European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2015/1591 programme.

European Language Resource Coordination
Connecting Europe Facility

ELRC Report on legal issues in web crawling

Authors:	Pawel Kamocki (ELDA) Vladimir Popescu (ELDA)
Contributors:	Isabelle Gavanon (FIDAL) Camille Gaffiot (FIDAL) Khalid Choukri (ELDA) Valérie Mapelli (ELDA)
Reviser:	Mickaël Rigault (ELDA)
Dissemination Level:	public
Date:	2018-03-22
Revision Date:	2021-02-09
Version	1.1
Copyright:	**© 2018 ELRC**

Service contract no.	SMART 2015/1591
Project acronym	ELRC
Project full title	European Language Resource Coordination
Type of action	Service Contract
Coordinator	Prof. Josef van Genabith (DFKI)
Title	ELRC Report on legal issues in web crawling
Type	Report
Contributing partners	ELDA, Prodromos Tsiavos
Task leader	ELDA
EC project officer	Susan Fraser, Aleksandra Wesolowska

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

Prof. Stephan Busemann
DFKI GmbH
Stuhlsatzenhausweg 3
Campus D3_2
D-66123 Saarbrücken, Germany

stephan.busemann@dfki.de
Phone: +49 (681) 85775 5286
Fax: +49 (681) 85775 5338

Contents

1 Executive Summary

The purpose of this report is to analyze the question whether and under what conditions web crawling operations can be lawfully conducted.

It starts with a general overview of web crawling (Section 2), which briefly presents the procedure and discusses possible scenarios for which crawled data can be used. Then, it proceeds to the legal analysis of the problem (Section 3), which takes into account such legal frameworks as copyright (3.1), the *sui generis* database right (3.2), digital rights management (3.3) data protection (3.4), contract law (3.5) and conflict of laws (3.6). The analysis is focused on EU law (with the laws of Germany and France often quoted as examples), but some questions specific to the US law are also discussed. Section 4 discusses possible sanctions for unlawful web crawling.

The conclusion proposes a roadmap – a set of recommendations that should be taken into account before the start of any web crawling operation.

1.1 Introduction to web crawling

Web crawlers (also referred to as web harvesters) are pieces of software which browse the Internet in a methodic manner. A crawler creates a copy of every web page that it encounters and follows all the links that these web pages contain, sometimes limited to the links pointing to pages situated within the same web site. Such copies can then be used e.g. to build indexes for search engines, or to train Machine Translation engines and many other Language Technologies.

Once the data have been crawled, they can be further processed, depending on the intended use. Various scenarios are possible: archiving only (1), data analysis (2), exploitation (3), sharing (4) and distribution (5). Each of these scenarios is further analyzed taking into account two hypotheses regarding 1) the characteristics of the entity that re-uses the data (private vs. public) and 2) the purposes for which the data are re-used (commercial vs. non-commercial).

1.2 Legal analysis of web crawling

1.2.1 Copyright

In many cases data that are subject to crawling are protected by copyright. Some exceptions to this rule include e.g. official works (in some countries) or purely factual statements (such as e.g. train schedules or data concerning web traffic) which are supposedly copyright-free. In principle, copyright-protected content cannot be reproduced or communicated to the public without the permission of the right-holders, unless the use is covered by a statutory exception. In the EU law, four exceptions seem relevant for web crawling: temporary acts of reproduction, research exception, private copy and quotation. Unfortunately, these exceptions allow for web crawling only in very limited circumstances. This is the case when:

- the reproductions made in the process are temporary (which is of very limited relevance for crawling activities); OR
- crawling is carried out for non-commercial research purposes, and it meets all the criteria set forth in the national transposition of the research exception (national transpositions in various EU Member States may e.g. only allow reproduction of excerpts of works, require payment of equitable remuneration or only allow communication within a strictly limited circle of persons); OR
- crawling is carried out for strictly private purposes (in which case it may enter within the scope of the private copy exception) and not in the professional context.

A new exception for Text and Data Mining (TDM), expected to be included in the new Directive on Copyright in the Digital Single Market, may provide a greater relief for web crawling. In some countries, such as Germany, exceptions for TDM exist already, but they are limited to non-commercial research. Moreover, TDM exception may allow for reproductions of content to be made, but the possibilities to share such reproductions under these exceptions are very limited. It shall be noted that typically these exceptions require “lawful access” to the mined data.

In the United States, crawling seems to be more largely allowed under the doctrines of fair use and implied license. Whether a particular set of crawling operations can qualify as fair use would have to be evaluated on a case-by-case basis, taking into account the specific facts of each case.

1.2.2 *Sui generis* database right

Many websites can meet the legal definition of a database and be protected by the *sui generis* database right. Therefore, in principle extraction and re-use of substantial parts of such websites needs to be authorized by the rightholders (the maker of the database).

However, non-substantial parts of protected websites (typically less than 10%) can be freely extracted (copied) and re-used (shared). On the other hand, repeated and systematic extraction of such non-substantial parts is prohibited.

The existing research exception may allow extraction (reproduction) of substantial parts of protected websites for non-commercial research purposes (providing that the source is indicated). However, it needs to be checked whether (and how) the exception has been transposed in the applicable national law. Re-used (sharing) of substantial parts of databases for non-commercial research purposes is not allowed.

It is useful to keep in mind that websites produced by US-based companies are not covered by the *sui generis* database right. The websites produced by such companies, however, may obviously still be protected by copyright.

1.2.3 Digital Rights Management

Digital Rights Management (DRM) can be defined as technological protection measures that prevent or restrict various acts not authorized by the rightholders. Mere circumvention of such measures is subject to sanctions, even if it is not followed by acts of reproduction and/or communication to the public. It is therefore important to keep in mind that for crawling activities to be lawful, the crawlers should not attempt to circumvent DRMs (such as paywalls, password protection or captcha challenges).

1.2.4 Personal data

“Personal data” is a broad concept that covers any information related to a natural person (regardless of whether it concerns his or her private or professional sphere of activities). The information is to be regarded as personal data not only if it is directly identifying (e.g. contains names), but also if it can be used to indirectly identify the person. Therefore, when particular kinds of websites (such as discussion fora, social media or even online shops where users can post reviews) are crawled, personal data are likely to be collected in the process.

In principle, processing of personal data requires the data subject’s (i.e. the person that the data relate to) consent. Some exceptions are possible, e.g. when the processing passes a “balance of interests” test (taking into account reasonable expectations of the data subject), or if the data were made manifestly public by the data subject.

The General Data Protection Regulation (applicable since 25 May 2018) also establishes further rules e.g. with regards to data minimization (only necessary, adequate and relevant data can be processed) or storage limitation (data cannot be stored for longer than necessary).

In order to comply with these principles, crawling operations would have to be designed in such a way as to only collect the *'necessary'* amount of personal data. This is why data-intensive language technologies should focus on non-personal (or anonymized) data.

