordance to the terms of the licence.

(d) In the case there are no conditions or where the conditions are not clear, the content aggregator/collector will have to request for permission, in a way that is well documented both as regards the communication stage and its result. The licences, consent or other permission obtained by the content aggregator/collector have to be assessed with regard to the acts of processing and dissemination that will take place at subsequent stages. They may also be amended at a subsequent stage, if the acts of processing or dissemination exceed those originally perceived and communicated during the licence obtaining stage.

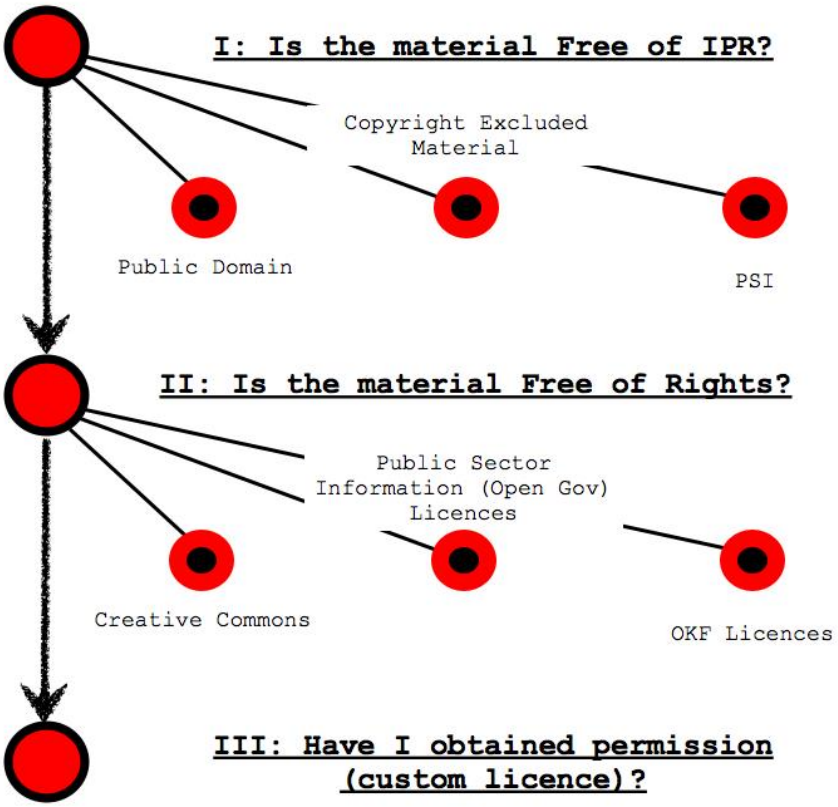

Figure III

A horizontal note at the stage of content collection is that, though it precedes the stages of content processing and dissemination, it cannot be properly assessed if the range of processing and dissemination acts has not be defined or at least projected. This is because both the case of performing acts without asking permission and the case of obtaining licences/consent have to be assessed on the basis of the intended processing and dissemination.

Stage II: Content Processing

At the stage of processing, all necessary permissions have been obtained or the fact that that there is no need for permissions has been established.

However, as mentioned above, it is essential to establish the range of processing acts before the permissions are sought or a judgement is made with regard to whether the processing

may take place without any permission. The most frequently attested acts of processing have as follows:

- **Copying**: it is impossible to make use of any LRs without copying them and hence, almost invariably, MT & MP will activate the provisions of Copyright Law.
- Making **derivative** works: in order to provide MT services, it is necessary to process the collected content in order to create different forms of derivative works, e.g. n-grams or different forms of corpora. These transformative uses of content activate mostly provisions of copyright law but could also have data protection or other regulatory regimes implications.
- **Linking** with other works: linking of content with multiple types of works. Such linking does not necessary activate copyright law, but could have data protection implications.
- **Other types** of processing: almost any type of processing will activate at least one type of regulatory regime, with data protection regulation to provide the broadest range of acts that are within its scope, as almost every act will amount to processing. However, whether a specific regulatory regime will be activated will depend on other factors as well, such as whether the data are personal or confidential, the type of subject matter and the permissions that have been in one way or another obtained.

Stage III: Content Dissemination and Re-use

This is the last stage in the life-cycle of the content collected for LR processing and at the same time a stage that may re-initiate the life-cycle through the re-use and enrichment of the relevant content.

More specifically:

- Similarly to stage II, the types of dissemination to a great extent define the type of permissions that should be sought in stage I. For instance, if the content is not to be shared as such, the legal risks are substantially reduced, as when compared with cases where the content is further made available as such. If the content is to be made available for re-use, the range of permissions that are to be sought will be far greater than the rights sought when the content is to be made simply available for end use, i.e. not for derivative uses or if it is not to be made available at all.
- Overall, we may identify the following types of content dissemination:

- The content is used for Machine Training and is never disseminated as such. However, it could be that the end result of the Machine Translation service is based on the original content or even comprise original parts of such content.
- The content is offered as part of a service, but in a transformed form, and hence it is not necessary that it will constitute a derivative work and in any fashion activate copyright law.
- The content is offered as a data-set, possibly with some form of amendments and annotations, but it may be clearly identified as the original material.

Depending on the type of content delivery the legal regimes activated and the type of permissions required differs. The course of action has to be decided on a case-by-case basis.

- Finally, the MT service provider will have to decide on two additional issues:
  - the licence which the content (that is the automated translation output) will be offered under
  - the terms under which the service is going to be offered.

In the following section we present the implications of LR-based MP and MT services for the following types of rights:

(a) IPR

(b) Data Protection

(c) Confidential Information

(d) Public Sector Information

3 Legal Issues of MT & MP

In this section we present various legal implications of LR-based MP and MT services. Objective of each sub-section is to illustrate the range of permitted acts, the types of permissions that need to be obtained and to suggest ways in which the obtained material may be further disseminated.

It should be stressed that this is a document that contains only legal information and not legal advice and should not be construed as such.

3.1 Copyright and Related Rights

This is probably the legal regime posing the greatest challenges and at the same time having the greatest implications in the way in which LRs are used in MP and MT contexts.

The section presents some key Copyright concepts and then moves to discuss the specifics of LR processing in a typical MP and MT scenario.

3.1.1 Key Copyright Concepts

3.1.1.1 Copyright/Author's Right

Copyright is a form of Intellectual Property (IP) that protects the fixation[1] of the original expression of an idea. Another definition is that of the right granted for the protection of literary, dramatic, artistic, musical and other works resulting from the author's own intellectual creation. It may be differentiated from Related or Neighbouring rights that include rights granted to contributors of subservient works (such as actors performing a dramatic work) or non-creative contributors (e.g. producers). It is also differentiated from the sui generis database right (SGDR), that is a property right of limited duration granted to the producer of a non-original database. The term copyright is used here as a catch-all phrase to describe rights granted to authors or rights-holders in any legal system.

3.1.1.1.1 Differences between Copyright and Author's Right

The most important differences between the two systems of protection are the following:

(a) Copyright gives greater emphasis on the protection of the work, whereas the Author's Right system gives more emphasis to the author. The most important implication of this difference is that in a Copyright system we may have a legal person as the first owner of the work, whereas in the Author's Right system, the norm is that the right originates in the individual.

(b) In the Copyright system the general rule is that the employer is the first owner of the work, whereas in the Author's Right system, the employee gets to be the first owner of the work, though there may be presumptions as to how it is transferred to the employer, achieving the same end-result.

[1] Fixation is a condition for Copyright Law in the UK (CPDA 1988)

(c) In the Copyright system the threshold of originality is generally lower than that of the Author's Right system: in the former, the "sweat of the brow" or "skill, judgement and labour test" conditions will suffice to grant a work, whereas in the latter, the work has to be the "author's own intellectual creation" or "bear the stamp of her personality". Whereas the two systems appear to have different tests, in most cases, a work judged as original in one system will probably be judged as original to the other system as well.

3.1.2 Related (or Neighbouring) Rights

Related rights are the rights granted for the protection of performers, producers, broadcasters and other non creative or subservient work producers. In some jurisdictions, notably in Commonwealth jurisdictions, the term copyright is used to cover both the rights of the authors and some or all the related rights.

3.1.3 Sui Generis Rights

These include rights that the legislator has chosen to differentiate from author's rights and related rights and characterise them as special nature rights, also known as sui generis rights. The one which is relevant for this report is the sui generis database right (SGDBR), found mostly in EU jurisdictions and related mostly to digital technology. Other sui generis rights are the right granted to the semiconductor topographies makers.

3.1.4 Technological Protection Rights

Some authors classify the rights to Technical Measures of Protection as a separate category of rights, whereas others number them among the economic rights comprising Copyright.

3.1.5 Differentiating Copyright from other Intellectual Property Rights (IPR)

Copyright and the rights presented above are to be differentiated from other forms of IPR, mostly patents, trademarks, industrial designs, trade secrets and other forms of IPR, including unfair competition, contracts and tortious acts. Whereas Copyright infringement may often involve the infringement of other IPR as well, these are beyond the scope of this paper and we will not endeavor in a full explanation of each one of them. Very briefly:

- Patents are monopolies limited in time granted to protect inventions and new methods of manufacture. In general they last 20 years from their registration.
- Trademarks are marks protecting the origin of a product or a service.
- Industrial Designs are designs used for industrial manufacture of article in quantity.
- Trade secrets are protected under the terms of confidential information and are often characterized as such through confidentiality or Non Disclosure Agreements (NDAs).

3.1.6 Copyright Boundaries

The copyright boundaries are set by a number of elements, the most relevant of which for this paper are as follows:

- Beneficiaries: these are the owners of the right. They are useful to know as they are the ones the licences (if required) should be obtained from.
- Subject matter of protection: it describes the categories of protected material. What is not protected (e.g. ideas, facts) is excluded from copyright protection.
- Moral and Economic Rights: these are the range of rights that can be exercised over the protected subject matter by the beneficiaries or the licensed rights holders.
- Limitations and exceptions: they include acts that because of different reasons (social, economic, practical) may be exercised by a user of the work without requiring the permission of the rights-holders
- Term of protection: it sets the term of the protection for copyright, related rights or sui generis rights. The general rule is that it is the author's life plus 70 years for literary and artistic works (Copyright) and 70 years from the making of recordings or other related rights protected subject matter. The SGDBR has a duration of 15 years from the creation or substantial modification of a database.

3.1.7 Public Domain

If a work is:

- exempted from Copyright; or
- outside the term of protection; or

- the acts related to it are within the limitations and exceptions; or not covered by copyright law

Then it is in the Public Domain.

If the work is licensed under an open licence, then it is considered to be in the *functional Public Domain,* i.e. while it is legally copyrighted, the range of rights granted by the licensor to the licensee are such that it is effectively as if it were in the public domain.

3.1.8 Works excluded from Copyright

There is no international harmonization as to the works that are excluded from Copyright protection. However, the following categories are highly likely to be outside the scope of copyright protection:

Works of applied art and industrial designs and models (art. 2(7) of the Berne Convention)

- Illegally made works:[2]
    - That infringe copyright (only in the US to the extent that the infringing material has unlawfully used other material (s.103(a) of the US Copyright Act). In other jurisdictions infringing material attracts copyright protection irrespective of its legality (e.g. UK, Germany, France)
    - That breach other types of laws, particularly property laws (e.g. a graffiti over someone else's wall - see "Berlin Wall Pictures" case[3])
- News of the day and press Information (art. 2(8) of the Berne Convention)
- Official material, which is frequently qualified as Public Sector Information (PSI) as well (see respective section).
    - In some jurisdictions, such as Italy, Germany and Greece, material such as laws, orders, government reports and decisions are excluded from copyright protection. Similarly, in the US copyright protection is not available for any works of the US government.
    - In the UK, the Crown has copyright in Acts of Parliament and other items of official material.

[2] It is highly unlikely to see such cases in the MT & MP context.

[3] "Berlin Wall Pictures", BGH, Feb.23, 1995: (1995) G.R.U.R. 673; (1997) 28 I.I.C. 282.

- Four points have to be highlighted here:
  - In most jurisdictions, government reports are copyrighted subject matter.
  - Foreign Official Material is protected outside the country of origin
  - Court judgements may be excluded from copyright protection; however, reporting of the judgements is not
  - Licensing seems to be a prevailing practice with regard to Official Material (or PSI) and this is always to be examined.

- Political Speeches and speeches delivered in the course of legal proceedings may be excluded from Copyright protection in accordance to the Berne Convention (Art.2bis(1))
- Seditious or Obscene Material may be refused protection by the courts, where the national laws allow them to take such a decision.

3.1.9 Systems of Copyright Exceptions

The system of copyright exceptions serves as a mechanism to ensure that access to certain types of works is ensured on considerations of public interest, balancing of rights, practical concerns or other. These appear as:

- permitted uses that do not require a special permission or a fee but are based on some requirements (e.g. attribution of the source, limited use of material etc.)
- compulsory or statutory licences.

The restrictions may be further classified in two categories:

- exceptions that are part of the copyright law
- exceptions that stem from other types of laws (e.g. competition law, property law, constitutional restrictions)

The most frequently found exceptions are:

- private copying
- time, space and format shifting
- criticism or review
- education
- preservation of copies or use of them by libraries
- research and private study

- certain aspects of use of computer programs
- use of databases, particularly the uses which are necessary for the regular use of the database by its lawful user
- circumvention of technical measures of protection
- text and data mining (in the UK as of 2014).

From the exceptions mentioned above, only the research and, more recently, the text and data mining exceptions are relevant for MT & MP purposes.

All limitations and exceptions have to be implemented in the national jurisdictions following the three step test (art. 9(2) of the Stockholm Act of the Berne Convention):

(a) the limitation or exception may only apply in certain special cases;

(b) the limitation or exception must not conflict with the normal exploitation of the work; and

(c) the limitation or exception must not unreasonably prejudice the legitimate interests of the author.

3.1.10 Fair Use vs. EU limitations and exceptions approach

The Fair Use approach, which is the US system for copyright limitations and exceptions contrasts to the EU system of limitations and exceptions in the following ways:

- the Fair Use system is doctrine-based and open ended: it is based on rules that are then further specified by the courts, whereas the EU system is based on a limited list of limitations and exceptions
- it is more flexible and future proof as it is not based on a particular technology or a closed list of exceptions.

The Fair Use system provides that the fair use of a copyrighted work for purposes such as criticism, comment, news reporting, teaching, scholarship or research is not an infringement (s.107 of the US copyright Act). The following factors have to be considered in order to decide whether a work infringes copyright or not:

(a) the purpose and character of the use, including whether such use is of a commercial nature or is of nonprofit educational purposes;

(b) the nature of the copyrighted work;

(c) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and

(d) the effect of the use upon the potential market for or value of the copyrighted work.

