SEVENTH FRAMEWORK PROGRAMME
THEME 3
Information and communication Technologies

PANACEA Project

Grant Agreement no.: 248064

Platform for **A**utomatic, **N**ormalized **A**nnotation and
Cost-**E**ffective **A**cquisition
of Language Resources for Human Language Technologies

D2.4 Annex 1

Issues related to Data Crawling and licensing

Dissemination Level:	Confidential
Delivery Date:	January, 28, 2013
Status – Version:	final
Author(s) and Affiliation:	Victoria Arranz, Khalid CHOUKRI Olivier Hamon (ELDA), Núria Bel (UPF), Prodromos Tsiavos (Legal advisor)

Table of Content

List of figures

I Executive Summary

The main objective of this report is to address a series of issues facing the Panacea consortium with regards to legal issues related to the sharing and distribution of LRs, workflows, and Web-services. This comprises the three major outcomes of the project: the Panacea Platform (the LR production factory and the associated workflow engine), the Web-services used/usable within the platform, and the Language Resources developed within the project. Although this report is an appendix to the Exploitation plan of the project[1], it was drafted to cover issues of interest to the whole community on legal aspects and to be disseminated as a separate/independent report.

II PANACEA objectives and legal framework

Panacea is the LR production factory that manages sets of workflows that target the production of Language Resources (LRs) through the exploitation of web-services.

The Panacea project achieved a number of objectives regarding the set-up of the factory. Among these objectives, we can list:

- ✓ Production of a set of language resources (e.g. monolingual, bilingual textual corpora, etc.), based on web sources.
- ✓ Exploitation of various tools (from within the consortium as well as from third parties), set-up as web-services, running either on a Panacea Partner server or on the original third party server.
- ✓ Design of workflows …

These led to valuable assets that we are planning to exploit.

I.1 Language Resources (LRs):

The Panacea Factory allowed to produce and validated a large set of useful language resources for the Machine Translation purposes. These comprise:

- Monolingual Corpora (EL, EN, ES, FR, IT)
- Monolingual Corpora N-grams (EL, EN, ES, FR, IT)
- Monolingual Dependency Parsed Corpora (EL, ES, IT)
- Bilingual Aligned Parallel Corpora (EN-FR EN-EL)
- Bilingual Glossaries (EL-EN, FR-EN, DE-EN)

[1] D2.4, Platform Software, Project Tools + Resources Licensing Policy and Exploitation Plan

- Monolingual lexica (EN, ES, IT)

I.2 Web-Services

Web-services2 are tools provided by the consortium or third parties that allow processing some data input and produce some data as output. The typical example of web-service is the language identifier. Such web-service uses an algorithm to identify the language of an HTML/XML/PDF/… file and is running on some servers.

It is important to understand how one can exploit such application and for what purpose. In addition, one need to understand what are the legal issues involved. It is not always the case that web-service use is governed by clear use conditions and terms, so one need interpret these or their absence with respect to the usages but also the rights on the outcomes.

I.3 Workflows

These are various workflows that implement the way web-services are chained as sequences, with specific input and output. For instance, a workflow that consists of a sequence of two "calls" to web-services and may take as input a set of URL and identify the language(s) in which they are written and hence give as out a table with pairs (URL, language). It could be chained with one that crawl all URLs that are in a given language.

Such work flow could be (identify-languages (list of URLs) + Crawl (URL, Language=Fr).

(1) (Input = set of URLs) ➔ Web-services 1: identify languages ➔ (output (URLs, languages)

(2) (input (URLS, Languages) ➔ Web-services 2: Crawl (URLs, Language=Fr) and store ➔ output (set of XML, HTML, PDF, … pages stored on a given address)

It is clear that more sophisticated workflows could be designed, requiring deep analysis and more creativity and hence triggering the application of copyright.

Some of the workflows are illustrated by the diagrams given below.

The first workflow (ILSP Basic NLP Tools), illustrated herein, accepts a list of URLs to XCES files with Greek paragraph-segmented content. It then uses web services for three basic ILSP NLP tools (Sentence Splitter and Tokenizer, FBT Tagger, Lemmatizer) to process the content. Each service first downloads all XCES files to the server and, on successful completion, returns to the client a list of URLs to XCES files with the automatic annotations (http://myexperiment.elda.org/workflows/20).