Moreover, even after consenting to the processing, data subjects have non-waivable rights in relation to their data (such as information, access and rectification), and data controllers and processors (i.e. persons or entities that define the purposes of processing) need to comply with complex obligations regarding organizational and technical means of processing (to implement data protection by design and by default, to carry out an impact assessment, to keep a register of processing operations, to notify breaches...). All these requirements are indeed difficult and costly to comply with.

It seems necessary, therefore, to resort to anonymization techniques. Ideally, data should be automatically anonymized already at the stage of their collection. The crawler should either omit personal data, or automatically anonymize them.

1.2.5 Contracts (Terms of Use, licenses, notices, waivers)

Most websites are available under conditions specified in a contractual instrument attached to the website (Terms of Use, public license). In principle, this instrument becomes binding once the website is accessed. The clauses of such contracts can roughly be divided into those that allow and those that prohibit crawling.

Regarding the first group of clauses, websites can lawfully be crawled if they are available under a public license (such as a Creative Commons license), providing that the conditions of the license are respected. Some notices may have effect similar to public licenses.

As far as the second category of contractual clauses is concerned, they can effectively prohibit any crawling (even allowed under a statutory copyright exception, unless this exception is expressly non-overridable by contractual clauses). However, the enforceability of such clauses may be doubtful (depending on the applicable law), especially if express acceptance of the instrument (ticking a box or clicking on a button) is not necessary to access the website.

1.3 Conclusion

The most viable way of making sure that the crawling operations are lawful is to perform an *a priori* clearance of the sources that are to be crawled. It shall be checked:

- whether the contents available via the list of URLs are protected by copyright and/or the *sui generis* database right;
- even protected, the content can still be crawled if it is available under a public license (such as Creative Commons) or with a notice that expressly allows crawling;
- if the contents are held by a public sector body, it is possible to request a license for their re-use (pursuant national rules on the re-use of Public Sector Information).

Only the sources that pass this validation procedure can be lawfully crawled. Even then, the data obtained in the process shall be anonymized before they are further processed.

2 Introduction to web crawling

2.1 Definition of web crawling

Web crawlers (also referred to as web harvesters) are pieces of software which browse the Internet in a methodic manner, starting from a set of seed Uniform Resource Locators (URLs). Typically, a crawler creates a copy of every web page that it encounters and follows all the links that these web pages contain, sometimes limited to the links pointing to pages situated within the same web site. Such copies can then be used e.g. to build indexes for search engines, or to train Machine Translation engines and many other Language Technologies (LT). In a way, crawling is somewhat similar to thorough web browsing by a human user (which also entails making of temporary reproductions necessary for visualizing contents in the browser). It may be seen as a cheap, fast and efficient way of obtaining data which can further be re-used for a very large variety of purposes. However, an essential difference between human browsing and crawling is that in the former case data are not persistently stored on the user's computer[1], while in the latter case data are stored on the user's computer persistent storage media (essentially, hard disk drives or solid-state drives).

2.2 Definition of crawled data

Digital data that can be obtained via web crawling (crawled data) are extremely varied. They include not only the web pages in their textual form (HTML), machine-readable documents (.doc, .pdf), images, audio and audiovisual recordings, but also information about the crawling process per se (timestamp of the crawling, particular URL for each page, page redirections, and HTTP status codes which give information about the server configuration). In order to speed up the crawling process, some categories of data can be filtered *a priori* (e.g. when the web crawler is programmed not to fetch images or certain categories of websites) or *a posteriori* (i.e. when some data are automatically deleted shortly after they are recorded).

2.3 Crawling processing stages

Irrespective of the use made of the crawled data, several essential processing stages take place:

1. A set of seed URLs (which may be defined by a human operator) is fed into the crawler. Usually, each seed URL represents the first page to be fetched by the crawler (e.g. the index page of a web site);

2. The crawler is launched starting from each seed URL and fetches the pages accessible *via* all the URLs present on the page accessible from the seed URL. This process is repeated for each page thus visited, until either a certain timespan has passed, or until a certain crawling *depth* has been reached, or until all pages accessible through the URLs in the same web domain as the seed URL have been fetched;

3. As the crawler fetches a web page accessible through an URL, it can either: (i) store the web page as it is (i.e. with all HTML contents, CSS, JavaScript), (ii) ignore certain parts of the site and do not fetch them (e.g. CSS, JavaScript, images);

[1] Note however that sometimes web browsers are configured in such a way that, in order to optimize web page access time and bandwidth, they store the contents of the visited page in a cache, which is, putatively, invalidated upon updates of the visited web pages, or after a certain period of time.

4. Once the content has been downloaded and stored, the crawler can perform several additional processing stages: (i) document[2] format transformation (e.g. from HTML to specific XML in order to facilitate further processing); (ii) HTML tags stripping so that only text is kept;

5. In parallel, the crawler can store several crawling and connection-specific metadata for logging purposes, viz. crawling speed[3], crawling time, URL redirections, HTTP status and error codes indicating timeout, unauthorized access, internal server error etc.

2.4 Potential applications of crawled data for language technology

Crawled data is mostly useful for training or evaluating statistical natural language processing systems that involve some form of text manipulation. For example, machine translation engines usually need three types of data:

(i) terminologies in the source language, in order to enhance the system robustness with respect to the input text, by finding substitutable expressions (synonyms are an extreme case);

(ii) phrase-aligned bilingual texts in order to perform the actual translation from the source language to the target language;

(iii) word n-gram counts in the target language in order to smooth the machine translation output and to prune less likely outputs.

Crawled data can help mostly with (ii) and (iii). For (ii), data crawled from multilingual web sites (i.e. web sites that have been internationalised) can be further aligned, first at the document level, based on some heuristics pertaining to the paragraph structure of the pages, to the URLs contained in the pages, and to the URLs that point to these pages, and then at the sentence level, by using specialised tools called sentence aligners.

For (iii), data crawled from monolingual web sites can be cleaned, curated and fed into an n-gram count estimator tool, so that, in the end, one obtains tables of word co-occurrence frequencies.

Another example can be automatic speech recognition, which also includes a natural language modelling stage akin to stage (iii) presented above. In this latter case, word n-gram counts serve the purpose of pruning the speech recognition decoding hypotheses so as to better match most likely occurring phrases.

2.5 Various processing scenarios

Once the data have been crawled, they can be further processed, depending on the intended use. Various scenarios are possible:

1. Archiving only. The data are preserved in their original form and archived for further re-use for historic or research purposes (web archiving). In this case the entire contents of the web sites are kept, including images, multi-media files, CSS stylesheets, JavaScript files, etc. The data is thus a verbatim copy of the crawled web site.

2. Data analysis. This usage, normally starting with data cleaning and curation (stripping off of HTML tags, discarding of ill-encoded text, etc.) can take various forms, going from simply collecting several statistics (word frequency, presence / absence or frequency of specified phrases, etc.), through adding linguistic information in a homogeneous way (regarding

[2] By “document” we understand the structured contents of each web page.