3.1.11 The user-creator rights movement

It is important to highlight that both Fair Use and limitations and exceptions do not constitute positive rights and can only be used as a defence against an action for copyright infringement. The lack of a harmonised international system of limitations and exceptions as well as the lack of a balanced approach with regard to the treatment of current and future creators has led to the emergence of the user-creator rights movement that aims at the establishment of international or regional instruments providing a harmonized horizontal system of positive rights for the users of content that potentially are future creators.

3.2 Licensing

3.2.1 Types of licences

This is the most important exploitation and dissemination instrument for copyright law. It is essentially a set of permissions covering the rights that the right-holder/licensor wishes to grant to the licensee. In the case of MP/MT, licences are important both as an instrument for receiving permissions (licensing in) and as instrument for granting permissions (licensing out). The following set of definitions provide the key types of licensing, making reference to PSI licences as well. PSI licences are of particular interest as they constitute a key instrument for the release of huge amounts of information by Public Sector Bodies (PSBs).

Non-transactional Licences are the licences that take effect through the actual use of the licensed material, without the need for any additional transactions.

Re-use Licences are the licences that allow the re-user to use the PSI in a fashion other than the one originally intended by the PSB, but are not necessarily non-transactional, neither open or standardised. These are licences used almost exclusively in the PSI context.

Re-usable Licences are standard licences that are publicly available and may be re-used by any licensor without any modifications. A highly re-usable licence is normally stored at permanent URI and has a community supporting its updating.

Standard Licences are licences that are addressed to a non-specified range of recipients and are not the result of individual negotiations between the licensor and the licensee.

Open Access (OA) is online access to peer-reviewed scholarly research. OA has two degrees: (a) gratis OA, which is online access to scholarly resources for free, and (b) libre OA, which is online access to scholarly resources for free, with some additional freedoms for the end/re-user, which are normally granted through Creative Commons or other Open Licences. These are described, in this Report, **as Open Access Licences**.

Open Licences are all standard, non-transactional licences that, to some extent, allow the end user to engage in the 4Rs, i.e. Re-use, Revise, Remix and Redistribute. Licences that allow all 4Rs under the sole condition of attribution or share-alike, comply with the Open Knowledge Definition and constitute Open Knowledge Definition (OKD) Conformant Licences. From the Creative Commons Licences, only Creative Commons Attribution and Creative Commons Share-Alike are Open Knowledge Definition Conformant Licences. From the Open Knowledge Foundation, the Open Data Commons Open Database License (ODbL) and the Open Data Commons Attribution (ODC-BY) licences are ODF conformant. According to the OKF, UK's OGL 2.0 is characterised as a conformant but "non-reusable" licence", in the sense that it cannot be re-used by any public sector body (PSB) in any EU Member State. However, it needs to be stressed out that the OGL was drafted to be used by any UK PSB, not just the UK government departments and agencies, and covers all PSI, including but not limited to Crown Copyright. The PSI released under the UK OGL 2.0 may be licensed under CC-BY[4] or ODC-BY[5].

Copyleft Licences are all licences containing terms that allow modifications to the licensed work on the condition that the work is further disseminated under the same terms and conditions. In the Creative Commons set of licences, all Share-Alike licences are copyleft licences.

Input Licences: are all licences acquired by a PSB in order to be able to release PSI.

Output Licences: are all licences under which PSB makes content available to re-users.

Implied licences: are licences that are not expressed in the form of a text but are rather implied by the conduct of the licensor. Examples of an implied licence would be allowing web robots (bots) and crawlers to obtain data or content from a web page by not instructing them otherwise through the html code of the relevant web page.

[4] http://creativecommons.org/licenses/by/4.0/

[5] http://opendefinition.org/licenses/odc-by/

3.2.2 Creative Commons Licences

All Creative Commons (CC) licences are comprised of a combination of the four high level Licence Elements. These are the core terms of the CC licences that may be combined with each other in order to produce the different CC licences. The CC licence elements are the following:

- Attribution
- Non-commercial
- Share-Alike
- No Derivatives

The licensor may choose a combination of the above in order to build the licence that suits most her needs.

In the rest of this subsection we present each of the elements separately.

Attribution (BY)

While licensors may choose any of the above elements, all the CC licences contain the attribution element. This has been the result of the need to accommodate moral rights within the CC licences and also an element that in the first year of the operation of the CC licences has been chosen by all CC licensors. One of the key functions of the CC licences is to allow the maximum dissemination of the work in order to increase the reputation of the creator and, in that sense, the Attribution element is a quintessential element of any open content licence.

In addition, for the rights holders and authors not wishing to use the attribution element, there is always the CC Zero tool that waives all economic rights and does not contain any positive attribution requirements. However, even in CC Zero it is possible to require attribution not in the form of a formal legal requirement but rather in the form of a soft norm that is attached to CC Zero.

The CC wizard allows licensors to expressly describe how they would like to be attributed. This is an important aspect of the CC licences as it allows the licensor/author to opt for the attribution model that is closer to her objectives and goals.

Non-commercial (NC)

The Non-commercial element is one of the most widely used. It grants the licensee permission to copy, distribute, display, perform, and remix the licensed work for non-commercial purposes

only. The Non-commercial Licence Element means that the licensee cannot use the work commercially unless she receives an additional permission from the licensor. However, because CC Licences are non-exclusive, the Non Commercial Licence Element would allow the Licensor to themselves commercially exploit the work, and grant licences to others to be able to use the work for commercial purposes.

Share-Alike (SA)

It is important to highlight that the Share-Alike element refers to derivative works. It means that if the licensor creates a derivative work and decides to further disseminate it or otherwise make it available, this needs to be done under the same terms and conditions as the original licence. For instance, if the original work is disseminated under a CC BY NC SA licence, then the derivative work also has to be disseminated under the same licence. The CC BY SA licence is the licence used by Wikipedia and it allows licensing of the derivative works under any licence that has been approved as CC BY SA compatible by Creative Commons. The SA element does not refer to the original works that are always licensed directly from the licensor and hence are always licensed under the same terms and conditions no matter how many copies are being made. The SA element is a viral element in the sense that it triggers a proliferation of the CC licences as more derivative works are produced and in that sense the CC SA licences are copyleft licences.

No Derivatives (ND)

The No Derivatives licence element is used in order to allow licensees to copy, distribute, display, and perform only verbatim copies of your work. The creation of derivative works is not allowed unless an explicit permission is obtained from the licensor. The ND licence element is used in case where the licensor does not wish to allow any changes made to the original work but want to encourage people to disseminate it as widely as possible.

Figure IV

List of possible combinations

The combination of the aforementioned elements results in the following possible combinations:

- Attribution
- Attribution Non-commercial
- Attribution Non-commercial Share Alike
- Attribution Non-commercial No Derivatives
- Attribution No Derivatives
- Attribution Share-Alike

Figure V

Summing up the basic features of the CC project

The CC project has features that may be useful in projects that go beyond the problem of Copyright law. They were originally built in order to accommodate the need for a user-centric system where the individual is able to re-use the content without having to obtain permission for every subsequent use of the work. This is particularly useful in cases of user generated content and there is the question whether elements of this model could be used for the management of personal data. The most important of the CC project features are as follows:

- It is a human centric project: unlike standard licence agreements issued by companies that have their own legal teams, the CC licences are public, that is, they are offered to copyright holders to license their works and are addressed to potential re-users of the work. That means, they need to be easy to read and understand both by the licensor and the licensee.

- It is a licensor centric – author respecting project: it respects the rights of the author in the sense that she decides whether to license the work or not and moral rights are respected in all licences
- It is machine readable: it acknowledges the fact that most of the searching of the works is being done automatically and in that sense search engines need to be able to "read" and recognise the respective licence elements. This facilitates the easy identification, re-use and marking of the work and thus is in accordance with current creative practices on the Internet.
- It is standardised and modular: the licences are standard and in that sense their operation is simple and clear and once interoperability is achieved this is valid for the entirety of the CC project. At the same time, the CC licences are modular, i.e. they allow a limited set of combinations that provide the necessary level of flexibility to prospective licensors and users.
- It has paid great attention to its organisational and institutional rolling out: the project has been very successful not only because of the features of the licence but also because it has been rolled out through universities and the legal professors that actually taught the use of the licences to future lawyers. The process of transferring the licences to different legal systems involved the key academic and often professional institutions in different countries and in that sense allowed the institutional acceptance of the licences. Any similar process in any other area, such as privacy should involve a similar procedure.

Figure VI

3.3 Copyright Analysis of a MT & MP scenarios based on crawling

3.3.1 Objectives

Objective of this section is to address a series of issues with regard to the crawling, extraction and re-use of data that may be found under different licensing schemes or without any licensing information on the Internet.

Exploring this issue invariably touches upon multiple types of rights (copyright, sui generis right, public sector information) as well as aspects of rights (copy, making derivative works, communicating to the public).

The section presents the different types and aspects of rights involved in a rather structured fashion, so that it is possible to draw generic conclusions with regard to the level of risks

involved when re-using material used for the construction and use of Language Resources (LRs) or Language Technologies (LTs).

More specifically, this section presents the key methodological aspects of our approach for analysing different cases of data re-use in the context of the QTLP project.

3.3.2 Methodological Approach

In order to address each of the above points, we adopt the following approach:

- First, we present a series of steps that describe in a generic fashion the acts that are to be legally assessed.
- Second, we examine the legal status of these acts.
- Third, we assess the risks and opportunities that these acts entail.
- Fourth, we make suggestions as to alternative or additional actions that could reduce risk or increase the production of value with regard to these acts.
- Fifth, we make overall suggestions as to how legislation should change in order to address the issue of content re-use in order to make different types of LR processing and the development of LTs possible.

This set of steps draws from cases presented in Section 5, where different use-case scenarios with regard to the re-use of data are explored. These cases are then grouped to provide some core suggestions as to how to process and re-use data in the context of language resources (LR) processing.

Copyright Law is not fully harmonised at the international level and, hence, it is extremely difficult to provide a generic answer for the entirety of the situations involving more than one jurisdictions, where possible acts of infringement take place. However, there are some common rules described in international treaties, mainly the Berne Convention,[6] the TRIPS agreement[7] and the WIPO treaties[8] that provide us with an understanding of copyright rules

[6] http://www.wipo.int/treaties/en/text.jsp?file_id=283698

[7] http://www.wto.org/english/tratop_e/trips_e/t_agm0_e.htm

[8] http://www.wipo.int/treaties/en/ip/wct/ and http://www.wipo.int/treaties/en/ip/wppt/

at the international level, whereas a series of directives at the EU level provide an even more harmonised legal regime for the Member States of the European Union.

The section makes reference primarily to the legal system at the level of the international treaties placing additional emphasis to the regime established by the relevant EU Directives and making references to Copyright Legislation in some of the key jurisdictions outside the EU in terms of where the greater volume of data processing takes place mainly US, Canada and Australia. The main focus of the section is Copyright Law, but there are also references to Public Sector Information and Data Protection/Privacy regulations, where that is applicable.

For reasons of simplicity we will refer to the entirety of these regulations as “copyright law” with additional references to specific legal instruments where this is deemed necessary.

3.3.3 LR processing and LT development

LR processing and development is to a great extent dependent upon the use of third party material that may be found on the Internet, in specific collections or final products (mostly books or other types of publications). The processing of content of different types is essential for the production of some key LRs, such as annotations, lexicons and mostly corpora. As a result, it is necessary to have a good idea of how material could be used and re-used.

The key questions in this context may be summarised as follows:

- What are the key types of use relevant for the development of LRs and how should be treated in order for the (re) use to be lawful?
- What are the key types of permissions necessary to perform actions on LRs?
- What are the best licences for making LRs available?
- How can we use exemptions, fair use/dealing, limitations and exceptions in order to process LRs?
- What are the legislative amendments necessary to advance LR related research?

The following section deals with the core issue of data mining and web crawling that is the most problematic in terms of dealing with the issue of content re-use for LR/LT use. We return to these questions again at the end of this section.

3.3.4 The issue of Data Mining and Crawling

Data Crawling or Mining may be defined as the act of collecting different forms of information from the public Internet in an automatic fashion (i.e. through bots) which is then stored and processed in different ways.

It is necessary that this description be broken down into distinct steps that will be subsequently assessed in terms of the degree to which they constitute violations of copyright law in different jurisdictions.

The material may be found either on the Internet or in specific repositories. In the former case, the likelihood of having a specific licence attached to the material is much lower than when finding the material in a repository. However, it is often the case that even material that is found in a repository is the result of crawling and data-mining, hence the issue of data-mining is a core issue for our understanding of how LRs are to be legally used.

More specifically:

(1) All the material found on the web is material that potentially constitutes protected subject matter. This is mostly due to the fact that copyright protection does not require any formalities for the protection to be granted and, hence, there is no record of whether the material is protected or who the owner is or when the protection is to expire.

Protected subject matter may fall under the following broad categories:

- Textual information (literary works)
- Pictorial (artistic works)
- Audiovisual works
- Sound Recordings
- Musical Works
- Data Bases and compilations

This section mainly focuses on literary works and databases, though it is also applicable in cases where the other types of works are being crawled.

(2) A portion of these works may be outside copyright protection either because the term of protection has expired or because it falls under categories of works that are by definition not protected in certain jurisdictions.

- In the first category (expired copyright), we find works that have been produced by creators that have expired over 70 years ago (e.g. in the case of literary works) or works that have been produced over 50 or 70 years ago (e.g. in the case of sound recordings). The term of protection is calculated on the basis of a variety of factors, mostly

  [a] the type of work (e.g. literary work vs. sound recording)

  [b] the type of rights subsisting over the work (e.g. copyright vs. related rights) and [c] the jurisdiction of where the rights holder seeks protection (e.g. Australia vs. EU vs. US).

- In the second category (exempted material) we find subject matter that by virtue of their nature are classified as not protected works. These will mainly involve works made by the public administration or the legislature and which for reasons of public interest remain outside the realm of protection of copyright law. Some of these works are universally outside the protection of Copyright law (e.g. statutes in the EU) and some others are outside the protection only within a specific jurisdiction (e.g. statutes in the US or Public Sector Information (PSI) in certain EU Member States). In addition, these types of works in some jurisdictions are presented as works outside the realm of copyright protection and in some other jurisdictions as falling under the limitations and exceptions to copyright law.