[2] (Wikipedia): A Web service is a software function provided at a network address over the web or the cloud, it is a service that is "always on" as in the concept of utility computing. The W3C defines a "Web service" as "a software system designed to support interoperable machine-to-machine interaction over a network". It has an interface described in a machine-processable format (specifically Web Services Description Language, known by the acronym WSDL). Other systems interact with the Web service in a manner prescribed by its description using SOAP messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards."

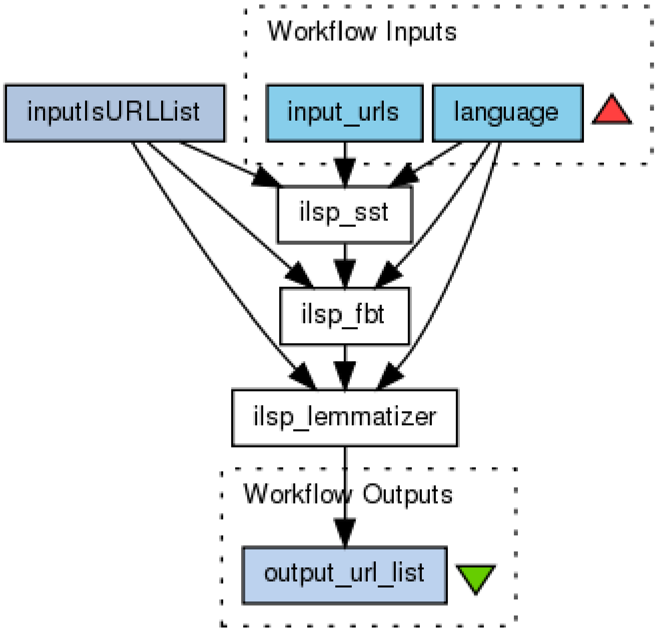

Figure 1: Diagram of workflow 1, ILSP NLP basic tools

The second workflow is about the "Freeling tagging for crawled data". This is the tagging workflow for crawled data using Freeling. Freeling is run using the "keeptags" option to remove boilerplate and to keep paragraph tags info from the input data. The output is converted to the Travelling Object format (TO1 xces) (http://myexperiment.elda.org/workflows/5).

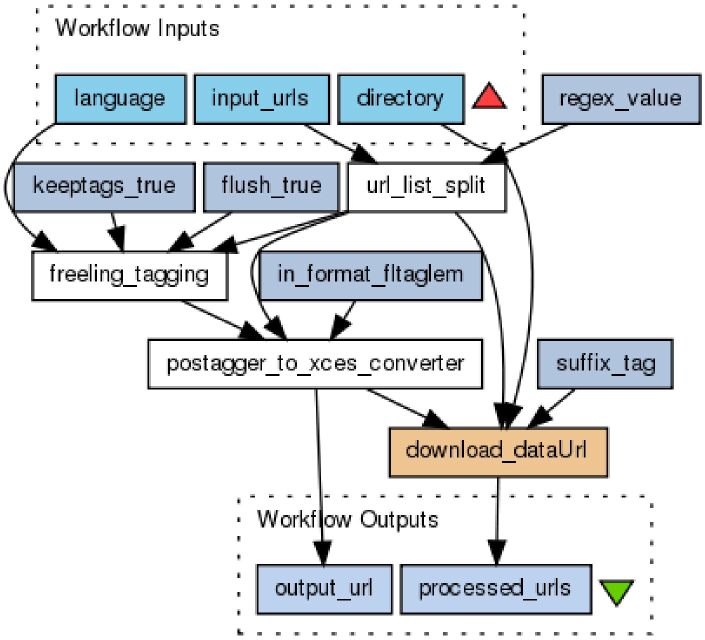

Figure 2: Diagram of workflow 2, tagging of crawled data

III Objectives of this report

The main objective of this report is to address a series of issues facing the Panacea consortium with regards to legal issues related to the sharing and distribution of LRs, workflows, and Web-services. As indicated above, among the assets of Panacea is the data produced using various web-services. The legal issues involved herein are about the possibility of crawling and using the data (both crawled data (raw/primary data), and processed data). Most of the raw data may be found under different licensing schemes or without any licensing information, on the Internet. More specifically, the report elaborates on the legal status of data crawled from the Internet both for internal use and to share with third parties. It also elaborates on the processed data (derivatives) that may be useful to the community, assuming one can release them without too much risk!