[3] If the bandwidth characteristics of the crawling agent are known, this metric gives relevant information on the connection particulars of the crawled server.

morphology, syntax, sentiment, discourse, etc.), to performing higher-level tasks (computing the overall subjectivity stance of the texts at hand, summarizing them, etc.).

3. Exploitation. In this scenario data, either primary (i.e. without annotations) or annotated as a result of the analysis specified at 2., can be used either directly, or subsequent to further processing, for training statistical language processing software tool(kit)s. For example, crawled and curated text data can be used for computing word n-gram counts, which are useful for smoothing speech recognition results, or for smoothing machine translation outputs. Or, the document-aligned data resulting from crawling multilingual web sites can be further aligned at phrase level and used for training machine translation engines. Another example of exploitation can also be the evaluation of automatic language processing systems, e.g. the output of a machine translation or speech recognition system is compared to reference text resulted from crawling, or annotated curated crawled data can be used as test data for e.g. automatic morphologic analyzers.

4. Sharing. In this scenario, data are shared, either in primary form (e.g. crawled and curated), or in analyzed form (annotated, either at a fine-grained level e.g. with per-word token morphological information, or at a coarse-grained level, e.g. with subjective information at the phrase or document level). This sharing process can take place either within the same entity (e.g. from the department that has performed the crawling, hence producing the data, to another department within the same organization), or with a limited group of entities. The usages can consist simply in evaluating existing software tools, or in using such tools in order to *derive* new content from the shared data via the tools, e.g. using a machine translation system trained with the shared data in order to produce new translations of new data, or using a speech recognition system whose language model is trained with crawled data, in order to transcribe further audio data.

5. Distribution. In this scenario, third parties gain access to digital copies of the data (organized in a dataset), for free or for a fee. These third parties can be either private or public. On the other hand, the usages of the distributed resources (according to what has been stated at 3.), can be for commercial purposes or for non-commercial purposes. As at point 4, the usages can involve deriving new data from the distributed data, or not.

Each of these scenarios should further be analyzed taking into account two hypotheses regarding 1) the characteristics of the entity that re-uses the data (private vs. public) and 2) the purposes for which the data are re-used (commercial vs. non-commercial).

3 Legal analysis of web crawling

3.1 Copyright

Definition and sources. Copyright is a form of Intellectual Property protecting *original works*. Copyright laws in different countries around the world share many similarities: it is so because the minimum standard for copyright protection is defined in international conventions, such as the Berne Convention (1886), the Agreement on Trade-Related Aspects of Intellectual Property Rights (1994), or the WIPO Copyright Treaty (1996). In the European Union, many aspects of copyright law are harmonized by the *Directive 2001/29/CE of 22 May 2001 on the harmonization of certain aspects of copyright and related rights in the information society* (hereinafter: the InfoSoc Directive), which establishes a minimum EU-wide standard for copyright protection and has been transposed in all the EU Member States*.* It is important to note that directives do not apply directly to the situation of EU citizens; instead, they need to be transposed into national law of every Member State. Therefore, in court an individual cannot rely directly on a directive, but only on its national transposition.

3.1.1 Scope of protection

Original works. Copyright protects the form of expression of *original works*. According to the definition adopted by the Court of Justice of the European Union a work[4] is original if it constitutes *"its author's own intellectual creation".* This means that in the process of creating the work, the author made some creative choices (of words, colors, light, sounds etc.) and by doing so marked the work with his personal imprint (i.e. no one else is likely to make the exact same choices and therefore create the exact same work). The genre of a work (written, graphic, audiovisual) is irrelevant, and so are its (artistic) quality and its character (although in some countries official works are excluded from copyright protection – cf. below) or its length (even titles and slogans can be protected by copyright if they are original). The excerpts of works can also be protected if they are original; the Court of Justice of the European Union (CJEU) ruled that snippets as short as 11 words can meet the originality criterion[5]. In our view, even shorter excerpts can attract copyright protection (although the shorter the excerpt, the harder it will be for it to meet the originality criterion).

Adaptations (including translations). Apart from original works, copyright may also protect original adaptations of other works (called *"derivative works"*). This broad category includes e.g. updated or extended versions of works, transpositions to another medium or technique (e.g. a movie based on a book) and translations. Derivative works are protected *"without prejudice to the original work"*. This means that exploitation of such works requires authorization both from the author of the adaptation and from the author of the original work.

Original databases and compilations. Databases and other compilations of works can also be protected by copyright if they are original in their selection and arrangement. This protection is limited to the 'envelope' and is independent from whether the elements that constitute the database are themselves protected by copyright. It means that e.g. a collection of public domain works (e.g. 19th century love poems) can be protected by copyright if the works are selected and arranged according to original (i.e. subjective) criteria; individual works included in the collection are still in the public domain, but exploitation of the whole collection is only

[4] Which can be roughly defined as any human creation (including software).

[5] This does not mean, however, that all 11-words-long snippets will indeed be original, but only that their originality cannot be ruled out because of their length; accordingly, shorter snippets can also be protected by copyright if they meet the originality criterion.

possible with the authorization of the rightholder. This protection should be clearly distinguished from the one granted by the *sui generis* right (see below).

"Sweat of the brow". Traditionally, in England copyright protection could have been obtained for *"industrious collection"* or *"sweat of the brow"*; in other words, copyright used to protect any result of hard work, even if it did not meet the criteria of originality. This was of particular relevance for compilations of facts, such as phone books, which – despite being unoriginal – could still be protected by copyright under the *"sweat of the brow"* doctrine. In the United States (which largely inherited English copyright laws from the colonial era), this doctrine was abolished by the Supreme Court in 1991[6]; in England however, this never really happened. It could be argued that *"sweat of the brow"* is not compatible with EU law, but after Brexit, it is not entirely unlikely that the UK will return to this traditional doctrine which considerably broadens the scope of copyright protection.

Exclusions. In the light of the above, raw facts (such as dates, prices, measurements) are not protected by copyright because they do not meet the criteria for protection (they exist objectively and therefore cannot result from human creativity). Moreover, in some countries certain categories of works, even if they seem to meet the originality criterion, are expressly excluded from copyright protection. This is the case of official works (such as texts of statutes and their official translations, administrative and court decisions etc.) in countries such as United States, Germany, Poland, Czech Republic, Spain, Italy and to a certain extent also France. In these countries such works (which are a part of what is referred to as Public Sector Information -- cf. below) are in the public domain and can be freely re-used.