(3) Depending on the type of work there may be different types of rights conferred to its creator, producer or performer. Hence, in the case of a literary work, copyright subsists as the main legal right; in the case of what is perceived as a single final work (e.g. an audiovisual work), multiple layers of works and rights may subsist (e.g. musical work, literary work, sound recording performance) with different durations and exceptions; in the case of a compilation of information, there may be different types of rights according to the type of creative input that led to the final work (e.g. copyright for the original compilation, sui generis database right for a database). The definition of the kinds of rights subsisting in a specific informational product depend on the jurisdiction where protection is sought, e.g. original databases are always treated as copyrighted

works in the US, whereas in the EU there are two types of rights, i.e. copyright for the original databases and the sui generis rights where only investment in time and labor has taken place. Finally, the level of originality required to grant protection may be different. For instance, in Australia the level of originality required to grant protection to a database is close to the definition of the non-original database in the EU, whereas in the US a greater level of originality is required. This means that the same work may have different levels of protection in different jurisdictions and, hence, what constitutes infringement in one jurisdiction may not have the same treatment in another. The most risk averse approach, hence, would be to take as a base line the highest level of protection (i.e. the existence of a sui-generis database right in all compilations of facts irrespectively of their originality) and act on the basis of very limited exceptions or a very narrowly construed fair dealing.

(4) It is necessary to specify the acts that are going to be performed upon the data and hence assess two factors: (a) the degree to which such acts fall within the acts restricted by copyright laws and (b) the extent to which such acts are visible enough to expose an organisation to the risk of legal action.

(a) In the case of web crawling, the acts would certainly include copying and processing of the relevant information and potentially the creation of derivative works and the communication to the public either of parts of the original work or a derivative work. Each of these acts needs special treatment:

- **Copying:** the act of crawling certainly involves the reproduction of content and hence activates the reproduction right. According to the Copyright Directive[9] any form of reproduction direct, indirect, temporary or permanent falls under the relevant economic rights of copyright and related rights holders and hence is regulated by copyright. In the case of crawling, the reproduction of the material could involve various quantities of material and could be temporary or permanent. In most of the cases of crawling for Language Technology purposes, the amount of material copied would be substantial both in qualitative and quantitative terms. It will be quantitatively substantial because otherwise there is not enough data for the LTs to perform operations that provide a meaningful result. It will also be qualitatively substantial, because it has significance for the entity

[9]Art. 2 of Directive 2001/29/EC of the European Parliament and the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society.

performing the processing and the parts of the material collected are by definition significant for the entity making the collection. The temporality of the copying is also a significant factor, but it seems that in the case of crawling for language resource processing there is very little of the temporary copying falling under article 5 of the Copyright Directive. This is because such temporary copying is allowed only in the case where it is used in order to facilitate either the transmission in a network between third parties by and intermediary or a lawful use that has no independent economic significance. It is almost impossible that even such a temporary reproduction in the LT context would fall under this exception since it does by definition have an economic significance. In any case, there are recent developments in national copyright legislation, particularly in Germany and France, where draft legislation has been proposed introducing a "snippeting right" with duration of a year. Under this new right news publishers would be able to license out snippeting rights for a royalty and start proceedings against those found to infringe their newfound neighbouring right. They would also be able to grant permission to reproduce to the relevant intermediaries for free. This is a trend that follows the two Infopaq cases decided by the European Court of Justice in 2009 and 2012 respectively and having been the result of heavy criticism by copyright academics and practitioners. In the Infopaq I case, the Court decided that snippets of 11 words may, depending on national law, be entitled to copyright protection under the European directives if they can be found to constitute an expression of the intellectual creation of their author. Accordingly, *originality* and not *substantiality* is the test that determines the copyright status of extracted parts of a work. In Infopaq II, the Court further noted that the transient copying exception to copyright enshrined in Article 5(1) of the Copyright Directive only applies if the act of temporary reproduction does not enable the generation of an additional profit beyond that derived from the lawful use of the protected work and does not lead to a modification of the work – under this interpretation the reproduction of news snippets by an automated process would not qualify as a protected use. Similarly, in 2011, the English Court of Appeal in Meltwater found that Meltwater News, an electronic media monitoring service, could be implicating its subscribers in copyright infringement by distributing sections that included the headline,

opening text and extracts from claimant Newspaper Licensing Agency (NLA)'s articles. Businesses that access press-monitoring services without a special web end-user licence may thus be in breach of publishers' content, notwithstanding any licence held by the press-monitoring agency. It becomes clear, hence, that in most cases of web crawling for LT purposes, the exceptions of art. 5 of the Copyright Directive would not be applicable, neither would the most of national laws in the EU accept it as falling within the realm of copyright limitations and exceptions.

- **Derivative works:** in most cases, the act of crawling will either mean the collection of only parts of the websites or will also include additional processing once the material is collected. As a result, derivative works will be created and hence additional permissions by the rights holders may be required. As demonstrated in the previous section, the act of creating derivative works cannot by construed as falling under the limitations and exceptions provisions and hence it will also require separate permission by the rights holders.
- **Communicating to the public:** the final part of a series of acts starting with web crawling and continuing with the processing of the collected data could be the communication of the results to the public. If what is communicated to the public is the actual data either in their original or their derivative form, then this constitutes yet another act restricted by copyright law. If, however, the end user is only the recipient of a web service that is the result of web crawling and processing of the relevant data without any direct communication of the actual web data to the audience in an identifiable form, then copyright law is not activated at all. It needs to be made clear that this is the case when no copying is involved. If this is not the case, then copyright applies.

(b) A separate question is the degree to which the act of crawling and any subsequent acts of data processing and dissemination are visible enough to expose the relevant entities to the risk of lawsuits. Unless the owner of each web site explicitly wishes the contents of her site not to be indexed or copied, the act of web crawling is part of the daily operation of a web site and hence it could be covered by an implied licence. Indeed, web sites need to be copied at least temporarily in order to be viewed and hence the simple act of web crawling may not be something that is noticed or objected by the web site owner. In addition, the processing of

information or the selected copying from the web site may occur in the site of the entity producing the LTs and hence not really perceptible to the rest of the world. If the LTs are offered as a service, the probability that a third person establishes a link between the infringement of a single web site and the final service offered to the end user becomes extremely low. Accordingly, unless the information crawled from a specific web site is substantial for the operation of the end user service, the legal risk drops dramatically.

(5) The previous analysis indicates that while the acts of web crawling and subsequent processing and communication of the relevant material constitutes copyright infringement and is unlikely to fall under the limitations and exceptions to Copyright law, the actual risk of legal action is fairly low and may be further mitigated through the following actions:

- It is necessary to identify big content providers whose content is crawled and is significant for the entirety of the collection of the entity that performs the crawling. This would be the case, for instance, of a big publisher or a newspaper licensing agency.
- If a commercial service is offered by the entity that performs the crawling, then it is good practice to contact the collecting societies of the jurisdiction in which it has its main place of operation and inquire whether there is a licence that actually covers the act of web crawling in its jurisdiction. However, if the LT provider is not using material from a specific jurisdiction or is mainly involved in non-commercial activities it may be more prudent to rely on an implied licence rather than seek for a commercial licence from the collecting societies.
- In some jurisdictions (especially in the US, where fair use is applicable), there is the concept of the implied licence with regard to web crawling. This legal construct relies on the fact that web browsing is not possible without the reproduction of the contents of a web site, that most of the owners derive value when their website is crawled or indexed and that there are technical measures to stop crawling, which are easy to apply and hence if they do not exist, imply that the web site owner wishes it to be copied. Objections to this line of argumentation include that, at least in the civil law jurisdictions, licences are very narrowly construed only to cover the explicit acts the rights owner would like to authorise. In that sense, web browsing or indexing or caching for a search engine is different from crawling for Language Technology purposes and the latter may not have been the intention of the web

site owner. In addition, if the LT provider profits out of this activity, this may prejudice the economic interests of the web site owner.

- In order to further reduce risk it is suggested that the LT provider:
  - only crawls sites where bots are allowed
  - has a notice publicly stating that its content only derives from web sites that do not prohibit crawling
  - provides a brief explanation as to how someone could stop her site from being crawled
  - produces a notice and take down procedure indicating under which circumstances the material will be taken down and for how long, what the decision making procedure is and an email address where relevant complaints could be addressed.
  - does not use the material for commercial purposes
  - the material provided through the LTs are in such a state or form that the original content cannot be re-constructed or its use substituted by the content provided by the LT.
- Finally, it is strongly suggested that the LT provider:
  (a) does not engage in acts of advertising the collection of web material unless necessary for the purposes of her work and only under the conditions stated in the previous bullet point
  (b) performs the processing of any collected content internally
  (c) does not offer any content or derivative content as such but only services that do not replicate the material collected but only produce a service out of its processing.

3.3.5 Grouping and understanding Cases of LR re-use

In order to better understand the ways in which we may make available LRs at the easiest and less risky possible way when they include third party material, is particularly helpful to make a typology of such material and how it could be redistributed. For this purpose, it is necessary to explore the questions raised in section C as parts of a three step process:

(a) Step A: understanding the type of material used

(b) Step B: understanding the limitations to re-use on the basis of Step A

(c) Step C: choosing the appropriate licence and way of releasing the LRs on the basis of steps A and B.

Step A: Material Used

The material used in LRs will almost invariably originate from different sources and the LR provider will not have the rights to release it without having a legal basis for its re-use. The material used may be classified in accordance to the types of rights or restrictions subsisting on it as follows:

(a) Generic Copyrighted material: this would be any type of material that is potentially under Copyright law irrespective of its source. Whether it is still copyrighted or not and whether its use within an LR constitutes a permissible use is something we have seen in detail under section 3.1 The rule of the thumb here, is that when the material is potentially copyrightable we ask three questions:

- Is it Public Domain material? (if yes, we use it, if not we proceed to the following question)
- Does its re-use fall under fair use/dealing or an exception? This is a rather rare case, as we have seen in section 3.1 (if yes, we use it; if not, we proceed to the following question)
- Is it licensed under a Creative Commons or other standard or custom open licence? If yes we use it, otherwise we only link to the material and we do not include it in a repository as such).

(b) Public Sector Information (PSI): this is material that has been produced by a Public Sector Body (PSB) and falls under the relevant PSI legislation. PSI is particular important as it comprises of large volumes of material that can be potentially re-used for the development of LRs. PSI legislation in the US exempts such material from being copyrighted in the US and licenses them under permissive and open licences (e.g. Creative Commons Zero or Attribution) outside the US. In other jurisdictions, such as Australia, Canada or New Zealand, a variety of Creative Commons licences is used in order to make such material available. In the EU, the PSI regulations and legislation have over 10 years of history and have recently been updated through the 2013 PSI Directive. According to the new PSI Directive all PSI made available has to be legally allowed to be re-used commercially or not commercially. The PSI 2013 Directive gives

the option to the Member States to choose whether to release PSI under an Open Government Licence or to exempt PSI from copyright law by law and only use a disclaimer to further disseminate the material. This practically, means that we are going to see an increase in licensing and notices in the PSI material over the course of the next two years (the implementation deadline for the new PSI Directive is the end of 2015) and hence more re-usable material for LRs. The classic limitations have to do with:

a. Attribution (of the web site, the information provider or the individual creator)
b. Non-endorsement
c. Differentiation between the original and the derivative work
d. Warranties and other disclaimers
e. Retaining copyright notices and disclaimers

Most of these conditions are easy to follow, though attribution and documentation of legal terms and conditions requires special treatment.

It is important to highlight that only because some material classifies as PSI, it does not mean that it is necessarily re-usable without conditions or only with attribution and notices conditions.

Step B: understanding re-use limitations

Re-use limitations stem from the types of rights subsisting on the material included in the LRs. Broadly speaking we may identify the following classes of cases:

(a) Re-use based on material without any copyright notices from the web: such material may be re-used after taking a series of measures to reduce risk. These include both actions before the re-use of the material and the structure of the web site/service through which the LR is to be offered:

    a. Check if there are any meta-data prohibiting crawling of the web-site
    b. Check if there are any legal conditions prohibiting re-use of the material
    c. Ensure the original material cannot be reconstructed from the LR
    d. Ensure the LR does not substitute in terms of use the original material

e. Ensure there is a notice and take down and an opt-out clause prominently featured in the web-site/service through which the LR is made available to third parties

f. Try not to use the material for commercial purposes or allow third parties to use it for commercial purposes

(b) Re-use based on material that is found on the web but which has a licence:

a. If the licence or terms of use do not allow derivatives or further distribution, we can only process the material and make it available through a service that does not give access to the original material. We have no limitations regarding annotations if the original work is not reproduced or identifiable.

b. If the licence allows derivative works, make sure to respect the conditions.

c. If the licence is an open licence:

i. Follow carefully the attribution clause

ii. Retain copyright notices and disclaimers

iii. Do not mix content with different SA licences (e.g. CCSA and OKF SA licences or CCBYSA and CCBYNCSA licences). You could, nevertheless, distribute material under different licences, if it remains separate and the licences are identifiable.

iv. Differentiate between the original and derivative

v. Follow the non-endorsement conditions

vi. Try to indicate what NC means, if the original licensor uses such conditions

(c) Re-use of material that is in print format or in digital format but constitutes a complete work (especially lexicons). Here the key question is the type of re-use made, especially if there is paraphrasing or copying of the structure of the lexicon. The following broad suggestions are made:

a. Avoid verbatim copying

b. Avoid replicating the structure of the lexicon unless there is no originality in it, e.g it is alphabetic.

c. When paraphrasing try to avoid replicating original elements in a specific definition, e.g. the order of explanations, use of characteristic words etc.

(d) Creating Derivative Works:

a. Translations, even of PD works, if they themselves are not in the PD, require special permission. See elements (a) and (b).

b. N-grams normally constitute derivative works. The more they are and the closer they get to the original work the higher the risk of copyright violation. The test followed under (a) will help in minimizing the risk from using n-grams.

c. Annotations: unless they reproduce part of the original work they do not constitute a problem. If they reproduce part of the original work, see how n-grams are treated.

(e) Anonymisation: in case the original resource contains personal data and in accordance with the specific personal data protection rules of the jurisdictions where the content is to be made available or where the data come from, there is need to either obtain permission from the data subjects or to anonymise the relevant data. The former is a rather expensive procedure, hence the latter is strongly suggested. Anonymisation, however, needs to be specified with regard to the entity that lawfully defines, the conditions for anonymisation and the permited uses after anonymisations has been completed. This is treated in greater detail in a separate section as this section only deals with Copyright related issues.