As said above, in addition to the data legal aspects, the report addresses other issues of interest to Panacea and the users of its Factory and related to:

- What is the optimal way to license web-services (See how to relate that to Panacea)
- What is the optimal way to license Panacea workflows
- What is the optimal strategy for integrating the PANACEA platform and the META-SHARE infrastructure?

IV 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 regards to these acts.

Copyright Law is not fully harmonized 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 jurisdiction, where possible act of infringement takes place. However, there are some common rules described in international treaties, mainly the Berne Convention, the TRIPS agreement and the WIPO treaties that provide us with an understanding of copyright rules at the international level, whereas a series of directives at the EU level provide an even more harmonized legal regime for the Member States of the European Union[3].

[3] 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.

The report 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 report 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.

V The issue of Data Crawling: Protected vs Unprotected data

The most important and crucial part of Panacea Data production is related to the data crawled/harvested from Internet. Data crawling may be generally 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. Other specific cases may fall under this definition, in particular "manual crawling" (with browse/store actions).

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

The data to be crawled may consist of PROTECTED MATERIAL OR UNPROTECTED material. All the material found on the web is material that potentially constitutes protected subject matter. This may fall under the following broad categories:

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

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 (protected material), 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 (unprotected 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). 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.

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 or 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.

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.**

b. **the extent to which such acts are visible enough to expose an organization to the risk of legal action.**

V.1 The Acts of Crawling

In the case of web crawling, the acts would certainly include:

1. copying and processing of the relevant information
2. potentially the creation of derivative works
3. and the communication to the public either of parts of the original work or a derivative work.

Each of these acts needs special treatment as elaborated upon in the coming sections.

V.2 Act of Copying

Copying: the act of crawling certainly involves the reproduction of content and hence activates the reproduction right.

According to the Copyright Directive[4] 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 (the data driven paradigm based on statistical modeling). It will also be qualitatively substantial, because it has significance for the entity 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 is clear that in the case of crawling for language resource processing there is very little of the temporary copying falling under article 5[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 an intermediary or a lawful use which 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

[4] 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.

[5] Article 5 of the copyright directive attached as an appendix to this report

monitoring service, could be implicating its subscribers in copyright infringement by distributing reports 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 license may thus be in breach of publishers' content, notwithstanding any license 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.

V.3 Act of realization of derivative work

Derivative works: in Panacea context, the act of crawling will either mean the collection and storage of parts of the websites or will also be followed by additional processing once the material is collected. As a result, derivative works[6] 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 be construed as falling under the limitations and exceptions provisions and hence it will also require separate permission by the rights holders.

V.4 Act of sharing with the community

The objective of crawling, in the context of Panacea, is to share the produced resources with the community or communicating to the public. The outcome of the series of acts starting with web crawling, and continuing with the processing of the collected data, targets 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 implements the web crawling and processing without any direct communication of the actual web data, then copyright law is not activated at all. This assumes the web-service will be running on the end user servers.

V.5 Visibility of the outcomes of crawling

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 license. 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 (assuming the LTs cannot offer possibilities of reconstructing the original data or parts of it (e.g. a list of words from a Textual Corpus). Accordingly, unless the information crawled from a specific web site is substantial for the operation of the end user service (and sources not identifiable), the legal risk drops dramatically.

[6] As a basic processing we can mention converting HTML , PDF files to plain text or XML, etc.

VI Fair Practices wrt to the outcomes of crawling

(1) 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 license 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 license rather than seek for a commercial license from the collecting societies. In many cases it's recommended to use the services of experienced agencies (ELRA, LDC, …) to negotiate such rights, including for commercial usage.

- In some jurisdictions (especially in the US, where fair use is applicable), there is the concept of the implied license with regards 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 (some) technical measures to stop (robot) 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, licenses are very narrowly construed only to cover the explicit acts the rights owner would like to authorize. 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: (a) only crawls sites where bots are allowed (b) has a notice publicly stating that its content only derives from web sites that do not prohibit crawling (c) provide a brief explanation as to how someone could stop her site from being crawled (d) produce 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.

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.

VII Issues related to WEB-SERVICES

The question of how web services should be legally treated in the context of an LT project is one that needs to take into consideration the four key components of any web service, i.e.

A. the data/ content it requires and delivers

B. the software that is packaged and set-up as web service

C. the web service as such, i.e. the use of a specific User Interface that reacts to a certain input and provides a specific output in the form and manner allowed by the web service.