> ***Taking into account the scope of the report, it should be assumed that in most cases data that are subject to crawling are protected by copyright. Websites normally contain copyright-protected works; moreover, their collection and arrangement can be protected as an original compilation. Some rare exceptions to this rule include e.g. official works (in some countries) or purely factual statements (such as e.g. train schedules or data concerning web traffic) which are copyright-free.***

3.1.2 Term of protection

Principle. In most countries (including all EU Member States and the US) copyright protection expires seventy years after the death of the author. In some other countries (e.g. in Canada) this term may be shorter, but no shorter than fifty years after the death of the author. Moreover, the exact term of protection may vary slightly depending on various factors[7]. Nevertheless, given that the Internet is a relative novelty, it should be assumed that the content of most websites is still in copyright. Exceptions may include e.g. websites with public domain literature (such as the Gutenberg project), which may nevertheless still be protected by copyright as compilations (if their selection and arrangement are original) or by the *sui generis* database right (see below).

3.1.3 Copyright ownership

Principle. In principle, copyright belongs to the author (or authors) of original works. It does not require any form of registration (unlike e.g. patents or trademarks). In some countries (such as e.g. the US, the UK or Ireland) copyright in works created by an employee in the course of

[6] Feist Publications, Inc., v. Rural Telephone Service Co., 499 U.S. 340 (1991).
[7] e.g. in the US works for hire (see below) receive copyright protection until 120 years after creation or 95 years after publication, whichever comes first; in Ireland, government copyright is limited to 50 years after publication…

employment belongs from the start (*ab initio*) to the employer (work for hire); in other countries (e.g. in France) this only concerns software or works created by civil servants.

Subsequent transfer. Just like 'regular' property, copyright can be subsequently transferred[8] by the rightholder e.g. to a publisher or a client (e.g. copyright in a website or a translation created by a freelancer).

3.1.4 Exclusive rights

Introduction. Copyright is a bundle of exclusive rights – this means that certain acts concerning copyright-protected works may in principle only be performed by the rightholder, or with his permission.

There are slight variations in how various countries define the exclusive rights of copyright holders; nevertheless, from the point of view of this report, two types of universally recognized prerogatives are relevant: the exclusive rights of reproduction and of communication to the public.

Right of access? Formally, access to works, if it is not accompanied by reproduction or communication to the public (i.e. reading literary works, looking at paintings…) is not an exclusive right and therefore does not require authorization. However, legal protection of Digital Rights Management (technological protection measures) against circumvention (see below) may have the practical effect of efficiently preventing unauthorized access to copyright-protected works.

Right of adaptation? The right of adaptation (i.e. the right to make adaptations or translations of copyright-protected works) is not harmonized by the InfoSoc Directive; nevertheless, it can be found in national laws of many Member States (e.g. in Germany[9], in France[10], as well as in the United States[11]). It should be noted, however, that adaptations will often require reproductions of elements of the original work (and therefore fall within the scope of the exclusive right of reproduction). Moreover, the international principle according to which copyright in adaptations is without prejudice to copyright in original works (so e.g. publication of a translation would require permission from the holder of copyright in the original work) remains unchanged. What may change, however, is whether the very *making* of adaptations/modifications/translations (even if they are not published or otherwise communicated to the public, e.g. a translation made for internal use only) can be prohibited by the author.

> ***There is a distinction between primary and secondary copyright infringement. Primary infringement consists of performing acts covered by exclusive rights without the rightholder's permission (e.g. a painter that copies a copyright-protected painting). Secondary infringement can be committed by someone who knowingly takes advantage of primary infringement (e.g. distributes illegal copies of works).***

[8] It is true that in some countries (such as Germany and Austria) copyright is not transferable; even there, however, an exclusive license can be granted, very much to the same effect as copyright transfer.
[9] Section 23 of the German Copyright Act.
[10] Art. L. 122-4 of the French Intellectual Property Code.
[11] Section 106(2) of the US Copyright Act.

3.1.4.1 Reproduction

Definition. Reproduction right is defined as: *"an exclusive right to authorise or prohibit direct or indirect, temporary or permanent reproduction [of a work] by any means and in any form, in whole or in part"*[12]. It shall be noted that the use of works in digital form will in principle always require at least a temporary and partial reproduction (cf. below about the exception for temporary reproductions). The activity of web crawling consists essentially of making reproductions of web content. Moreover, if the crawled websites contain unlicensed content (i.e. content uploaded without the rightholder's permission), the person that conducts crawling activities may also be liable for secondary infringement.

3.1.4.2 Communication to the public

Definition. The right of communication to the public is an *"exclusive right to authorise or prohibit any communication [of works] to the public, by wire or wireless means, including the making available to the public of their works in such a way that members of the public may access them from a place and at a time individually chosen by them [i.e. uploading on the Internet]"*.[13]

Notion of a public. It shall be noted that only communication of works to *a public* can be prohibited by rightholders. This leaves some leeway for interpretation. It is widely admitted that communication of works within a circle of *"family and friends"* does not constitute communication to the public. It is also unclear whether communication to an individual person (even from outside the circle of *"family and friends"*) constitutes communication to the public.

Also, communication of a work within a company is not communication to the public (unless it serves to attract customers, e.g. music played in a pub).

Recent interpretation by the CJEU. In the last couple of years, the right of communication to the public (article 3(1) of the InfoSoc Directive) has been subject to detailed interpretation by the CJEU (Court of Justice of the European Union). In particular, it has been ruled that:

- in principle, linking[14] to web content (and framing[15] of such content) does not amount to communication to the public, as there is no new public;
- however, this does not apply when the link leads to unlicensed content (i.e. a content that was uploaded without the permission from rightholders, e.g. a recent blockbuster movie) and the provider of the link is aware of this; the knowledge of the unlicensed character of the content is presumed when the link is provided with a profit-making intention[16] (so e.g. when a link to unlicensed content is on a website containing commercials, the owner of the website would have to prove that he was not aware of its unlicensed character);
- also, links that circumvent access restrictions[17] or facilitate access[18] to unlicensed content are acts of communication to the public;

[12] Art. 2 of the Directive 2001/29/EC.

[13] Art. 3(1) of the Directive 2001/29/EC.

[14] CJEU, case C-466/12 (Svensson)

[15] CJEU, case C-348/13 (Bestwater)

[16] CJEU, case C-160/15 (GS Media)

[17] idem

[18] CJEU, case C-527/15 (Filmpeler); CJEU, case C-610/15 (Ziggo)

> ***It appears therefore that those who share crawled data (which may contain links to unlicensed material), especially with a profit-making intention, may be liable for secondary infringement of copyright.***

3.1.5 Copyright Exceptions in the European Union

Introduction. In order to strike balance between the interests of rightholders and those of users, legislators provide for copyright exceptions. Users that perform acts covered by exclusive rights cannot be liable for copyright infringement if they meet the criteria of a copyright exception.

Copyright exceptions under EU law. Art. 5 of the InfoSoc Directive provides a limitative list of copyright exceptions; however, only one of these exceptions (for temporary acts of reproduction -- see below) is mandatory. The remaining exceptions are optional, e.g. they do not need to be transposed in national laws of the Member States, or can be narrowed down in the transposition process[19]. Therefore, there are important differences between the scope of optional copyright exceptions in various EU Member States. On the other hand, national legislators cannot adopt copyright exceptions that are not included in art. 5 of the InfoSoc Directive.