Step C: understanding licensing of LRs containing third party works

Overall, the licensing of LRs containing third party material is only possible when Steps A and B have been successfully completed, i.e. there is an understanding of the legal basis under which the material is to be used and the limitations such legal basis entails. In broad terms, we have the following scenarios with regard to the licensing options with regard to such LRs:

(a) 3rd party material is without a licence: provided that I follow the conditions under Step B(a), I can release it under any licence. The use of an NC licence is suggested in order to substantiate that no commercial harm is being done to the rightsowner of the original content.

(b) There is a licence attached to the material:

a. If the licence is a custom made licence that does not allow redistribution or transformative uses, the only possible way to use the LR is through some form of service that does not give access to the LR itself or if the original content

cannot be identified within the LR. Otherwise, it can only be linked from the LR to the original content or not disseminated at all.

b. If the licence is an open licence, standard or custom:

   i. You probably cannot relicense the original content, and license derivatives only under the conditions of the licence

   ii. Even when not additional conditions regarding the derivative work are provided, try to include the attribution, reference to the original licence and the relevant disclaimers

   iii. In the case of the derivative ensure you differentiate between the original and the derivative. Attribution to the original would in most cases suffice

   iv. Ensure you comply with the SA/copyleft conditions of the original content, i.e. do not remix the derivative work with any content under a licence other than the original.

(c) In the case the LR contains personal data, it is strongly suggested that there is a relevant indication (data protection notice) as well as a notice whether the data set has been anonymised (anonyimisation notice)

3.3.6 Conclusions

Overall, there seems to be great need for a horizontal legal intervention at the legislative level to clarify some of the key issues examined in this section particularly with regard to the re-use of material mined from the Web. IFLA's[10] suggestion or the UK Copyright data mining exception[11] is toward the right direction, which is introducing a clear exception for web crawling/data mining with the view of LR processing, with a specific clause of how to treat commercial uses (possibly through extensive collective licensing). With regard to PSI, there is also need for legislative solutions at the Member State level, making open of the PSI by default and only resorting to licensing in cases where there is no legislative tradition of exempting PSI from copyright (e.g. the UK).

[10] http://www.ifla.org/publications/ifla-statement-on-text-and-data-mining-2013

[11] http://www.ipo.gov.uk/copyright-summaryofresponses-pdf

3.4 Data Protection

In this section we present initially some key concepts of data protection, then some data protection instruments that resemble creative commons, and finally, the types of uses of data that do not require consent from the data subject. It is important to note that in the case of data protection, the consent is required not from the data processor or controller but rather from the data subject. This practically means that it needs to be sought from a different person and possibly at a different time. Various schemes that are based in the concept of consent commons try to accommodate this problem by searching for such permissions when the data controller performs the original data collection.

3.4.1 Key Data Protection Concepts

Personal Data

Personal data is any information relating to an identified or identifiable natural person (data subject): an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one of more factors specific to his physical, physiological, mental, economic, cultural or social identity (Data Protection Directive Art. 2(a))

Data Controller

The person who, either alone or jointly or in common with other persons, determines the purposes for which and the manner in which any personal data are or are to be processed.

Data Processor

Any person, other than an employee of the Data Controller who processes the data on behalf of the data controller.

Processing of Personal Data

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 (Art. 1(b) of the Data Protection Directive)

Data Subject's Consent

Any freely given specific and informed indication of his wishes by which the data subject signifies his agreement to personal data relating to him being processed.

3.4.2 Key Data Protection Principles

In accordance to Art.6(1) of the Data Protection Directive,[12] personal data should comply with the following principles:

(a) processed fairly and lawfully;

(b) collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes. Further processing of data for historical, statistical or scientific purposes shall not be considered as incompatible provided that Member States provide appropriate safeguards;

(c) adequate, relevant and not excessive in relation to the purposes for which they are collected and/or further processed;

(d) accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that data which are inaccurate or incomplete, having regard to the purposes for which they were collected or for which they are further processed, are erased or rectified;

(e) kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed. Member States shall lay down appropriate safeguards for personal data stored for longer periods for historical, statistical or scientific use.

3.4.3 Sensitive Data

These are normally defined in the relevant national legislations. The UK list of Sensitive Data (Section 2 of the Data Protection Act) is indicative of what they include:

Sensitive personal data means personal data consisting of information as to:

[12] Directive 95/46/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 **http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML**

(a) the racial or ethnic origin of the data subject,

(b) his political opinions,

(c) his religious beliefs or other beliefs of a similar nature,

(d) whether he is a member of a trade union (within the meaning of the Trade Union and Labour Relations (Consolidation) Act 1992),

(e) his physical or mental health or condition,

(f) his sexual life,

(g) the commission or alleged commission by him of any offence, or

(h) any proceedings for any offence committed or alleged to have been committed by him, the disposal of such proceedings or the sentence of any court in such proceedings.

3.4.4 Historical Evolution of Personal Data Protection and its implications for MT & MP

The terms data protection and privacy are often used interchangeably, particularly in the management literature [see for example (Kuner 2003)][13]. However, in the legal literature and in the European context in particular, the two terms are considered very closely related but not identical. Privacy could be described as "a condition or state in which a person (or collective entity) is more or less inaccessible to others, either on the spatial, psychological or informational plane" (Bygrave 2002, p.23),[14] whereas data protection is defined as "a set of measures (legal or non-legal) aimed at safeguarding persons from detriment resulting from the processing (computerized or manual) of information on them" (Bygrave 2002).[15] Data protection may be regarded a narrower concept than privacy in the sense that the former is closer to a right of "informational self-determination", whereas the latter relates to the protection of an individual's "personal space" (Kuner 2007)[16]. In this document, we focus specifically on the data protection regulation that involves computerized information processing.

[13] Kuner, Christopher. *European Data Privacy Law and Online Business*. Oxford: Oxford University Press, 2003.

[14] Bygrave, Lee A. *Data Protection Law: Approaching Its Rationale, Logic and Limits*. The Hague; London: Kluwer Law International, 2002.

[15] Ibid. p.22

[16] Supra Note 13, p.3

Three aspects of Data Protection (DP) regulation are important in relation to MT & MP processing, as being very close to the evolution of technology and the respecting regulatory responses:

- First, the processing of personal data is one of the activities most heavily influenced by changes in information and communication technologies. Different stages in the evolution of ICTs have provoked different modes of processing and communication of personal data. From the mainframes of the 1970s that gave raise to the first DP regulations, to personal computing, the Internet and then web 2.0 technologies, data protection regulation has known consecutive related changes. As we will see in sections 3 and 5, the technological changes have been in the direction of greater decentralisation of processing, storage and communication of information and accordingly the regulatory instruments used have been of similar nature.
- Second, the data protection regulatory framework is a relatively new one, the first regulation of that type appearing in Germany in the 1970s.[17] It belongs to a broader category of regulatory instruments that have as their objective to protect the weaker part in a transaction. The individual is seen as being threatened by the processing of its formation by the state or by private entities and hence needs to be protected through specific regulatory means. Though the objective of regulation is to protect the individual, the recipient of regulatory power are the entities that process data. However, as the processing of personal data becomes more pervasive, it is a question both of regulatory resources and strategy how the relevant regulatory framework is to be structured. The challenge is how DP regulation is to be transformed in such a way so that it manages to protect the individual while not restricting the free flow of information necessary for the operation of social and economic activity in contemporary societies.
- The third aspect of the DP regulatory framework is closely related to the second: the transactions related to personal data processing are highly standardised and normally regulated through formalised agreements. As in the case of Intellectual Property Rights regulations, the use of End User Licence Agreements is deployed in order to structure legally the transfer and processing of personal data from the data subject to the data processor. These EULAs have two distinctive characteristics: First, they are standardized; second, they are unilaterally defined by the strongest transacting party; and finally, while they presuppose the consent

[17] Supra Note 13.

of the individual whose data are collected and processed, this consent is problematic in great many ways. The individual does not have the opportunity to negotiate the terms of the contract or to withdraw the consent if the circumstances change.

The first data protection law in Europe appears in the 1970s in the German federal state (Land) of Hessen as a result of the increasing threats to personal data from computerised data processing (Kuner 2003, p13). About the same time the Younger Committee on Privacy proposed ten guiding principles for the use of computers that manipulated personal data that eventually lead to the publishing of a White paper, the setting up of the Lindop Committtee that provided information about the setting up of a Data Protection Authority and the compilation of Codes of Practice for different sectors of the business community (Carey et al. 2004. pp.1-3). Similar considerations have given rise to the introduction of the Council of Europe Convention of 1981 that provided in turn the impetus for the Data Protection Act 1984 constituting the first UK data protection Bill. Further to this, the Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data led to the passing of the Data Protection Act of 1998, which came into force on 1 March 2000. Finally, the Directive on the Processing of Personal Data and the Protection of Privacy in the Electronic Communications Sector (2002/58/EC) introduced in 2002 applies “to the processing of personal data in connection with the provision of publicly available electronic communication services in public communications services in the Community” that led to the introduction of the 2003 Privacy and Electronic Communications Regulations published by the UK Department of Trade and Industry. In all the aforementioned instruments the relationship between technology and the normative content of the various regulatory instruments is apparent.

From the 1970s mainframes that increased the risks of data processing by the government to the 2000s Internet-based technologies proliferating the possibilities of data processing and transfer, the data protection regulation is closely linked to the evolution and development of information and communication technologies and could be described as a form of info-regulation. It is important to note that personal data regulation, very similarly to Intellectual Property Rights regulation is a pure form of info-regulation as it is not merely technology but information and its processing that it evolves around and is seems very likely to follow the trends of development for information processing and communication.

The nature and characteristic features of information processing at different stages of their development directly influences both the content and the structure of data protection regulation. Some examples from the early data protection regulation in the UK and EU are indicative of this trend.

First, in terms of content, the White Paper introduced by the British government after the Younger Report[18] explicitly stated that: *"the time has come when those who use computers to handle personal information, however responsible they are can no longer remain the sole judges of whether their own systems adequately safeguard privacy" (paragraph 30).* In this passage it appears that the self-regulatory capacities of the relevant information processors are not deemed adequate for the required level of data processing. The White Paper justified such position on the basis of specific features of computer technology relating to improvements in (a) data maintenance and retention (b) ease of access to personal data (c) ease of data transfer (d) combination of data that would not be otherwise possible (e) storage, processing and transmission of data in ways not possible in the past. The technologies in place where these reports and white papers were issued describe a form of data processing that is to be primarily conducted by formal organisations of significant size, so that they are able to cover the costs of information computerisation. Such organisations would be the government itself and any commercial organisation of substantial size.

The fact that the government was deemed as one of the primary sources of risk for personal data protection is not only a result of its historical role but also of the nature of the information systems introduced to governmental agencies at the time. This is expressive in the choice of introducing independent DP authorities entrusted with the responsibility of personal data regulation rather than making that a function of the state. Also the fact that the use of Codes of Practice as one of the first measures to be adopted for the regulation of personal data is indicative of the ability, at least in principle, of the commercial sector to deal with this issue. At this stage the lay person is only the passive subject of data processing, something being in total accordance with the information technologies available at the time.

[18] http://www.jstor.org/discover/10.2307/1093890?uid=3738128&uid=2&uid=4&sid=21103789521363

Finally, the interest in introducing DP regulation at a Europe-wide level is also indicative of the importance of the free flow of data within the framework of the then European Community and the appreciation of such flow as one of the elements contributing to the creation of the single European market. The market as an implied regulatory modality appears thus as one of the silent regulatory forces that has even partially formed the way in which the EU data protection regulation was to be further developed: to create an essentially single regulatory framework for all European States that belong to the Community then and the Union later so that the seamless participation to economic activities was possible.

As technologies of processing and reproduction of digital content became cheaper, the regulation changed in order to accommodate the proliferation of entities holding data (data users according to the UK Data Protection Act of 1984). These entities required to register with the Data Protection Authority. As we move to the UK Data Protection Act of 1998, we see that there are two broad trends: the individual rights awarded to the person whose data are processed (data subject) are increased, the scope of what constitutes data processing expands (including manual processing as well) and there is more emphasis on data transfers and the regulation of data exports.

The maturing of the European common market and the phenomenon of information processing outsourcing that started growing in the 1990s has had a direct impact on the regulation of personal data exports and processing. The introduction of personal computing in the 1980s and the Internet in the 1990s has fundamentally changed the organisational structure of the multinational corporation and made it possible to process data in places different from those of offering the services. Different forms of “sourcing”, from outsourcing to off-shoring have entailed the introduction of geographically dispersed corporate structures and accordingly influenced the way and locus of employees’ personal data processing. This becomes extremely important in the case of LR-based MT & MP where the re-use of data is necessary and hence all types of personal data processing becomes instrumental for the operation of such projects.

These developments have impacted the personal data regulation accordingly not only in terms of their actual normative content, which remained to a great extend the same following the original data protection principles, as it has influenced the structure and organisation of regulatory instruments. Decentralised information processing has led to decentralised and

volatile organisational forms that required more flexible regulatory means. These trends are expressed in the following forms of data protection regulation, that is, the extensive use of self-regulation, safe-harbor provisions, technical standards and contracts.

The US safe-harbor system allows personal data to be transferred under a presumption of adequacy to US-based companies that agree to be bound by the system. Interestingly, the safe-harbor basis may be found in a variety of documents ranging from the European Commission decision adopting the safe harbor system to a set of Frequently Asked Questions and the safe harbor principles. The US safe harbor system is characterized by a series of features that make it a very interesting regulatory species as it provides a standard to be followed by the companies deciding to follow the system rather than a comprehensive set of a priori rules. The companies interested in following the system must ascertain whether they are eligible for participation to the safe-harbor scheme. Secondly, the companies must determine which dispute resolution and enforcement system they want to be subject to in relation to their safe harbor system participation. As Kuner notes *"This can broadly be either (1) a private, self-regulatory mechanism, such as membership in a self-regulatory group, or development of the company's own privacy policies that comply with the safe harbor principles, or (2) a 'legal' or government mechanism, such when the company is subject to 'statutory, regulatory, administrative or other body of law (or of rules) that effectively protects personal privacy', or if it commits to work with the EU DPAs. The company must also send a written confirmation to the US department of Commerce (which can be done online) signed by a corporate officer stating its commitment to the safe harbor principles. Finally, the company must disclose its commitment to the safe harbor principles, such as by stating in its publicly available privacy policies that it participates in the safe harbor arrangement"* (Kuner 2007, pp. 139-140).

Another example of the changing nature of regulation is art. 26(2) of the Data Protection Directive 95/46/EC, which stipulates that transfer to countries outside the EU may be authorised when there are safeguards that may "result from appropriate contractual clauses". The interesting aspect is not merely that the transfer of data is possible under contractual arrangements, but also that these may appear in two forms: (a) Model contracts, which constitute standardised sets of clauses approved by the European Commission and (b) Ad Hoc contracts, that have to be approved by or notified to the PDA of the Member State from which data are to be transferred. The choice of contractual arrangements that are to be

approved either by the European Commission or individual PDAs is an indication of a model of regulation that adopts both the centralised model of the independent authority and the more flexible model of contractual arrangements.