D. The server on which the web service runs

Each one of these components needs to be treated separately in order to provide the full scale of the options available with regards to the terms and conditions under which a web service should be offered. The approach we follow in this section is that by analyzing each of the three constituent parts of a web service, we are able to identify the elements that the web service Terms and Conditions should have and the choices the web service providers and third parties would have to make to define how the web service is to be offered. Some parts are also related to data issues that have been treated in previous sections but are given herein for completeness.

<u>(A) Data/ Content:</u> As mentioned above, a web service both requires some form of input and provides some form of output to the end user. This means that it is necessary to ensure a smooth flow of rights along with the data in the web service and to the end user. However, methodologically, we should start with the forms of data output, as this is going to define the range of rights necessary at the stage of the data input. We also need to assess the type of data processing that takes place in order to ensure again that all necessary rights have been obtained or that the legal risks have been minimized.

(A.1) **Data Output**: There are two types of data output that may come out of the web service.

[I] The first type is data or resources as such. This is particularly the cases where the web services include services such as viewing or downloading. In such cases, the basic rule to be followed is that the data or resources should be delivered to the end user with a license that confers the same or less rights than the one obtained by the web service provider. In other words, the service provider cannot give to the end user rights that she

has not obtained. Another approach that is followed particularly in the cases where third party content is provided, is that the service provider either has a re-distribution contract with the content/ data provider that sets the terms and conditions under which the material is to be provided or the service provider does not obtain any license at all but rather just passes the resource to the end user with a license of the choice of the content/ data provider. Depending on the type of the license under which the content/ data is provided to the end user, it may be that the license also covers the retaining of the data/ content on the service provider's servers or that the content/ data is delivered to the end user through a web service that draws data from different source that do not necessarily reside on the web service provider's servers. We will return to this issue, when we cover the data input side.

[II] The second type of data output is processed data. In this case, the service provider obtains data from different sources, processes them and then provides the end user with a final result. This means (a) that the service provider needs to have the right to process the data, which is a question we explore in the following point, and (b) that depending on how different and distinct from the original data the service output is as well as on the license under which the original content was made available, the web service provider has the ability to choose the kind of license she wishes to make the new data output available with.

(A.2) **Data Processing:** When the data have been collected as input and before the result is send to the end user, they are processed by the software that is operated by the service provider. This practically means that the service provider needs:

(a) either to have obtained the licenses necessary to process the data

(b) To be informed of the license obtained by the end user that allow him/her to make collect/process such data

(c) to operate on data that are in the public domain

(d) the processing has to fall under the limitations and exceptions or fair use/ dealing.

This essentially brings us back to the discussion of the legality of web crawling presented above. The safest strategy is to differentiate the data/ content in accordance to the categories mentioned above, always seek to have an explicit than implied license and only rely on limitations and exceptions if this is the last resort to make the processing necessary for offering the related web services.

1. **Data Input:** This is the perhaps the most important stage in the chain of acts that have to be taken into consideration by the web service provider as it defines both the range of actions with regards to processing and the types of data outputs the web service may provide. Data input may belong to the following categories:

   (I) Input by the user (push model). This may be data manually entered to the system, uploaded as files or provided to the web service through aggregation or APIs. In all three cases the web service provider needs to take reasonable measures in order to ensure that the data provided by the user are granted with such a license so that processing and data output is actually legal. This has to be communicated to the end user through the web service terms and conditions. However, the service provider has no special obligation to clear rights over the input content/ data if the end user agrees with the terms and conditions and declares that all rights have been cleared. A process of notice and take down would be particularly helpful in cases where a third party identifies a rights violation. In any case the service provider has to inform the data/content provider regarding the type of processing (especially if it involves personal data), the degree to which such processing will take place on the web service provider's site or somewhere else, what kind of processing will take place, if the data are going to be retained and to whom the data are to be shared with or disseminated to.