Relevant exceptions. Four exceptions seem relevant for web crawling: temporary acts of reproduction (1), quotation (3), private copy (4) and the exception for research (2). It shall be noted that due to the optional nature of most of these exceptions, their implementation in various Member States may vary substantially; since it would be impossible to compare the laws of all the Member States, our analysis is in principle limited to German and French law, although references to other countries will also be made on several occasions.

3.1.5.1 Temporary acts of reproduction

Mandatory exception. The exception for temporary acts of reproduction is the only mandatory exception in the InfoSoc Directive. This means that it can be found (in more or less the same form) in national laws of every Member State.

Purpose. As explained in the recital 33 of the InfoSoc Directive, the main purpose of this exception is to allow web browsing without the necessity to obtain a permission to view every website.

Source. The source of the exception is art. 5(1) of the InfoSoc Directive:

Temporary acts of reproduction (...), 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 (...).

Five elements. The exception can therefore be analysed into five conditions which all need to be met:

[19] It is important to keep in mind that individuals cannot rely directly on directives, so non-transposed exceptions are of no practical use.

- The reproduction needs to be temporary. This means that copies cannot be permanent (intended for long-term preservation). This rules out most web crawling activities, the purpose of which is to create permanent copies of web content. However, for the sake of completeness, and because in our view it is in some cases possible to structure the web crawling process in such a way as to make it compatible with the exception, we present below the four remaining conditions.

- The reproduction needs to be transient or incidental. A reproduction is transient *"if its duration is limited to what is necessary for that process* [that it is an integral part of] *to work properly, it being understood that* [the reproduction shall be deleted] *automatically, without human intervention, once its function of enabling the completion of such a process has come to an end"*[20]. However, for the exception to apply the reproduction does not have to be "transient" -- it can be just "incidental" which means that it *"neither exists independently of, nor has a purpose independent of, the technological process of which it forms part"*[21].It has to be automatically deleted "after a certain time"[22]. The CJEU ruled that cache copies (even though they can be stored for many months) are "incidental"[23].
- The reproduction needs to be an integral and essential part of a technological process. This condition is met when reproduction is carried out entirely in the context of the implementation of a technological process for which it is necessary, i.e. the technological process could not function correctly and efficiently without it[24]. Moreover, it has been ruled that while the reproduction has to be deleted automatically upon completion of the process (see above), the process itself can be activated manually[25].
- The sole purpose of the reproduction needs to be a lawful use of the work[26]. This condition is interpreted rather broadly by the CJEU. In short, a use if lawful if it is not restricted by European or national law. For example, the CJEU ruled that reception of encrypted TV broadcasts in a private circle is a lawful use (because a private circle is not a "public"), and so is drafting of summaries of newspaper articles[27]. On the other hand, streaming of unlicensed video materials is not a lawful use, and therefore it does not enter within the scope of the exception[28]. It seems that the creation of statistical language models would also qualify as lawful use[29].
- The reproduction cannot have independent economic significance. The last condition is similar to the third one; in order for this condition to be met, the reproduction has to be inseparable from the technological process that it is part of. Moreover, it has been ruled that the condition is not met when the process involves modifications of the

[20] CJEU, case C-5/08 (Infopaq), para 64.
[21] CJEU, case C-360/13 (Meltwater), para 43.
[22] *Meltwater*, para 26.
[23] *Meltwater*
[24] CJEU, case C-302/10 (Infopaq II)
[25] *Infopaq II*, para 32
[26] Or "a transmission in a network between third parties by an intermediary", which is irrelevant for our study.
[27] *Infopaq II*, para 42-45; it shall be noted, however, that this concerned Danish law in which there is no exclusive right of adaptation; if the case concerned a country that recognized such an exclusive right (e.g. Germany) the outcome may have been different (see above).
[28] *Filmpeler*
[29] Of all the scenarios discussed by the CJEU, creation of statistical language models seems to be the closest to drafting summaries of news articles.

reproduced work[30], which (arguably) precludes any form of annotation of the reproduced data. In most cases, reproductions made in the process of web crawling will not meet this criterion.

The Meltwater (UK) case. An interesting case decided in 2013 by the UK Supreme Court[31] concerned the use of crawling and scraping of news websites by the operator of a paid news aggregator service that allowed its users to read extracts of the crawled articles. The court ruled that the copying activities were covered by the exception for temporary acts of reproduction because they served a lawful purpose (transmission of copyrighted works in a network) and because they were temporary (they were automatically deleted after a certain time). Moreover, it was ruled that the users of the service did not need a license to view the articles, as mere reading of copyright-protected content is not within the ambit of exclusive rights. It shall be noted that if the users wanted to perform other acts on the reproductions (other than mere reading, e.g. data analysis), they would probably need a license to do so and so the decision would have been different.

> ***Typically, reproductions made in the process of web crawling are not temporary, and they are intended to be used outside of the crawling process (which means that they have independent economic significance). Therefore, they cannot be covered by the discussed exception****.*

However, in our opinion it is not impossible to organize the crawling process in such a way as to comply with the conditions of the exception. This would be the case if:

- the process is launched manually (which is allowed by CJEU case law);
- web contents are automatically reproduced and analysed (but without being modified) in order e.g. to derive a statistical language model (which, arguably, would constitute a "lawful use");
- the data are not used for any other purpose; no sharing is possible;
- upon completion of the process, the reproduced data are automatically deleted (but then can be reproduced again from the same source, as long as it is available).

It goes without saying that such a process would not be optimal, particularly from the point of view of reproducibility of results. Therefore, it shall be concluded that the discussed exception is of very little (if any) significance for web crawling activities.

3.1.5.2 Research exception

a) General research exception

Source. According to art. 5(3) of the InfoSoc Directive: *"Member States may provide for exceptions (...) to the rights [of reproduction and communication to the public] [for] use for the sole purpose of (...)*[32] *scientific research, as long as the source (...) is indicated (...) and to the extent justified by the non-commercial purpose to be achieved"*.

[30] *Infopaq II*, para 54.

[31] Public Relations Consultants Association v The Newspaper Licensing Agency Ltd ([2013] UKSC 18)

[32] The full text mentions "*illustration for teaching or scientific research*"; this expression is syntactically ambiguous and is interpreted differently in different language versions. In our opinion (based on the English version of the Directive) the word "illustration" refers only to "teaching" and not to "scientific research". However, e.g. in the French version the word "illustration" refers clearly to both "teaching" and "scientific research" (*…à des fins exclusives d'illustration dans le cadre de l'enseignement et de la recherche…).* This considerably narrows down the scope of the exception; according to the second interpretation ("use for the sole purpose of illustration for (...) scientific research"), the exception seems completely useless for web crawling activities (because crawled data are not used for illustration, but rather as a foundation for research).