The advent of the Internet has introduced even greater possibilities of collecting data and hence the possibilities of personal data violations. In that sense there is greater emphasis in the security element. Need to control the flows of data and of course more players that need now to be aware of their capacity of data processing and hence need to comply to personal data regulations. At the same time there is more ability to the individual to monitor and in principle control the use of her data. In that sense there is need to push more control over the individual. The individual seems to need to have more control over the way in which her data are used. This is expressed particularly in the various forms of consent required for the processing of data, especially as required in the Directive 2002/58/EC on Privacy and Electronic Communications.

Web 2.0 and 3.0 mark the introduction of a huge impetus for sharing, re-using and processing each other's personal data. What does this mean in practice for data protection is that the concept of personal data protection and consent need to be radically reconceptualised. Each person needs to be made aware of the value of her personal data, where they are stored, how they are processed and how the regulation should deal with this problem. The monitoring costs seem to be tremendous and the concept of consent seems to devolve in a world where anyone seems to be both processing and giving away personal data. In this context we move from a need for self-regulation in the industry to need for mass-micro auto-regulation. In such context the need for the individual to appreciate the importance of her personal data and be able to regulate it accordingly is crucial. For that purpose it is not any more enough just to provide such means to the individual but also to make sure that the individual is equipped with the knowledge to actually manage her personal data accordingly.

It is important to note that the different phases of development are not linear and mutually exclusive. Different sources of threat and accordingly different regulatory measures need to co-exist and co-develop and this is reflected in the co-existence of different regulatory means.

3.4.5 Standard Licensing Models for Personal Data

In this subsection we present a number of projects that get inspired by the Creative Commons idea in order to support the re-use of personal data without requiring any additional permissions.

3.4.5.1 Consent Commons

The Consent Commons project takes the idea of Creative Commons that there is a permission by the licensor on the work that allows it to be used by everyone under specific terms and applies it to the case of data protection. It thus assumes that at the time of obtaining copyright permissions, the time of the Intellectual Property Rights (IPR) clearance, there will also have to be clearance of the privacy rights. The interesting part here is that these permissions will have to be built upon the CC licences. This means that the data subject will have to give permission for her data to be re-used as the chosen CC licences prescribe. The consent may cover only personal data but also tissue and in that sense it will have to be compliant with data protection law, the human tissue legislation and any other confidentiality agreements that may be in place.

The Consent Commons idea originates in the area of Open Educational Resources (OER) and was funded by the UK Joint Information Systems Committee. It has as its main objective the re-use of medical educational material and specifically clinical images and could potentially also include research.

There is a number of issues associated with the Consent Commons project:

- the level of awareness of whether and how consent should be acquired is considerably low. Very frequently the data protection law does not really require consent and if the medical practitioners could avoid the effort to obtain consent, they probably would not use the Consent Commons approach.
- The transition from providers of clinical services to the Higher Education (HE) sector is not always without friction. A patient may be happy for her data to be used in the context of a treatment or even research, but it is not clear whether this would also be desirable in the case of educational purposes, especially if the material is to be openly re-used without an obligation to notify the data subject for the use of her medical image. These are not just personal but sensitive data and the consent is really essential.

- There may be ownership of the IPR without consent clearance or the other way around. This means that if the content is free to move there need to be procedures that ensure that either the permissions are in place or that there is a clear way to get back to the IPR owner and the data subject to obtain the necessary permissions.
- There are tracking issues with regard to how the content is to be used. Even when the permission is given by the data-subject it will be necessary to make sure that an as broad consent as one that allows re-use is legal and that there is some tracking of how and where the content is being used. This may lead to a need to retract permissions, which is not necessarily in tune with the current CC licences or the open content ethos and practice.
- There is no clear guidance as to how consent is to be obtained or clear policies as to how to do this without violating both the legal framework (data protection and human tissue act) as well as the codes of conduct and professional rules and codes of ethics in different contexts.
- There are no clear policies by the providers of clinical services or the people in medical education as to how consent is to be obtained and the licensed and consented content is to be further re-used. This is reflected in the lack of process and legal instruments that could further support their use.

Consent Commons come to give a solution in the following respects:

- they complement the CC licences covering the data protection side of things
- they provide a set of principles that could then be used to build processes and tools upon them
- they also provide a three layer structure as the CC licences
- they could potentially provide revocation of consent though it is questionable whether this would match the CC structure and philosophy
- it has different levels of release of personal data corresponding to different types of consent:
    - fully open access
    - sharing between trusted partners
    - open but the specific users have to be approved

- restricted access

- some of these modes of Consent Commons are CC compliant and some others are not.

3.4.5.2 Privacy Commons

Privacy Commons is an umbrella project aiming at hosting different projects with the objective of tackling the following problems:

- the existing privacy policies by different data controllers are not clear and understood by the end-users
- the privacy policies are not machine readable so that they could be easily and quickly read, compared, tracked and understood
- overall, the privacy policies are data processor rather than data-subject centric

The solution proposed by the Privacy Commons project is to standardise and modularise privacy policies, so that they may then be easily expressed in icons and become machine readable. The Privacy Commons project is implemented by different companies that wish to adhere to such principles. An example of the icons used would include the following basic elements:

- notifying the data-subject that her personal data are collected
- if there is a specific type of information that is collected, this also has to be made clear (e.g. banking information)
- if aggregate statistics are collected, this is also made clear to the data-subject
- another very important element is whether the company will or will not disclose the data to third parties; in the current implementation of the Privacy Commons project it is not clear how the data are further disseminated or to whom
- a separate icon signifies whether the data are to be sold or not; again, the entities to which they are to be sold are not made explicit
- finally, another key element of the project is to indicate whether there is user, corporate or shared ownership over the data. This may relate to the way in which IPR are handled but also how the personal data are to be shared. The interesting part here is that we see a transition from personal data to property rights. Even though it is not clear

whether the individual has ownership over her personal data or not -and in most legal system it does not have such ownership- it is being informed as to the property status of such data. This will allow the individual to make an informed decision as to whether she would like her data to be processed or not.

3.4.5.3 Privacy Icons

Privacy Icons is a project specifically aimed at addressing the problem of privacy policies lacking clarity and not focusing on what the end users wants. The objective of the project is to produce icons that allow the data subject to understand how her personal data are to be processed and further disseminated. The privacy commons project is primarily addressed to organisations that wish to improve the communication of their data processing principles. These organisations are to adopt range of icons that represent the basic features of their privacy policies. Such features may include:

- whether the data are to be used only for a specific purpose or for all possible purposes, including others than those originally intended
- whether the data will be bartered or sold
- whether the data will be passed to advertisers or not
- how long the data are to be kept (ranging from one month to indefinitely or only for the duration that is necessary for the intended processing to take place)
- whether the data will be given to law enforcement only when a legal process is followed or when they are given to law enforcement irrespectively of any process followed.

Two basic issues related with the Privacy Icons project are the following:

- how will the project fit all possible privacy policies: the answer to this is that by following a “lego” like approach, the data controller is able to express its policy in an accurate and at the same time understandable fashion
- the problem of the “evil” icons, that is, why would a data controller adopt icons that would show that it processes data in a way that may be contrary to the interests of the data subject. Precisely because no data controller would ever accept that, it is assumed that a privacy icons model could identify and rate privacy policies. If a policy does not accept to be rated, then it would be assigned the worse possible values by

default. Of course, this is an approach that could be both misleading and not feasible and hence it is least likely to be adopted.

3.4.5.4 Identity Commons

The Identity Commons is an umbrella project including a number of smaller projects all aiming at creating a user-centric, identity infrastructure and to address the resulting social trust issues. Though it does not directly deal with privacy issues, it is a project that has a great impact on the protection of personal data and includes projects having a specific focus on privacy. The Identity Commons projects include:

- Data portability
- Higgins
- ID-Legal
- Identity Gang
- Information Card Foundation
- Internet Identity Workshop
- Kids Online
- OpenID Foundation
- OSIS
- Pamela
- Project VRM.

3.4.5.5 Conclusions

All these projects follow the same basic principles, which are:

- Self-organization
- Transparency
- Inclusion
- Empowerment
- Collaboration

- Openness
- Dogfooding

Summing up the four aforementioned projects they have the following characteristics:

- they are user centric
- they contain mechanisms both for protecting personal data and identity
- the have also an organisational perspective focusing also on the implementation of the project
- they all use techno-legal mechanisms such as meta-data, tagging and APIs
- they all assume that the content will flow over the public internet

3.4.6 Legal Basis for the Use of Personal Data

Personal Data could be processed on the following premises:

(a) that free, prior and informed consent has been obtained for a specific re-use purpose and for a specific duration

(b) that there is another legal basis for the processing

Consent may be obtained either through a standard release statement, such as those presented in section 3.2.2, which resemble the Creative Commons mechanism or by asking for a specific permission for the processing of Personal Data. In the latter case, the request for such permission should state the purpose and nature of the processing, its duration, the data retention period and the degree to which personal data are to be further disseminated either as such or through a value added service.

The most common legal basis for LR-based MT & MP will be that of research. The conditions for such processing are:

- that it is only for research purposes
- that the data is anonymised
- and that all necessary measures for the protection of people involved are taken. This last condition is to be interpreted in different ways according to the case law in the

relevant country, however, it should definitely include certain measures to prevent further dissemination of the personal data.

The problem with such definition is that a large amount of LR based research may end up into commercial purposes that go beyond the original research purpose. It needs to be noted that this is not necessarily a problem, if the commercial service does not make use of the personal data or does not further release them.

Even in cases where sensitive data are to be used, processing is possible provided that:

- access is only provided at the site where the data are stored
- only data that are necessary for the research purpose are extracted
- do not record data that may be used to identify living persons
- retain the data anonymised for subsequent uses (e.g. publications)

The problem with such conditions is that they are very difficult to be implemented in an MP/MT scenario. For this reason, and despite the obvious problems with such an approach, it is always suggested to ensure that the consent of the data subject is obtained when sensitive personal data are involved.

3.5 Public Sector Information

Public Sector Information is of particular relevance in the MP/MT context, particularly since a large amount of the content used for such purposes is Public Sector Information.

As **Public Sector Information** is deemed any form of document held by Public Sector Bodies (PSBs) of Member States (Art. 1(1) of the PSI 2013 Directive (2013/37/EU)).

According to Art.2(3) of the PSI 2013 Directive:

"document" means:

(a) any content whatever its medium (written on paper or stored in electronic form or as a sound, visual or audiovisual recording);

(b) any part of such content;

As a result, a vast amount of information released publicly on the Internet by different EU Member States is within the scope of the PSI Directive and as such it is eligible for re-use by third parties both for commercial and non-commercial purposes.

While the discussion regarding disclosure and making available of PSI is dominated by the question of PSI licensing, which is covered in section 4, it is also likely that no licence is either necessary or possible to be used. This section covers three main cases:

(a) PSI exempted from copyright protection

(b) PSI on Public Domain works (copyright has expired)

3.5.1 Works not granted copyright or exempted from copyright and similar rights protection

A licence will not be necessary in the absence of copyright or similar rights (including neighbouring rights, the sui generis protection of databases or other similar rights) for that specific subject matter.

There are two main cases of PSI as copyright exempted subject matter:

3.5.1.1 PSI as exempted subject matter per se

Some jurisdictions do not assert copyright in PSI, so no license would be needed. The clarity of the regime varies in different jurisdictions: In some cases, this is clearly stated in the law, in some others this is information that could be provided by the National Copyright or Intellectual Property Office and in some others it is less than clear what the PSI status is and it has to be established on a case-by-case basis.

In such cases, the relevant national copyright law will make explicit reference to the types of works and uses of such works that are non-copyrightable subject matter. These will in most jurisdictions be text used to exercise the administrative powers or to offer public service. There are explicit references to judicial, legal and administrative texts, but the limits of the exemptions are to be defined on a national jurisdiction basis.

Content Provider's Perspective: The content provider should at least use notices to indicate the copyright status of the work as Public Domain. The Creative Commons Public Domain Mark (PDM) could greatly contribute towards such directive.

User's Perspective: The user could either make an assessment with regard to the copyright status of the work or seek for relevant notices. It is suggested that the user uses one of the Public Domain Calculators that are available to assist her in making the relevant decisions. If the work is not copyrighted, the user may use in any way she wishes. Because legislative material is very likely to fall under this category in most civil law jurisdictions, it is important to assess the status of the work at the country of origin.

Tip: It could well be the case that the individual works are copyright-free, but their compilation is protected under copyright or the sui-generis database right. This is the case with most commercial legal databases. Hence, always pay attention to the source of the material: it is more likely to be Public Domain if it comes from an official or public web site.

3.5.1.2 Non-copyrightable subject matter released as PSI

The second case is different from the first case, as it is the nature of the PSI material rather than its use as PSI that renders it non-copyrightable.

These are mostly cases where the subject matter does not fulfil the requirements of originality and form to attract copyright protection or specific types of information, such as factual information, raw data or traditional knowledge. Such types of content constitute non-protectable subject matter and hence do not require any type of licensing, irrespectively of whether they are PSI or not.

However, it needs to be highlighted that once arranged in a systematic or methodical way, the resulting set of information may be protected under copyright (if the arrangement passes the criterion of originality) or the sui generis database right (if they constitute non-original compilations).

Content Provider's Perspective: The content provider should identify such material and convey its legal status to the end user through some form of notice. However, because of the existence of the sui generis database right, which provides even mere information with a shell of a property right, it is in very few cases that non-copyrightable subject matter will be released as such. An interesting case in the scope of the PSI 2013 Directive will be the release of traditional knowledge material from museums, archives or libraries, including oral history and songs, which may be of great interest for MP purposes, however, even such content may attract copyright through its packaging (sound recordings, transcriptions etc.).

User's Perspective: The user will have to assess whether the material is not copyrightable or not. This is often a difficult and ambiguous exercise and is –generally- suggested either to seek for some sort of notice or use the risk mitigation techniques presented in section 3.7.

3.5.2 Expiration of the Copyright Term (Public Domain Works)

A licence will also not be necessary when the copyright or similar rights term has expired. Works no longer protected by copyright because of the expiration of the economic rights term should be treated as public domain works and therefore should be freely re-used. The economic rights granted under the copyright regime typically expire 70 years after the death of the last co-author or 70 years after the publishing or recording, but rules may vary and the term of protection may be greater in special cases. Public Domain calculators are being developed to help assessing whether a work will be in the Public Domain in a particular jurisdiction.