(II) Input by collecting data from third sources (pull model). This is the case where the web service provider actively seeks to obtain content from third parties either through specific licensing schemes or through a fair use/ dealing regime. Again, the preceding analysis regarding clearing of rights and what applies to web crawling is useful. The service provider is always encouraged either to use material that fall in the public domain or obtain explicit licenses and only do web crawling after mitigating risks as indicated above. Obtaining the necessary licenses also includes making redistribution agreements, passing licenses to the end user or use Open Public Licenses such as Creative Commons or FOSS licenses. It is important to note that the license should not only cover redistribution but also the types of processing that the web service requires. In the case where the processing is not perceptible to third parties and the service provider provides a data output that is based on the original work but is not its derivative work, while a license allowing adaptations of the work is technically required the risk of a lawsuit is extremely low. In all cases documentation of the due diligence steps is highly recommended. Another scenario here is the case where the web-service provider is not given Data to process for the user but a list of URL and web sites. The web service would then crawl and process them for the user.

(B) Software: The software which is used to offer the necessary web services needs to be licensed to the service provider under a license which allows him/her: (a) to adapt the software so that it may be offered as a service and (b) to offer the software to third parties for commercial or non-commercial purposes. All software is accompanied by licenses that set the conditions of its use and is either accompanied by the source code or not. In any case the web service provider will have to obtain as many rights from the software provider as possible since it is frequently necessary to make changes to the code and is certain that the software or some of its specifications are to be used by third parties. It is, hence, strongly suggested that either the service provider obtains a license giving her all necessary rights to materialize the web service or opt for a Free/ Open Source License. In most cases, the service provider is also the owner of the web-service and the server on which it runs so the legal risks described below are minimized

Even though, the terms and conditions of the web service as such have to contain elements related both to point (a) and (b) above. More specifically, they should include terms with regards to the licensing of the data input by third parties, make reference to the licensing conditions under which the software is used if it is provided by a third party and define the licensing terms or make reference to specific third parties' licenses with regards to the data output. It should also contain specific points regarding:

1. The liability of the service provider, which will have to be waived
2. The service level which is going to be provided to the end user
3. A set of warranties and disclaimers with regards to the service which essentially has to be offered as is. The disclaimer should also extend to the software and the content, if they are provided by a third party.
4. What constitutes fair use of the service in order to avoid abuses or overload of the service?
5. A waiver of liability for acts of the end users.
6. Interoperability provisions and terms and conditions under which the Application Interface may be made available to third parties.

VIII Legal status of Panacea Workflows and relationship with web services

The workflows presented in the context of the PANACEA project are effectively XML documents that may invoke web services but only as they are executed. As an XML document, a workflow is legally speaking a literary work or a database and in that sense it constitutes content that may be licensed under any relevant public open license. Creative Commons is a family of standard licenses that could actually be used in order to license XML documents and in that sense it is suitable for a workflow as well. The kind of CC license chosen for the workflow is clearly related to the kind of protection or the kind of end result that the owner of the workflow needs to achieve. More details are given in Appendix B to this report. Broadly speaking, Panacea consortium agreed to go for the most permissive licenses when it comes to its workflows. Some restrictions may apply, hence the use of different variants of the Creative Common licenses.

It is important to consider the workflows as one way to promote the Panacea Factory and as such consider the most permissive licenses for their distribution and redistribution.

IX Integration of PANACEA into META-SHARE

PANACEA has a great interest in its service being integrated within the META-SHARE platform, as it will allow them a broader audience and will maximize the re-use of its material. This is more a strategic concern than simply legal. As of today, META-SHARE remains a platform usable to share Language Resources and therefore not ready to interoperate with the Panacea platform. Panacea will exploit the meta-share platform capacity to deposit and share language resources with respect to the set of resources produced within Panacea.

In order to achieve a stronger integration in the near future with the least possible frictions and achieve the maximum returns the following points have to be taken into consideration:

1. All content and data contained in PANACEA have to be gradually cleared as recommended herein

2. Such content has to be documented following the META-SHARE meta-data schema, including for the legal part. This means that particularly in the case of data obtained through web crawling, if they are offered as a corpus, there have to be accurate licensing information as to what the re-users of such corpora are allowed to do and what not.

3. In the case where content is provided through a web-service and the web-service is treated as an LT, there needs to be a specific set of Terms and Conditions under which they are to be offered as described above. These Terms and Conditions have to accompany the Tool and be properly documented within the META-SHARE platform.

4. In the case where workflows are offered as a separate resource there needs to be licensing information with regards to the workflow and an understanding why a specific type of license is chosen over another one.

5. In all cases where dual licensing is offered there needs to be an assessment of the business model behind the choice of the specific licensing model. It is broadly suggested that the licensor decides where the primary source of economic value derives from.