Necessary elements. The Directive only allows for exceptions for uses that meet three criteria:

- research has to be the sole purpose of reproduction and/or communication to the public (reproductions cannot be used for any other purpose);
- research has to be non-commercial; the interpretation of this condition is particularly problematic. It seems that research is commercial if it is directed towards monetary or economic advantage[33]; public research activities (carried out at universities and public institutions) will probably be qualified as non-commercial in most cases outside public-private partnerships. Activities aimed e.g. at developing a Machine Translation system for commercial use (for a client) or to reduce operating costs of a company (e. g. to replace an external service) shall, in our view, be regarded as commercial and therefore falling outside the scope of the exception;
- the source (including the name of the author) has to be indicated (unless this is impossible). In the case of crawled data this condition seems easy to meet; however, mentioning the name of the author, i.e. a physical person that created the work, may require some additional efforts. The InfoSoc Directive stipulates that the requirement does not apply if its fulfillment is impossible, e.g. the name of the author is not indicated on the website (some national transpositions, however, seem less lenient in this respect);
- the acts of reproduction or communication to the public shall be justified by the purpose. Arguably, this does not imply a strict necessity test (*sine que non*, i.e. the purpose cannot be achieved without these specific actions), but only require that the acts be useful or relevant to achieve the purpose. In other words, a use is “justified by the purpose” not only if it is strictly necessary, but also if it is useful, i.e. it facilitates the achievement of the desired result. Still, the application of this requirement to data-intensive research will always be controversial as it is difficult (especially *a priori*) to draw a line between useful and useless data.

National implementations. As mentioned above, national legislators often implement the research exception in a narrow way. In particular, national implementations may contain additional requirements:

- regarding the parts of works that can be used; e.g. in German law in most cases only excerpts of works can be used (up to 15% of a work can be reproduced and communicated to the public; if there is no communication to the public, up to 75% of a work can be reproduced); this is particularly problematic from the point of view of web crawling,
- regarding the public to which the work can be communicated (which will often be limited, like in France or in Germany, to a group of people directly concerned by the research activities, e.g. members of one research team);
- regarding the beneficiaries of the exception, which may be limited to research institutions (e.g. in Poland or in Austria);
- regarding remuneration: German and French law both require that equitable remuneration be paid to collecting societies for the uses allowed by the exception (the money are then redistributed among authors). The amounts may not be prohibitively high (although in the case of web crawling they may become so if they are calculated on a “per work” basis), but specific agreements need to be negotiated between users and collecting society, which may be time-consuming.

[33] Cf. Creative Commons licenses and tools the definition of commercial use in CC licenses.

Nevertheless, some countries (such as Estonia or the UK) have relatively broad research exceptions (to the extent allowed by the Directive). In such countries, web crawling for non-commercial research may not infringe copyright. In most countries, however, this will not be the case. In order to enter within the scope of narrowly construed research exceptions, the crawling process would have to be modelled in a specific way (e.g. copying only 75% of contents of every web page?). In no case can commercial use be allowed by the exception.

> ***In sum, the traditional research exceptions do not provide much relief for web crawling activities. However, this may change if new exceptions are adopted, such as those for text and data mining purposes.***

a) New data mining exception

Context. In recent years, copyright aspects of data mining have been extensively discussed in various fora. This led some national legislators (most notably in the UK, in France and in Germany) to adopt specific exceptions for such activities (covering not only “data mining” in the strict sense, but also other forms of digital data analysis, such as those necessary for the creation of statistical language models). It shall be noted, however, that these exceptions cannot exceed the research exception in the InfoSoc Directive. National TDM exceptions are relatively new and have rarely (if ever) been tested in court.

In the UK. Section 29A of the Copyright, Designs and Patents Act allows a user having lawful access to a work to make digital reproductions thereof for the purposes of data mining for non-commercial research. No communication to the public (or any other form of “subsequent dealing” with the copies) is allowed.

In Germany. Section 60d of the German Copyright Act (which entered into force in March 2018) allows users to make reproductions of works in order to compile a corpus to be used for data mining for non-commercial research purposes. Necessary modifications of works are also allowed; on the other hand, the source needs to be indicated. The reproductions can be shared with a strictly limited public (e.g. members of the same research team), or communicated to the public for research evaluation purposes. However, at the end of the research project, reproductions must be deleted or transferred to a library or an archive for long-term storage. Moreover, equitable remuneration needs to be paid to a collecting society (cf. above about general research exceptions).

In France. In France, a data mining exception was adopted in 2016. Its wording is very unclear (hopefully to be clarified by a soon-to-be adopted decree or collective agreement), however it seems that it only allows for mining of scientific articles that the user has lawful access to. Therefore, it is not useful for web crawling activities.

Lawful access. The abovementioned exceptions require lawful access to the work[34]. This requirement seems relatively straightforward -- a user has lawful access to a work if he has the “right to read” it without circumventing any legal or technical protection measures. Therefore, a user with lawful access to the Internet has lawful access to the content that is openly available on this network (i.e. without a paywall or password protection), or at least to everything that was lawfully uploaded (i.e. by the rightholder or with his consent)[35]. Nevertheless, it can also be argued that access is lawful only if it is expressly authorized by

[34] In English and French law the requirement of lawful access is explicit; in German law it does not appear, but can be implied (see CJEU C-435/12 (ACI Adam) which seems to infer this requirement from the three-step test (art. 5(5) of the InfoSoc Directive)).

[35] It may be argued that otherwise (e.g. a poem was published online without its author’s permission) there is no lawful access and the user may still be liable for secondary infringement.

the rightholder; this interpretation can, in our opinion, hardly be defended, especially given that in principle the rightholder has no exclusive right to access (i.e. to read) his work (cf. the Meltwater (UK) case mentioned above). Therefore, in our view, only the first interpretation is the right one.

A mandatory EU-wide data mining exception? The Directive on copyright in the digital single market (DSM), expected to be adopted in 2018, will probably contain a mandatory data mining exception whose scope will go beyond the current research exception (i.e. allowed data mining also for commercial research purposes, or even for purposes other than research). For now, however, the content of this exception is impossible to predict. Typically, the deadline for implementation of a directive is 2 years after its adoption, so the new DSM directive is unlikely to take effect before 2020.

> ***For now, it is hard to say to what extent data mining exceptions can allow web crawling for non-commercial research purposes. It is possible, however, especially in Germany, that once the amount of equitable remuneration is fixed, the new data mining exception will provide much relief for those carrying out web crawling activities.***

3.1.5.3 Quotation

Relevance for web crawling. *Prima facie*, the quotation exception does not seem relevant from the point of view of web crawling activities. However, given that crawling consists of making reproductions of web content, it can be assimilated (from the point of view of copyright law) to making quotations. It is therefore useful to discuss the quotation exception in this report.

Source. According to article 5.2 d) of the InfoSoc Directive, *"Member States may provide for exceptions (...) to the reproduction right [for] quotations for purposes such as criticism or review, provided that they relate to a work (...) which has already been lawfully made available to the public, that (...) 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"*.