Content Provider's Perspective: A good practice here is not to license content that is in the public domain. The Content Providers often do not have a clear understanding of the copyright status of the information they release. It is strongly suggested to make an assessment of the duration of the information to be released and mark the works accordingly.

User's Perspective: The lack of relevant documentation and harmonization in the term of protection of different works in different jurisdictions is likely to cause significant implementation issues with regard to the assessment of whether a particular work belongs to the public domain or not (due to term expiration reasons). The Public Domain Calculators may be useful with regard to an initial assessment of the copyright status of the work, but a risk mitigation strategy should always be applied.

Tip: The risk of infringement is reduced as we get closer to the term expiration, the work is of low commercial value and the use is non-commercial.

3.5.3 Limitations and Exceptions

A different, but related case, is when the PSB needs to use copyright material to perform its public task or where a court requires to have access to copyrighted subject matter in order to issue a decision. These are cases, where no permission or input licence is required for the PSB or court of justice to perform its mission or task, as it will normally fall under the limitations

and exceptions, fair dealing or fair use rule and could hence be used without any additional permissions.

Content Provider's Perspective: When such a material is to be disclosed or made available for re-use this cannot be done if it contains third party copyrighted material. While the exception will cover the use of the PSI, it will not necessarily cover its re-use. This is the reason why it is strongly suggested that PSBs mark PSI containing third party material with some sort of meta-data or notice regarding the third party material.

User's Perspective: The lack of a harmonized copyright limitations and exceptions regime across the EU has as a net effect an increasing uncertainty as to what falls within their scope. It is not clear whether the material, its use or the entity that performs it are such that they are considered as falling within the limitations and exceptions. The disparity between the fair use, fair doctrine and limitations and exceptions systems, further complicate the situation, making the request for a licence a safer option.

3.5.4 Use of Marks and Notices

It is highly recommended that, when PSI material is not covered by copyright or other similar rights or when it contains third party copyrighted material, the relevant marking is in place. This will increase legal certainty and allow the lowering of transaction costs.

This can be achieved in a standardized fashion by using the Creative Commons Public Domain Mark or by drafting an ad hoc notice.

Content Provider's Perspective: Using a standardised tool such as the Public Domain Mark developed by Creative Commons provides the text in a language that is accompanied by metadata, valid across jurisdictions and translated in many languages. According to Creative Commons, the Public Domain Mark "is intended for use with old works that are free of copyright restrictions around the world, or works that have been affirmatively placed in the worldwide public domain prior to the expiration of copyright by the rights' holder." The Public Domain Mark tool provides the ability to generate HTML code to inform the public (and search engines) of the public domain status of the work. The Public Domain Mark enables a person who wishes to mark the work as being in the public domain to include optional useful information, such as:

- Name of the work, e.g. title of the dataset;
- Name and URL of the author, e.g. the division or department releasing the PSI and the source page;
- Identifying individual or organisation, in case this information differs of the above, e.g. a higher level of the PSB which should be contacted for further information.

User's Perspective: It is always preferable to search the material through search engines that allow the identification of the relevant licensing form or copyright status of the material or use relevant APIs or other technical means.

3.5.5 Licensing of Public Domain material released as PSI?

PSBs should refrain from using licences for PSI, which is in the Public Domain. Such licences would create restrictions upon the use of works that are no longer protected by copyright or similar rights and can be freely used without any conditions.

Furthermore, since no copyright exists in a Public Domain work, there is no legal basis to license it. The PSI2013 Directive explicitly makes reference to the possibility of releasing material without any conditions, and the case of Public Domain material clearly falls under such case.

In addition, it is not recommended to add a licence (and therefore restrictions where none should apply) to the digitised reproductions of analogue non-copyrightable data or Public Domain works. The mere act of digitisation is not a source of new rights and keeping digitised versions in the Public Domain will guarantee they remain free to use as the original work. Digital reproductions of works which are in the Public Domain must also belong to the Public Domain. Use of Public Domain works must not be limited by the addition of unnecessary licensing requirements. In some countries, the threshold for originality is low, and digitisation might open a claim to copyright, but it is not recommended to enforce that right.

Content Provider's Perspective: Refrain from using any licence for PD material; instead use notices where applicable.

User's Perspective: Ensure that the material used is indeed in the PD irrespective of the licensing scheme. Check if there is any additional form that may revert the resource to

copyrighted material, e.g. book format protection, database right or digitization (depending on the jurisdiction).

3.5.6 Concluding Remarks and Recommendations for PSBs licensing

Overall, the material released by PSBs as PSI may be used in a number of occasions for MT & MP purposes without requiring additional permissions or even a licence, either because of it belonging to the Public Domain or because it falls within relevant limitations and exceptions. When a licence is required, the normal copyright rules should apply as stipulated in section 3.1, where reference to PSI and the related licensing is also made.

Directive 2013/37/EU (the New PSI Directive), and previously Directive 2003/98/EC, allows for the release of PSI for re-use under a licence or without conditions (art. 8). This practically means that a Member State may choose to release PSI for re-use without a licence if this PSI is:

(a) in the Public Domain (e.g. because the duration of the copyright has expired)

(b) is exempt from Copyright law.

The experience of the open licensing community, even outside the realms of public data regulation, favors maximum simplicity in the release of public data. Such simplicity is best served when any type of work is made reusable without any limitation, or with very few limitations. This helps to ensure licence compatibility and increases the re-use of the content by the industry and the civil society. In turn, this best serves the objectives of the Directive, i.e. growth/job creation and the objective of the Digital Agenda 2020 for greater transparency in the activity of the Public Administration.

In the open licensing community this is amply demonstrated by the recent statement issued by Creative Commons after the 2013 Global Summit, effectively supporting the development of positive user rights in copyright law, rather than relying on list-based exceptions or open licensing as sufficient solutions. The need for copyright reform is beyond the scope of this report. However, the fact that there is an urgent need for reducing uncertainty and complexity with regard to copyright limitations and exceptions--and that licensing is a patch rather than a fix to the problem--points at the direction of a legislative solution at the Member State level, something that the new PSI Directive makes possible and something that is followed by a number of Member States. This approach combined with an “openness by default” policy may allow the maximum benefits from opening PSI while substantially reducing transaction and clearance costs for potential re-users.

The following recommendations could strongly support the use of PSI-based LRs for MP/MT purposes:

Recommendation 1

(a) The adoption of a legislative solution, instead of licensing, could further reduce transaction costs, i.e. openness by default of PSI in the form of a *law* without further requirements of issuing or adopting of a specific licence:

- introduce a positive, actionable user-creators' right to PSI;
- harmonise exemptions and exceptions to any kind of protection so that such right applies across the European single market,
- in order to ensure interoperability make sure that no conditions are attached to the re-use right other than the ones mentioned in the Directive, i.e. acknowledgment of source and acknowledgment of whether the document has been modified by the re-user in any way;
- provide clear report as to how the positive PSI right operates; these reports could be provided by the competent for the implementation of the PSI directive Public Sector Bodies through circulars and then to the re-users through notices.
- make a registry for the information that is to be used under a closed/all rights reserved/re-use but not open licensing scheme. Ensure that such registry is regularly updated and provided as (linked) open data in the national data.gov (open data) portal. This should not introduce additional costs to the PSB, as it could be a simple spreadsheet file (e.g. in csv format), it would be necessary if the PSBs were to charge anyhow, and it could make use of the already existing infrastructure (data.gov site).
- clearly indicate when personal data are included in PSI, ensure that any re-users are obliged to also use a personal data notice (indicating the original processing purpose) when such data still exist and provide report as to how to resolve the conflict between personal data protection and the right to re-use;
- have a single institution responsible for collecting, coordinating and administering open licensing, even where it evolves subject matter outside the scope of PSI (e.g. research data, broadcasters' data etc). Ensure that there is harmonisation of open access policies in PSI, the cultural, educational, research and science sectors; This could be done at the Member State level through an intra-ministerial committee headed by the

PSB responsible for the implementation of the PSI Directive. EU-wide coordination could be done through the introduction of a Working Group or standing committee to support the review of Art. 13(2) of the PSI 2013 Directive.

- issue technical report ensuring the legal documentation of the relevant data sets (e.g. under which regime/law are made available).

  Use report for the following issues:

  - regarding notices to be placed on PSI. This could be done through *smart notices*, i.e. not legaly binding notices that refer to specific legislative provisions, that are permanently stored in the same way as the Public Domain Mark, and added as meta-data or mark to the relevant PSI. In fact, if PSI is equaled to Public Domain, then the CC PDM could be used, as in the case of Europeana.
  - how to mark that the PSI contains personal data and how to resolve the conflict between personal data protection and PSI re-use.
  - on how notices are to be retained by the re-users
  - on how to obtain consent by the data subject for re-use of her information at the point of collection
  - on obtaining maximum IPR from third parties

(b) Use the licensing instrument with the least legal friction, i.e. the CC0 Public Domain Dedication (https://creativecommons.org/publicdomain/zero/1.0/)

(c) Use standard open public licences, especially CC BY, which is the solution with the least legal friction if recommendations (a) and (b) cannot be followed.

National governments and public sector bodies (including Galleries, Libraries, Archives and Museums, as well as other public interest institutions) still prefer to use their own Open Government Licences (OGL), as this gives them more control over the wording of the licence and the licence update process. However, this poses increasing challenges as it requires a dedicated team of experts for the creation and maintenance of the licence as well as continuous monitoring of updates of other standard public licences and extra care in the wording so that interoperability between different licences of the same type is achieved. While this is possible, it is extremely difficult to achieve for all six types of Creative Commons licences.

Hence, it is advisable, that if governments or public sector bodies insist in making their own licence, to only create an Attribution (by this we mean “Attribution-only”) licence.

Recommendation 2

Governments are strongly advised NOT to use:

- NoDerivative licences, since they substantially erode and effectively annule the scope and ambit of the re-use of the material and hence the application of the PSI directive.
- NonCommercial licences, since it is extremely difficult to achieve a common definition of NonCommercial both within their own jurisdiction and--even worse--in other jurisdictions
- ShareAlike licences, since they require that the derivative work will be released under the exact same licence, which make legal interoperability extremely difficult.

It is also important to note that the only types of licences that conform to the Open Definition (http://opendefinition.org/okd/), i.e. a definition of how all types of knowledge could be disseminated and re-used with the minimum possible restrictions are the following:

- CC0 Public Domain Dedication or Public Domain Dedication and License (i.e. blanket waivers when construed as a licence)
- Attribution licences
- Attribution ShareAlike licences

However, as mentioned above, ShareAlike licences may be very problematic in terms of compatibility as they require that data-sets are remixed using licences with the same terms and conditions. In addition, CC Zero may be the preferred tool for releasing public data since it is the preferred tool for Europeana and produces the least possible interoperability frictions.

It is strongly advised that the European Commission does not create its own Open Data licences. The reason is that another licence would only add to the problem of “licence proliferation”, i.e. the problem of having multiple licences with similar terms but with potential incompatibilities that do not allow the seamless re-use of different data sets. For this reason, the suggested direction is towards a European licence standard rather than a standard European licence.

Recommendation 3

It is strongly advised that even when governments choose to produce their own version of an Attribution (or other open) licence, they should always try to apply it in the following fashion:

- retain a clear versioning/date scheme, i.e. each licence should have a specific version and a date of that version. When changes are made, these should be made public and the version number should change.
- store the licences at a permanent location and link the licensed material with the URI of the licence

3.6 Confidential Information and Other Rights/Agreements

MT & MP of LRs may be hindered if the content that is used falls under the category of trade secret or confidential information. Such could be the case particularly with regard to technical information, manuals or technical descriptions. While it could be that such information is not marked as such, it is highly unlikely that processing of such material could happen without prior knowledge of the nature of the material by the entity conducting the MP/MT. It is hence highly advisable that the MP/MT service provider obtains the relevant permission before it conducts any processing or indeed dissemination of the material.

Other types of rights that could subsist in the content to be processed, such as cultural protection or geodata regulation is highly unlikely to influence any of the processing options, as these types of regulations normally involve material that is not relevant to MP/MT (e.g. maps or digital representations of monuments).

3.7 Risk Mitigation Strategies and Measures

The aforementioned analysis is illustrative of the uncertainties, costs and, hence, risks related to the collection and re-use of LRs for MP/MT purposes. At the same time, a closer look at the patterns of LRs use indicates that there is a possibility to substantially reduce legal risks by adopting certain strategies that allow the provision of the intended services without exposing the LR processor to substantial legal risk.

Such strategies and measures are based on three premises:

(a) First, that the essence of LR processing does not relate to the processing of a resource, whether textual or audiovisual, as something that is to be consumed by a human-being but rather as data to be processed by a machine. As a result, it is very likely that the original resource will not necessarily be made available to the audience or end-user as such; instead the resource will be modified, often into something that is either unrecognisable or not even copyrightable, and will feed into a service that will not leave traces of the original resource when offered to the end-user.

(b) Second, while copyright infringement covers both introvert (e.g. copying or modifying) and extrovert (e.g. publishing, making available, distributing) acts, the latter (rather than the former) are a greater source of legal risk.

(c) Third, in cases of personal data processing, the link between a specific person and the relevant data is not what is valuable for LR purposes. It is not the information *per se,* but rather language elements that make it a useful resource for MP/MT. As a result, it is possible to provide services without requiring to provide the personal data as such.

The three main risk mitigation strategies may be summarized as follows:

(a) Provide the service rather than the data: as mentioned above, the key objective of any LR-based MP/MT is the provision of a service rather than the data as such. Even in terms of the content processing, what is of importance are structural elements, series of words in n-words, relationships and syntactic elements that provide semantic information etc. Such a strategy is mostly useful with regard to copyrighted content, though it may also useful with regard to content containing personal data. Overall, a service provision strategy may lead to two results:

   a. Either reduce the likelihood of infringement by focusing on mere acts of copying and/or modification rather than distribution, since what is produced is so far away from the original work that it does not classify as a derivative work at all; or

   b. Reduce the risk of legal action by providing access to the service only rather than the content as well, the use of which remains not visible for the wide public.

(b) Anonymise or pseudonymise personal data that are then to be released as MP/MT data: While the area of anonymisation and pseudonymisation is vast and often contested, it is also an area that has great potential for obtaining access to personal data for research purposes or for allowing the release of a service without providing

access to such personal data. It is not the objective of this section to explore the problems of anonymization, especially with regard to the question of whether they remove any personal data from a data-set or piece of information and to what extent such mechanisms truly protect the data subject. It is rather to highlight the elements that reduce the legal risks. These are:

a. Imposing anonymisation obligations as an access condition

b. Imposing anonymisation obligations as a release condition

In both cases the objective is to extract the maximum value from the processing of the data, while maintaining its value.