As indicated above, given that the Panacea core value is in creating a community of re-users the whole stack of workflow, web service-content and related software have to be offered for an as open license as possible.

Overall, there has to be a good understanding of the types of value that are to be served with each of the licenses and a wide range of options offered to the PANACEA members to choose. In that sense, the integration of its services with META-SHARE will provide the PANACEA members with a set of ready-made resources that could be easily used and processed to adapt to their needs.

X Concluding remarks and recommendations

The legal analysis of the previous sections 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 actions described below.

In order to maximize benefits and minimize risks PANACEA needed to address:

(a) Data crawled by the PANACEA FACTORY and passed (with/without additional processing/values) to third parties

(b) Status of web services offered to third parties so as to produce customized data from third parties sources (e.g. Crawled by PANACEA or by the end-user) and the status of the associated data (input /output)

(c) Status of work flows exploited by the Panacea FACTORY to manage the combination of web services so as to produced high value resources.

More specifically:

- At the stage of web-crawling or collecting of third party material, it is necessary to obtain all necessary rights, to avoid violating existing licensing terms and have an easily enforceable and very visible notice and take down procedure. The due diligence steps in clearing the material have to be clearly documented and archived. This allows sharing the produced data (Panacea data) with third parties under licenses that necessarily comprises a NoRedistribution and a clause on deletion of resources if owners require so.
- At the stage of processing, the web service providers will have to assess the degree to which the processing itself has to be made public. If it is only the results of such processing that are important for the audience, the sources of the material and the kinds of processing may be better if they are not made public. This will drastically decrease risks of legal action against the web service providers.
- At the stage of communicating and offering the material to the public, it is important that the provider has a clear understanding of it business model before choosing the relevant license. Broadly speaking, the primary source of value has to be identified and accordingly to decide whether she will opt for a license that allow re-use or not as well as how to charge for different layers of the service stack (i.e. the content, the service or the software).

At all stages it is necessary that the flow of rights is not interrupted, i.e. that the rights offered from one stage to the other always follow two rules:

(a) that you cannot give more rights than the one you have obtained; and

(b) that in order for (a) to happen you need to have proper documentation of licensing (e.g. CC licensed) and IPR regimes (e.g. material in the Public Domain) for all types of material collected.

If these rules are observed the legal risks are substantially lowered or may be easier to assess, case by case. Most importantly the documentation proper of the legal rights over the existing material allows for informed decision making with regards to the most appropriate business model to follow.

XI Appendix A: Article 5 of the EU copyright directive

Article 5

Exceptions and limitations

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

(a) a transmission in a network between third parties by an intermediary, or

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

2. Member States may provide for exceptions or limitations to the reproduction right provided for in Article 2 in the following cases:

(a) in respect of reproductions on paper or any similar medium, effected by the use of any kind of photographic technique or by some other process having similar effects, with the exception of sheet music, provided that the rightholders receive fair compensation;

(b) in respect of reproductions on any medium made by a natural person for private use and for ends that are neither directly nor indirectly commercial, on condition that the rightholders receive fair compensation which takes account of the application or non-application of technological measures referred to in Article 6 to the work or subject-matter concerned;

(c) in respect of specific acts of reproduction made by publicly accessible libraries, educational establishments or museums, or by archives, which are not for direct or indirect economic or commercial advantage;

(d) in respect of ephemeral recordings of works made by broadcasting organisations by means of their own facilities and for their own broadcasts; the preservation of these recordings in official archives may, on the grounds of their exceptional documentary character, be permitted;

(e) in respect of reproductions of broadcasts made by social institutions pursuing non-commercial purposes, such as hospitals or prisons, on condition that the rightholders receive fair compensation.

3. Member States may provide for exceptions or limitations to the rights provided for in Articles 2 and 3 in the following cases:

(a) use for the sole purpose of illustration for teaching or scientific research, as long as the source, including the author's name, is indicated, unless this turns out to be impossible and to the extent justified by the non-commercial purpose to be achieved;

(b) uses, for the benefit of people with a disability, which are directly related to the disability and of a non-commercial nature, to the extent required by the specific disability;

(c) reproduction by the press, communication to the public or making available of published articles on current economic, political or religious topics or of broadcast works or other subject-matter of the same character, in cases where such use is not expressly reserved, and as long as the source, including the author's name, is indicated, or use of works or other subject-matter in connection with the reporting of current events, to the extent justified by the informatory purpose and as long as the source, including the author's name, is indicated, unless this turns out to be impossible;