Minimum requirements. The analysis of the text reveals that quotations in order to be lawful need to meet at least the following requirements:

- they need to be justified by purposes "*such as* criticism or review"; this short list is by no means limitative (as the words "such as" clearly indicate), and other similar purposes (including research) are also possible;
- only lawfully published works can be quoted (this would exclude e.g. an original love poem published on Facebook by the addressee without permission of the author);
- the source, including the author's name, needs to be indicated; the kind of information about the source that need to be provided (according to fair practice) differs between domains (e.g. in a scientific publication it is customary to mention the name of the author, the title of the quoted work, its specific edition and its publisher, year of publication etc., in other contexts, mentioning the name of the author may be sufficient).

Additional requirements in national laws of certain Member States. Just like in the case of the research exception, in most national laws the quotation exception is narrower than allowed by the InfoSoc Directive. Additional requirements may include:

- the length of quotations; e.g. in France it is not allowed to quote entire works (even if they are short), while e.g. in Germany entire works can be quoted (if it's justified by the purpose and character of the work);

- the “substitutability”; in other words, a quotation cannot be a substitute for the original work (in which case the market value of the original work would be significantly harmed -- cf. about fair use below); this is an explicit condition in many jurisdictions (it can also be inferred from compliance with “fair practice” required by the Directive); it is expressly mentioned in Italian law;
- originality of the quoting work; in other words, for quotation to be allowed, the quoting work should meet the criterion of originality, even if the quotations are removed from it. In countries that include this condition, simple compilations of quotations (e.g. with no added commentary) are not allowed.

This last condition makes the quotation exception quite useless for web crawling activities, as crawled data are normally not included in original works (no original or creative commentary is added). However, the CJEU recently ruled[36] that EU law does not require that the quoting work be original; therefore, national legislators are allowed not to include this condition in national copyright laws. Nevertheless, it seems to us that this requirement is present in most countries (Slovakia being a notable exception), and therefore the quotation exception does not allow simple compilation of (excerpts of) copyright-protected works. Even without this requirement, it is doubtful whether the quotation exception can provide any relief for web crawling activities.

3.1.5.4 Private copy

Source. Art. 5(2) b) allows Member States to provide for exceptions to the reproduction right *“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 (...)”*. The fair compensation is paid via a special tax (private copy levy) on blank supports (paper, blank CDs etc.) and devices that can be used for copying, which is then distributed among rightholders by collective management societies.

Meaning. The private copy exception allows individual users (only natural persons, i.e. no legal persons (companies)) to make reproductions of works that they have lawful access to (e.g. that are publicly accessible on the Internet) for personal and non-commercial use only. This seems to exclude any use of such reproductions in professional context. The reproductions cannot be communicated to the public.

Significance for web crawling. The private copy exception may allow web crawling, but only for strictly personal and non-commercial purposes. Arguably, web crawling by ‘citizen scientists’ for purposes of their own personal research can be allowed under this exception (especially if it is limited just to the first scenario described in above of this report, i.e. “archiving only”). However, given that communication of private copies to the public or any use of such copies in professional context (e.g. by professional researchers in language technology) is prohibited, the significance of this exception for web crawling is very limited.

3.1.6 Fair use in the United States

Introduction. The United States traditionally adopts a different approach to copyright limitations and exceptions than continental Europe. The anglo-saxon doctrine of *fair use* allows for much more flexibility in deciding whether specific uses should be exempted from the obligation to obtain permission from rightholders.

Source. Fair use is codified in section 107 of the United States Code, according to which:

[36] CJEU, case C-145/10 (Painer)

“the fair use of a copyrighted work (...) for purposes such as criticism, comment, news reporting, teaching (...), scholarship, or research, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include—

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

(2) the nature of the copyrighted work;

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

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

Four factors. There is a lot of leeway in the interpretation of the four factors of fair use. However, some general principles can be identified:

- the first factor (purpose and character of use) takes into account whether the work is used for a transformative (i.e. new, original) or a derivative purpose, i.e. whether the use generates an added value for the society. This is arguably the most important factor in fair use cases concerning Internet uses. Most notably, it has been ruled that image search engines use indexed images for a transformative purpose (to provide information)[37];
- commercial purposes do not preclude fair use (e.g. a parody of a popular song recorded on a commercial album can still qualify as fair use[38]);
- uses of non-fiction works (e.g. news articles) are more likely to be qualified as fair (in a sense, the protection of fiction works is therefore stronger);
- there is no quantitative test (e.g. maximum length of snippets) to evaluate whether a given use is fair; this has to be done strictly on a case-by-case basis; on one hand, quotation of several paragraphs of a book can be infringing (especially if the excerpt can serve as a “substitute” for the whole book – cf. above about the quotation exception), on the other -- in some cases the use of entire works can be fair[39];
- arguably, the fourth factor (which can be derived from the three other factors) is generally the most important one in fair use cases.

Flexibility of fair use. Section 107 of the United States Code is just a guideline; in deciding fair use cases, judges are free to weigh the four factors as they see fit, and to take other factors into consideration[40]. This provides for great flexibility, but also for lack of legal certainty: apart from few obvious cases (e.g. use of snippets of text for classroom teaching) it is impossible to predict whether a given use can be qualified as fair -- only arguments for and against can be provided, which then would have to be weighed by a judge. Given high costs of litigation, many (potentially very informative) fair use cases are settled out of court. Even Google tried for many years to settle their dispute with the Authors’ Guild (see below) out of court, even though at the end it turned out they had a good case for fair use.

> ***Therefore, it cannot be said that web crawling is generally allowed under fair use; instead, everything depends on the circumstances of every specific crawling activity.***

[37] Kelly v. Arriba Soft Corporation, 336 F.3d 811 (9th Cir. 2003)

[38] Campbell v. Acuff-Rose Music, 510 U.S. 569 (1994)

[39] Sony Corp. of America v. Universal City Studios, Inc., 464 U.S. 417 (1984) (Betamax)

[40] Harper & Row v. Nation Enterprises, 471 U.S. 539 (1985)

In our opinion, however, several cases can be quoted as valid precedent in favor of crawling activities:

Google v. Authors' Guild (Google Books). In the *Google Books* case[41] it was ruled that mass digitization of books and display of snippets on a commercial website is fair use. The judge ruled that the use made by Google was transformative, and that it did not harm the market value of digitized books (instead, in many cases it actually helps increase sales of these books). The decision (which came as a surprise to many) is now definitive[42].

iParadigms. In *iParadigms*[43] it was ruled (following a simple "the purpose justifies the means" principle) that unauthorized reproduction of student essays for purposes of plagiarism detection constitutes fair use.

Kelly. In *Kelly* (see above) it was ruled that the use of images in a commercial image search engine serves a transformative purpose; mostly because of that, the use was found to be fair.