(c) Shuffling or scrambling data: Objective of this strategy is to reduce access to the original information as such, i.e. as information that is addressed to a human reader. This is particularly relevant in cases where the value of MT & MP is not dependent upon the complete sequence of words and sentences or the value of the LR is not dependent upon its human-readability. Such a strategy is helpful in order to prevent canibalising the market of e.g. a work of literature when it is distributed as part of a corpus and is either part of the language processing copyright exceptions (e.g. in the UK) or the fair use doctrine (e.g. in the US). What is important to note here is that such an approach would only have an effect in jurisdictions where the lack of competition with the original work is deemed as a reason for allowing the use of the work without requesting permissions from the rights holder. In other cases, it is less likely to pass the test of the normal exploitation of the work, since even the use of e.g. the syntactic elements of a text could be considered a source of potential income for the rights-owner.

An interesting application of the aforementioned premises and strategies is that of the MetaShare NoRedistribution (NoRed) licences.[19] While this is a mechanism that is very different from the original purpose of the risk mitigation strategies, because it is the expressed will of the content provider rather than a way in which the re-user of the content could use it without asking for a permission of the rights-holder, it is expressive of a broader approach regarding the use of content for MP/MT purposes. More specifically, it is indicative of the type of re-use that is accepted by the content providers even when they have the option to restrict (or open the content) as much as they desire. Under such licences the recipient of the licence

[19] http://www.meta-net.eu/meta-share/licenses

may use the material and –in some cases- even to make derivative works. However, it is not possible to further disseminate the work in its current form. These are the cases, where the content provider is not really interested in restricting the MP/MT market, but rather the emergence of products competing with the content itself. This realization of the existence of two different markets or two different classes of use value sets allows us to revisit the strategies we suggested and provide a final generic rule as to how the resources could be used. That is, that the content should be used in such a way that it does not affect the original market of the work. The means for achieving that would be to ensure that it is not available as such to the end user, either through the provision of a service than the content or it anonymisation or its distortion (scrambling).

4 Concluding Remarks

What has been clearly shown in this report is that LR processing for purposes of MT & MP is a complex issue that requires a number of permissions to be obtained in order to occur at the lowest possible transaction cost. At this stage the following options are in place:

- Unrestricted use on the basis of copyright and data protection exceptions: this is the best possible solution, as it incurs the lowest possible transaction costs. However, the operation of such a regime is far from being a reality. The legal uncertainties and inconsistencies are such that render most of such exceptions practically inapplicable.
- Use of standard licensing regimes allowing open access, such as the one provided through the Creative Commons and various Privacy Commons projects. This is a solution that provides legal certainty, but is dependent upon voluntary action and is, hence, fragmented and not uniform in its application.
- Requesting ad hoc licences: this is the legally safest solution but has the greatest transaction cost and is not certain that it will lead to the identification of the rights-holders or the obtaining of the necessary permissions.

Overall this report proposes to take a clear and unambiguous position towards the introduction of positive user rights

(a) at a global level, i.e. through WIPO,

(b) at a regional level, i.e. through and EU Directive and

(c) at a national level by supporting positive user rights policies, where these exist (e.g. Public Sector Information rights).

Using open or extended collective licences has been a key element of a more balanced copyright policy for the past ten years. However, it is high time to move forward seeking change at the legislative rather than the contractual level. Open licensing has been tremendously successful in terms of making the business case for open data and content, but much less successful in terms of reducing clearance costs and uncertainty inherent in a copyright system that requires no formalities or registration for protection and that still relies in the EU context on an exclusive set of limitations and exceptions rather than a fair use doctrine.

What ten years of experience with open content licences has amply demonstrated is that users are no more passive recipients of content but rather active creators, especially in the context of language resources, and, hence, they require positive rights rather than mere exceptions, particularly when such exceptions are list- and not doctrine- based.

The re-use of LRs is not something that may be easily solved through licensing, since it poses an excessive clearance and licence choice cost upon the Language Technologies providers, it does not solve the problem of orphan works and does not take into consideration the issue of works belonging to rights-holder other than the LR providers.

Hence, the overall direction needs to shift and have three successive goals:

(a) the overarching goal should be the establishment of positive user rights

(b) if this is not possible, the aim should be the transition to a fair use doctrine in the EU

(c) if this is not immediately achievable, extensive collective licensing and linking of funding with open licensing should be sought

While licensing remains an option, it is the very last option we should consider. It currently serves Member States with developed documentation of their works and a positive IPR trading balance, but is highly unlikely to offer any substantial benefits to Member States with nascent information society services and negative IPR trade balances.

5 Appendix: Use Cases

In the Appendix, we present a series of LR use cases, classified on the basis of the type of material and type of use. The use cases are numbered and referenced in the main document of the report, where they seem more relevant. A composite case may appear in multiple different sections of the document:

The following classes of use cases are illustrated:

Class I: PSI

Class II: Open Data and Generic Web Page Crawling and Distribution of Derivative Works

Class III: Distribution of Translated Data

Class IV: Data Anonymisation

Class V: Lexicon Distributions

Class VI: Annotated Dataset Distributions

5.1 Public Sector Information (PSI) Use Cases

5.1.1 Case #1: Uploading-copying Public data (normally under PSI directive) to a repository

Case description	
Actor	Repository manager
Intended use	Upload the ECDC TM dataset to my repository
Conditions	The dataset is accompanied by this licence: http://optima.jrc.it/Resources/ECDC-TM/2012_10_Terms-of-Use_ECDC-TM.pdf
Question	Do I upload the dataset with the original licence to my repo? or Just describe it with metadata, add attribution info, and link to the original site for downloading?
Suggested legal solution	
Legal position	This is a specific type of Public Sector Information (PSI). This is PSI that belongs to one of the EU institutions (e.g. JRC, EC etc.) and hence its distribution is covered by EC Decision 2011/833/EU which is then

	specified in the terms found under the relevant URL. The EC re-use decision effectively allows all types of re-use as long as reference proper is made. The licence describes how the attribution should be made. In our case, the decision defines that the attribution should be made as follows: Copyright (c) EU/ECDC, <YEAR>. In case of the unmodified version, the warranty notices should also be kept.
Suggested course of action	Upload the dataset but: (a) follow the attribution notice (b) keep the copyright and no warranty notices. Practically, the attribution data should be included in the dataset metadata, whereas the copyright licence should also somehow be retained or linked to.
Type of Terms and Conditions	Attribution, Non-endorsement, no warranties, copyright notices.
Legal basis	Copyright Law, 2011/833/EU

5.1.2 Case #2: Uploading-copying Public data (normally under PSI directive) to a repository

Case description	
Actor	Repository manager
Intended use	Upload the JRC-Aquis dataset to my repository
Conditions	The dataset is accompanied by this "Usage Conditions/Licensing Issues" text at http://ipsc.jrc.ec.europa.eu/?id=198#c2726
Question	Do I upload the dataset and also provide a link to the original site or Just describe it with metadata, add attribution info, and link to the original site for downloading?
Suggested legal solution	
Legal position	The licensing information is slightly confusing, since it accepts the AC corpus as being in the public domain, but then moves to impose conditions to its access (attribution and non-endorsement). This may be construed as having individual legislative documents as under the PD and the complete database (corpus) under copyright (or the sui-generis right) and having thus the licence applied only to the copyrighted parts of the corpus. It is also debatable whether the corpus would fall under Decision

	2011/833/EU. In any case, the licensing terms would satisfy the conditions of the re-use decision. Note that the Eurovoc Thesaurus does not fall under the Re-use decision and a special permission is required regarding its re-use.
Suggested course of action	Keep the attribution notice and metadata together with the non-endorsement note. You may upload and share the material through your repository as long as you adhere to these conditions.
Type of Terms and Conditions	Attribution, non-endorsement, copyright notices
Legal basis	Copyright Law, 2011/833/EU

5.1.3 Case #3: Uploading-copying "Open" data to a repository

Case description	
Actor	Repository manager
Intended use	Upload the Opensubtitles (http://opus.lingfil.uu.se/OpenSubtitles.php) dataset to my repository
Conditions	I want to copy it from the (http://opus.lingfil.uu.se/OpenSubtitles.php) site which says "IMPORTANT: If you use the OpenSubtitle corpus, please, add a link to http://www.opensubtitles.org/ to your website and to your reports and publications produced with the data! I got the data under this condition!"
Question	Can I upload the dataset and link to the original site? If yes, which is the original site in this case: http://opus.lingfil.uu.se/OpenSubtitles.php OR http://www.opensubtitles.org/en? Or Just describe it with metadata, add attribution info, and link to the original site for downloading?
Suggested legal solution	
Legal position	This is a case where we have a copyrighted work (subtitles and subtitles database) that is protected under copyright and is licensed under a custom made open licence. Custom made open licences are licences with minimal conditions (i.e. attribution and copyleft) that were made for a specific work or set of works but could potentially interoperate with other open licences.

	The specific licence only requires reference to the original site, i.e. http://www.opensubtitles.org/, to allow all uses of the work. No further attribution or use of notices is required, since the attribution through the URL is meant to cover them all.
Suggested course of action	Upload and share after including URL (i.e. http://www.opensubtitles.org/) as instructed in the licences
Type of Terms and Conditions	Attribution
Legal basis	Copyright Law

5.2 Open Data and Web crawling Use Cases

5.2.1 Case #4: Uploading-copying "Open"/"Public domain"/web crawled data to a repository

Case description	
Actor	Repository manager
Intended use	Upload the SETIMES (http://www.setimes.com/) dataset to my repository
Conditions	I want to copy it from (http://opus.lingfil.uu.se/SETIMES2.php), where it states "A parallel corpus of news articles in the Balkan languages, originally extracted from http://www.setimes.com. The corpus is compiled by Nikola Ljubes(ic' and is taken from http://www.nljubesic.net/resources/corpora/setimes provided under the CC-BY-SA license". On the original setimes.com site, in the disclaimer section it states "Copyright Information. Unless a copyright is indicated, information on the site is in the public domain and may be copied and distributed without permission. Citation of the original source of the information is appreciated. If a copyright is indicated on a photo, graphic or other material, permission to copy these materials must be obtained from the original source."
Question	Do I upload the dataset and link to the original site? Which is the original site in this case? The http://www.setimes.com/, OR http://opus.lingfil.uu.se/SETIMES2.php OR

	http://www.nljubesic.net/resources/corpora/setimes (the url stated in the note of opus) Or Just describe it with metadata, add attribution info, and link to the original site for downloading?

Suggested legal solution	
Legal position	This is a case, where the material itself has a "general rule with exceptions" clause. The general rule here is that the material is in the public domain and the exception is defined by the individually licensed material. This is a rather common construct that allows simplicity and flexibility at the same time. For this approach to be operational, it is necessary that individual material is licensed appropriately. The statement amounts to a waiver in the US and a full license of the economic rights in the EU. Note that the attribution requirement is not a legal condition but rather a soft norm. Reference to a CCBYSA or any other copyright licence is limited to specific items in the database. Attribution is to be made through the use of the URL. If specific attribution requirements are attached to specific material, these have to adhere to the specific licensing conditions (e.g. those of the CCBYSA licence). In the absence of more specific conditions, attribution to individual items has to be done by reference to the URL of the source.
Suggested course of action	Include the material in the collection under a Public Domain Mark (PDM) or CC0 mark. If possible include the original licence as well.
Type of Terms and Conditions	PD, Soft Attribution Requirements, General Rule
Legal basis	Copyright Law

5.2.2 Case #5 Distributing web crawled data I

Case description	
Actor	Researcher-Resource Compiler & Provider

Intended use	Do I distribute a dataset I have compiled using automatic crawling techniques? Being aware of the legal constraints, I have also crawled legal metadata where available.
Conditions	Many of the pages crawled are available under different CC licences
Question	Can I make the whole set available under one CC licence? If yes, which one? Or Do I partition the dataset according to licences and distribute the dataset as a bunch of subsets?

Suggested legal solution	
Legal position	Material crawled from the Web is treated differently in different jurisdictions. Overall, it is suggested to clear before you publish. You cannot use a single CC licence if the material is licensed under multiple CC licences or no CC or no licences at all.
Suggested course of action	Include the material in the collection but only after clearing the content. If clearance is not possible, offer services on the basis of the content and not the corpus itself. When clearing make sure you do not change the CC licences of the source or provide a CC licence unless you have the appropriate licences or rights.
Type of Terms and Conditions	Multiple
Legal basis	Copyright Law

5.2.3 Case #6 Distributing web crawled data II

Case description	
Actor	Researcher-Resource Compiler & Provider
Intended use	Distribute a dataset I have compiled using automatic crawling techniques. Being aware of the legal constraints, I have also crawled legal metadata where available.
Conditions	Although many of the pages crawled are available under (some) CC licence(s), some others do not mention anything about terms of use
Question	Can I make the whole set available under one licence? If yes, which one? Or

	Do I partition the dataset according to licences and distribute the dataset as a bunch of subsets, leaving out those pages that do not contain info as to terms of use?
Suggested legal solution	
Legal position	Material crawled from the Web is treated differently in different jurisdictions. Overall, it is suggested to clear before you publish. You cannot use a single CC licence if the material is licensed under multiple CC licences or no CC or no licences at all.
Suggested course of action	Clear rights and provide only the content for which you have the appropriate licences or rights in accordance to such licences. Note that in most cases you are only re-distributing and not sublicensing or re-licensing material. Hence, do not change the licence unless there is an understanding as to the types of sub-licences and range of re-licensing allowed by the original licence.
Type of Terms and Conditions	Multiple
Legal basis	Copyright Law

5.2.4 Case #7 Distributing web crawled data III

Case description	
Actor	Researcher-Resource Compiler & Provider
Intended use	Distribute a derivative of a dataset I have compiled using automatic crawling techniques. Being aware of the legal constraints, I have also crawled legal metadata where available.
Conditions	Since many pages did not contain legal info, plus for a number of other reasons, I decided to build a language model out of it (2-5 grams)
Question	Can I distribute it under the licence of my choice? Or Do I obey the terms of use for those data for which there is legal info (=restrictions)?
Suggested legal solution	

Legal position	Creating a language model out of material would most probably fall under fair use doctrine unless the original work may be constructed out of this or it may deemed as an unauthorised derivative work. There are issues with regards the definition of the derivative work but in all probability creating n-grams would constitute derivative work precisely because of the dependence on the original work.
Suggested course of action	Follow the terms of the licences, where these exist. Try to group material in accordance to the licences they belong to. More specifically: (a) if they are CCBY material you may use the derivatives with every other licence (b) if they are SA material make sure they are grouped with the same type of -SA material (e.g. CCBYNCSA only, CCBYSA only) (c) you have no restrictions on the use of derivative works under a CCBYNC licence, so you may follow the CCBY rule (d) you have no restrictions with regard to CCZero and PDM works (e) there are multiple compatibility issues with regard to OKF Open Data Commons Attribution licences as they refer to the database and not its content, something which is difficult to apply in practice (f) Open Knowledge Foundation Open Database Licence (ODbL) cannot be mixed with any other SA licence (g) all (Attribution) Open Government Licences have the same treatment as (a). If you cannot clear, provide a service based on the n-grams rather than the n-grams themselves.
Type of Terms and Conditions	Multiple
Legal basis	Copyright Law

5.2.5 Case #8 Distributing web crawled data IV

Case description	
Actor	Researcher-Resource Compiler & Provider
Intended use	Distribute a subset of a dataset I have compiled using automatic crawling techniques. Being aware of the legal constraints, I have also crawled legal metadata where available.