(d) quotations for purposes such as criticism or review, provided that they relate to a work or other subject-matter which has already been lawfully made available to the public, that, unless this turns out to be impossible, the source, including the author's name, is indicated, and that their use is in accordance with fair practice, and to the extent required by the specific purpose;

(e) use for the purposes of public security or to ensure the proper performance or reporting of administrative, parliamentary or judicial proceedings;

(f) use of political speeches as well as extracts of public lectures or similar works or subject-matter to the extent justified by the informatory purpose and provided that the source, including the author's name, is indicated, except where this turns out to be impossible;

(g) use during religious celebrations or official celebrations organised by a public authority;

(h) use of works, such as works of architecture or sculpture, made to be located permanently in public places;

(i) incidental inclusion of a work or other subject-matter in other material;

(j) use for the purpose of advertising the public exhibition or sale of artistic works, to the extent necessary to promote the event, excluding any other commercial use;

(k) use for the purpose of caricature, parody or pastiche;

(l) use in connection with the demonstration or repair of equipment;

(m) use of an artistic work in the form of a building or a drawing or plan of a building for the purposes of reconstructing the building;

(n) use by communication or making available, for the purpose of research or private study, to individual members of the public by dedicated terminals on the premises of establishments referred to in paragraph 2(c) of works and other subject-matter not subject to purchase or licensing terms which are contained in their collections;

(o) use in certain other cases of minor importance where exceptions or limitations already exist under national law, provided that they only concern analogue uses and do not affect the free circulation of goods and services within the Community, without prejudice to the other exceptions and limitations contained in this Article.

4. Where the Member States may provide for an exception or limitation to the right of reproduction pursuant to paragraphs 2 and 3, they may provide similarly for an exception or limitation to the right of distribution as referred to in Article 4 to the extent justified by the purpose of the authorised act of reproduction.

XII Appendix B: Creative Common licenses and Potential application to Panacea workflows

We indicate the following types of CC licenses and the key functions they could perform:

(a) Creative Common Zero: This is a legal tool that may be construed as a waiver of copyright in the jurisdictions where this is allowed and as a license of all economic rights where waiver is not allowed. It is a good option, if the objective of the licensor is to maximize the reuse and mixing of different workflows. It makes sense only if there is no economic benefit from providing access to the workflow and the value derives from the collection of workflows and their constant reuse.

(b) Creative Commons Attribution: It allows all possible uses as long as the source of the workflow is attributed. It serves the same functions as CC Zero, only it obliges the re-user to attribute the maker of the original workflow and the practicality of this attribution has to be taken into account when deciding whether it is going to be used or not. This is a licence particularly popular with data providers and after the issuing of v.4.0 of the CC licences, where the sui generis database right is also going to be included as part of the licensed subject matter, is most likely to become the standard for governmental data.

(c) Creative Common Attribution ShareAlike: this is a copyleft license, i.e. a license which will allow the re-user of the workflow to make any amendment he/she wishes to make and further re-distribute these amended versions, but under the same terms and conditions as the original license. This is a license that is aims at creating communities that collectively improve workflows, very much in the fashion that Free/ Open Source Software communities operate. It is not necessarily as a popular license for commercial re-users as CC-Zero or CC-BY that allow re-use with the minimum restrictions and may be complemented with a commercial license, i.e. a license allowing the re-use of the workflow without an obligation to share it further under the same terms and conditions as the original work.

(d) Creative Commons Attribution NonCommercial ShareAlike: this copyleft license allows all non-commercial uses including the production of derivative workflows, but such derivative workflows will have to be further distributed under the same Non-commercial terms. In that sense it makes sense only of the workflow provider either has a specific reason for not allowing commercial use or there a dual licensing approach is followed where a license allowing commercial use if offered for a premium.

(e) Creative Commons Attribution NoDerivatives: This license allows only the redistribution of the workflow without allowing any changes to it. This would be a good license candidate if the objective is to maximize the use of workflow but would not like it to be altered by any third party.

(f) Creative Commons Attribution NonCommercial NoDerivatives: This license operates in the same way as the Creative Commons Attribution NoDerivatives but also prohibits commercial use. It is, hence, necessary that the licensor has a reason not to allow commercial uses or provides also a commercial license for a premium.