Conditions	Since I only want to distribute a few hundred sentences, I tend to be agnostic to the legal conditions. Therefore I have picked these hundred non-consecutive sentences and packed them in a dataset that I want to distribute for technology evaluation.
Question	Can I distribute it under the licence of my choice? Or Do I make sure that I have the right to distribute even sentences extracted from a dataset?
Suggested legal solution	
Legal position	The question relates to the degree to which (a) the work is a derivative (b) the use of the work may fall under limitations and exceptions. The very act of crawling may constitute copyright violation in a number of jurisdictions. Especially in Europe it is difficult to fit it under any specific limitation or exception, though there is active work on introducing a data-mining exception (see e.g. UK copyright amendments).
Suggested course of action	Distribute the material under any CC licence if the original material cannot be reconstructed. NC licences could denote that the material should not be used for commercial purposes adding thus to a due process strategy.
Type of Terms and Conditions	Multiple. NC element suggested.
Legal basis	Copyright Law. Emphasis on limitations and exceptions.

5.2.6 Case #9 Distributing web crawled data V

Case description	
Actor	Researcher-Resource Compiler & Provider
Intended use	Distribute a subset of a dataset I have compiled using automatic crawling techniques. Being aware of the legal constraints, I have also crawled legal metadata where available.
Conditions	Since I only want to distribute a few hundred paragraphs, I tend to be agnostic to the legal conditions. Therefore I have picked these hundred paragraphs and packed them in a dataset that I want to distribute for technology evaluation or development (e.g. parameter tuning).
Question	Can I distribute it under the licence of my choice?

	Or Do I make sure that I have the right to distribute the paragraphs extracted from the dataset? If yes, then does reshuffling the paragraphs help circumvent the problem?
Suggested legal solution	
Legal position	See Case #8. Reshuffling paragraphs reduces legal risk since it reduces the possibility of having the work reconstructed or substituted.
Suggested course of action	See Case #8.
Type of Terms and Conditions	Multiple. NC element suggested.
Legal basis	Copyright Law. Emphasis on limitations and exceptions.

5.2.7 Case #10 Distributing web crawled data VI

Case description	
Actor	Researcher-Resource Compiler & Provider
Intended use	Distribute a subset of a dataset I have compiled using automatic crawling techniques on AUTOMOTIVE websites.
Conditions	Most probably the websites and relevant pages are copyrighted
Question	Can I distribute it under the licence of my choice (possibly adding a phrase *a la Wacky* "If you want your webpage to be removed from our corpora, please contact us.")? Or Do I make sure that I reshuffle the sentences in the paragraphs, before distributing? Or Just refrain from distributing it?
Suggested legal solution	
Legal position	See Case #9. Pay particular attention to the Terms of Use on the relevant websites. While Fair Use/Doctrine conditions may be on your side in many jurisdictions, these may be circumvented through contractual terms

	expressed in the Terms of Use of the relevant web-site. If no specific conditions exist, see analysis and suggestion under Case #9. The existence of a notice and take-down procedure is always helpful and is suggested in all web-crawling corpora.
Suggested course of action	See Case #9. Include a notice and take down procedure.
Type of Terms and Conditions	Multiple. NC element suggested.
Legal basis	Copyright Law. Emphasis on limitations and exceptions.

5.2.8 Case #11 Distributing web crawled data VII

Case description	
Actor	Researcher-Resource Compiler & Provider
Intended use	I want to build a corpus *a la Wacky* (http://wacky.sslmit.unibo.it/doku.php)
Conditions	For sure I cannot guarantee the copyright-free"ness" of the crawled data. Does a sentence as the one at the end of http://wacky.sslmit.unibo.it/doku.php?id=corpora save me?
Question	Can I use such a phrase in tandem with a Notice and Take Down policy such as the META-SHARE MoU? Or Do I refrain from doing anything with it?
Suggested legal solution	
Legal position	See sections regarding webcrawling. The last sentence "If you want your webpage to be removed from our corpora, please contact us." is a good practice though not perfect under the current copyright system. It amounts to a notice and take down procedure though not very accurately described.
Suggested course of action	(a) Check meta-data that stop crawlers in relevant web sites (b) try to crawl CC-licensed pages or PD pages (c) have an idea of the extent of your non-licensed part of your corpus (d) list the websites you crawled from (e) use CC NC licences (f) have an opt-out or notice and take down procedure clearly stated on your web site (g) make sure that the original content cannot

	be fully re-constructed or even when it is, that it cannot substitute the original content.
Type of Terms and Conditions	Multiple. NC element suggested.
Legal basis	Copyright Law. Emphasis on limitations and exceptions.

5.3 Distribution of Translated Data Use Cases

5.3.1 Case #12 Distribution of Translated Data I

Case description	
Actor	Researcher-Resource Compiler & Provider
Intended use	Distribute a parallel corpus (original & translation in one or more languages) consisting of public original text, translated in languages A and B by a company.
Conditions	Although according to PSI the content should be publicly used, the following link states almost the opposite: http://www.ermis.gov.gr/portal/page/portal/ermis/termsOfUse.
Question	Can I distribute it under the licence of my choice, presupposing that the Terms of Use are violating the PSI directive? Or Do I contact the content owners? Or refrain from distributing and using it altogether?
Suggested legal solution	
Legal position	There are two issues here: one is how I can use PSI in general, the second is whether I can consider PSI as PD or licensed under a very permissive licence (e.g. CCBY- like) by default. The question of corpus translation depends on our reading of PSI terms of use since it constitutes a transformative use or derivate work. PSI in most countries is not open by default. Greece is possibly one of the worst cases since the licensing proper of PSI requires a ministerial degree. Another rather confusing for

	the re-users aspect of PSI is that in many EU Member States (MS), it is not clear whether PSI falls within Copyright at all (some types of PSI are exempted from copyright) or the use of PSI constitutes one of Copyright's limitations and exceptions or it is licensed and under which licence it is licensed. In the absence of a harmonized regime for PSI, the re-user needs to go back to the terms and conditions of the specific web site under which the content is provided. In the absence of such conditions, the re-user will have to check the specific jurisdiction's conditions regarding PSI's status or seek for the appropriate permission, the latter often being an easier option.
Suggested course of action	Check copyright or use terms and follow them. In the absence of conditions, check if PSI is exempted material or falls within the Copyright limitations and exceptions in the jurisdiction where it is made available. In the absence of any other specific rules you cannot translate and you need to seek permission.
Type of Terms and Conditions	Multiple. See specific Legislation.
Legal basis	Copyright Law, PSI laws

5.3.2 Case #13 Distribution of Translated Data II

Case description	
Actor	Researcher-Resource Compiler & Provider
Intended use	Distribute a parallel corpus which has been crawled from the web (original & translation in one or more languages) from http://www.michael-culture.gr/mpf/pub-mpf/index.html
Conditions	There are no terms of use, no licence info whatsoever
Question	Can I distribute it under the licence of my choice? Or Do I refrain from distributing and using it altogether?
Suggested legal solution	
Legal position	See analysis in Case #12. This is PSI and the ministry is obliged to licence the material (both for commercial and non-commercial uses), though not

	necessarily under a Creative Commons licences. It also has to explain the charging basis.
Suggested course of action	Actively ask for a licence. Try to group the material per ministry in order to reduce transaction costs.
Type of Terms and Conditions	Multiple. See specific Legislation.
Legal basis	Copyright Law, PSI laws

5.3.3 Case #14 Distribution of Translated Data III

Case description	
Actor	Researcher-Resource Compiler & Provider
Intended use	Distribute a parallel corpus (original + translation in one or more languages) consisting of copyright-free original text (e.g. a novel over 70 years since publication), translated in languages A and B
Conditions	Although I know I can distribute the original novel (or not?), I know nothing about the translation of this literary work
Question	Can I distribute it under the licence of my choice? Or Do I contact the translated content owners (if any)? Or Do I refrain from distributing and using it altogether?
Suggested legal solution	
Legal position	If the material is in the public domain, I am fully entitled to make translations and distribute under any licence I deem appropriate. If the translation is being done by a third party, then permissions should be sought as it constitutes a new work.
Suggested course of action	Ask for a licence from the translators.
Type of Terms and Conditions	The translator's restrictions/terms and conditions.
Legal basis	Copyright Law

5.4 Data Anonymisation Use Cases

5.4.1 Case #15 Anonymising Data Sets

Case description	
Actor	Researcher-Resource Compiler & Provider
Intended use	Distribute a corpus which contains person names and other personal data
Conditions	I have a licence for disseminating the material but the personal data should not be exposed
Question	Can I distribute the corpus after anonymising the personal data Or Should I refrain from disseminating it?
Suggested legal solution	
Legal position	Anonymisation is useful only to the extent that (a) the corpus is still useful and relevant after the anonymisation (b) if it is clear who and under which conditions may proceed with the anonymisation (c) if the data protection provisions allow anonymisation and the subsequent distribution of data.
Suggested course of action	See how anonymisation relates to the desired use.
Type of Terms and Conditions	Data protection rules/Terms and Conditions
Legal basis	Data protection law

5.5 Lexicon Distribution

5.5.1 Case #16 Lexicon Distribution (based on other datasets) I

Case description	
Actor	Researcher-Resource Compiler & Provider
Intended use	Distribute a lexicon I have created on the basis of web crawled data (lemma selection, lemma frequency)
Conditions	I have no (or little) information on the legal info of the web crawled texts, since I only used the dataset to extract specific information from it

Question	Can I distribute it under the licence of my choice? Or Do I check again all the sources and ask for permission, which can be extremely time-consuming Or refrain from distributing it?
Suggested legal solution	
Legal position	See Cases #8 - 10
Suggested course of action	Key suggestions: (a) the closer to the original the greater the risk (b) if the original can be substituted by what I have produced, then I should not distribute my work (c) if the original cannot be recognised or reconstructed the risk is too low (d) ideally I should not be using it for commercial purposes (e) if the risk is high include a notice and take-down notice
Type of Terms and Conditions	Multiple. NC element suggested.
Legal basis	Copyright Law. Emphasis on limitations and exceptions.

5.5.2 Case #17 Lexicon Distribution (based on other datasets) II

Case description	
Actor	Researcher-Resource Compiler & Provider
Intended use	Distribute a lexicon I have created including web crawled data used in the form of examples
Conditions	I have no (or little) information on the legal info of the web crawled texts, since I only used the dataset to extract specific information from it
Question	Can I distribute it under the licence of my choice? Or Do I check again all the sources and ask for permission, which can be extremely time-consuming Or Refrain from distributing it?

Suggested legal solution	
Legal position	See Case #16.
Suggested course of action	Same as Case #16. The inclusion of my own or cleared content further reduces the risks of infringement.
Type of Terms and Conditions	Multiple. NC element suggested.
Legal basis	Copyright Law. Emphasis on limitations and exceptions.

5.5.3 Case #18 Lexicon Distribution (based on other datasets) III

Case description	
Actor	Researcher-Resource Compiler & Provider
Intended use	Distribute a lexicon I have created which includes definitions (paraphrased but also verbatim) from other lexica (one or more)
Conditions	The licence of the original lexica is not always clear (e.g. there is nothing in the *Dictionary of modern Greek*, while the web site for the Triantafylides dictionary includes the following terms of use: http://www.greek-language.gr/greekLang/terms/index.html)
Question	Can I distribute it under the licence of my choice? Or Do I ask for permission from sources? Or Simply state the sources (attribution-like) Or Refrain from distributing it?
Suggested legal solution	
Legal position	Inclusion of definitions from another lexicon will most probably constitute extraction of substantial parts from another database, even if these are few. The reasons is that to extract them they are significant and thus by definition substantial. For these, permission should be sought. Apparently,

	the fewer the definitions, the lower the risk. If the definitions are paraphrased and there could not be a way to define a term differently and the structure of the lexicon is not copied, then there should be no problem, since facts and ideas are not copyright protected. The prohibition of paraphrasing in the terms and conditions of a web-site is only valid to the extent that your paraphrasing has copied original elements in the expression of the original definition. Otherwise, ideas are not protected.
Suggested course of action	Paraphrase where there is not other way to say something. Avoid copying any original elements. Do not do excessive verbatim copying.
Type of Terms and Conditions	Multiple.
Legal basis	Copyright Law. Emphasis on limitations and exceptions.

5.6 Annotation Cases

5.6.1 Case #19 Distribution of Annotations of a Data Set I

Case description	
Actor	Researcher-Resource Compiler & Provider
Intended use	Distribute the annotations to an original dataset (e.g. images, broadcasts) incl. links to specific instances of the web available dataset (e.g. the hand on an image, time sequence of a film)
Conditions	I have asked for permission for the originals but received no reply
Question	Can I distribute it under the licence of my choice? Or Do I ask for permission from sources Or Simply state the sources (attribution-like) Or Refrain from distributing it?
Suggested legal solution	
Legal position	The annotations, to the extent they are my own original work or constitute works that are not copyright protected (e.g. a URL) can be distributed

	under any licence of my choice. With regard to the inclusion of extracts from the original work see above.
Suggested course of action	Do not ask permission for linking but only for re-using original content. When the annotation includes original extracts it may still be deemed fair use/dealing or falling within the limitations and exceptions as a form of critical approach -in the broadest sense- of the original work (see e.g. Greek law 2121/1993, art.19).
Type of Terms and Conditions	Any licence of my choice.
Legal basis	Copyright Law. Emphasis on limitations and exceptions.\