On the other hand, some other cases can be used to argue against web crawling:

Meltwater (US). The *Associated Press v. Meltwater*[44] decision concerned the use of crawling and scraping of news articles (involving copying of up to 60% of the articles) by the owner of a news aggregator service. The court ruled that the use was not transformative, and that it had substantial effect on the market; therefore, it could not qualify as fair use. It shall be noted that in a parallel case (based on the same facts), the UK Supreme Court reached a different conclusion and ruled that the use was within the scope of the copyright exception for temporary acts of reproduction (see above). Both decisions are definitive.

3.1.7 Copyright and fundamental rights

It is sometimes argued that application of strict copyright rules can be a prejudice for fundamental rights (such as freedom of speech, information or research[45]). It should not be forgotten, however, that copyright in itself -- as a form of property -- is a fundamental right and therefore should not automatically submit to arguments based e.g. on freedom of research. Rather, the fundamental rights of copyright holders should be balanced against those of users of copyright-protected works. This reasoning may sometimes lead to conclusions that are interesting from the point of view of web crawling activities.

For example, in Germany the Federal Constitutional Court ruled that copyright exceptions shall be interpreted in such a way as not to limit fundamental rights recognized by the German constitution[46]. The case concerned the freedom of artistic expression (a work made entirely of quotations of other works), but the reasoning can probably be extended to freedom of research and teaching (as it is also expressly recognized by the same article of German constitution).

In France, after a 2015 ruling by the Court of Cassation[47], it seems that lower courts are obliged to perform a "balance of interests" test in order to determine in every specific case whether the application of copyright rules is a proportional limitation on freedom of expression.

[41] Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015)

[42] The Supreme Court refused to grant certiorari to the case: https://www.authorsguild.org/wp-content/uploads/2015/12/Authors-Guild-v.-Google-Petition-w-Appendix.pdf.

[43] A.V. v. iParadigms 562 F.3d 630 (2009)

[44] Associated Press v. Meltwater U.S. Holdings, Inc. (S.D.N.Y. March 21, 2013)

[45] Cf. art. 13 of the Charter of Fundamental Rights of the European Union.

[46] BVerfG, Beschluss vom 29. Juni 2000, 1 BvR 825/98 (Germania 3)

[47] Cour de cassation, civile, Chambre civile 1, 15 mai 2015, 13-27.391

The cases decided so far concerned artistic works, and the relevance of this new trend for web crawling (as well as its longevity) is highly uncertain.

> ***In our opinion the argument based on fundamental rights (e.g. freedom of research) is currently not a very reliable defense and should rather be used as a last resort and not as a foundation for a sustainable business model.***

3.1.8 Implied license

A crucial decision from the point of view of search engines in the US is the *Field v. Google*[48] case. The court ruled District Court of Nevada ruled that by uploading copyright-protected works on the Internet without taking steps to prevent indexing by search engines[49], the rightholder grants search engines (and Google in particular) an implied non-exclusive license to create cache copies of these works. In other words, such uploading can reasonably be interpreted as a permission to index. This is an important precedent in cases concerning web crawling in the US.

It shall be reminded here that a license is a contract by which the licensor authorize a licensee to perform certain acts that are restricted by an exclusive right. A license shall be distinguished from a copyright transfer (or assignment) agreement, as no rights are actually transferred to the licensee. Under US law, transfer of copyright ownership as well as an exclusive license (i.e. a license that guarantees the licensee exclusivity to perform certain acts) need to be in writing[50]. In contrast, a non-exclusive license does not have to be in writing, which means that it can be implied (e.g. deducible from the parties' behavior in certain circumstances). According to the Field v. Google case, such an implied license to index (as indexing necessarily involves reproductions, it normally requires a license) is granted to Google simply by uploading copyright-protected content on the Internet without taking steps to prevent indexing.

The decision also had some impact in Europe; in Germany, the Federal Court of Justice (BGH) ruled (in cases concerning Google Images)[51] that the act of uploading pictures on the open Internet (without any measures to prevent indexing) can be interpreted as "implied consent" (*konkludente Einwilligung*) for them to be indexed by search engines. Intriguingly, this holds true even if the pictures were uploaded without the rightholder's permission[52]. This implied consent, however, can be revoked by an explicit action of a rightholder at any moment, in which case the image has to be removed from the search engine. In a recent case concerning Google Images[53], BGH seems to have abandoned the doctrine of "implied consent" and instead based its analysis on CJEU's case law concerning linking (see above).

The key element of the implied license doctrine seems to be the notoriety of search engines such as Google, and the ease to prevent indexing (or at least indicate the intention to prevent it). In other words, every Internet user is supposed to know that content uploaded on the Internet can be indexed by search engines, and if he does not take measures to prevent it, it means that he accepts it. It is uncertain whether such reasoning can also apply to other forms of crawling activities (cf. above about the Meltwater (US) case). Moreover, the doctrine of

[48] Field v. Google, Inc., 412 F.Supp. 2d 1106 (D. Nev. 2006)
[49] Such as a robot.txt file or 'noindex' values.
[50] Section 204 of the US Copyright Act.
[51] BGH, 29.4.2010, Az. 1 ZR 69/08 (Vorschaubilder)
[52] BGH, 19.10.2011, I ZR 140/10 (Vorschaubilder II)
[53] BGH, 21.09.2017, I ZR 11/16 (Vorschaubilder III)

implied license has only been formally recognized in the US; in many EU countries a copyright license needs to be explicit (i.e. written) in order to be valid.

> ***In conclusion, it seems that the implied license doctrine cannot be a basis for crawling activities in Europe.***

3.1.9 Conclusion

In sum, from the copyright standpoint crawling is lawful if:

- the crawled contents are not protected by copyright (e.g. official works in some jurisdictions[54]) or copyright in them expired (in most countries, the term of copyright protection is 70 years after the death of the author);
- the user is granted a permission (i.e. a license) that covers crawling activities (e.g. content is available under a public license such as CC BY 4.0; licenses will be discussed below).

Currently, copyright exceptions in the EU allow for web crawling only in very limited circumstances. This is the case when:

- the reproductions made in the process are temporary (which is of very limited relevance for crawling activities); OR
- crawling is carried out for non-commercial research purposes, and it meets all the criteria set forth in the national transposition of the research exception (national transpositions in various EU Member States may e.g. only allow reproduction of excerpts of works, require equitable remuneration or only allow sharing within a strictly limited circle of persons); OR
- crawling is carried out for strictly private purposes and not in the professional context (in which case it may enter within the scope of the private copy exception).

Under US law (which applies to activities taking place on the US territory[55]), crawling seems to be more largely allowed under the doctrines of fair use and implied license. Whether a particular set of crawling operations can qualify as fair use would have to be evaluated on a case-by-case basis, taking into account the specific facts of each case.

[54] It shall be noted that in continental Europe copyright rules shall generally be interpreted according to the *in dubio pro auctoris* principle; i.e. if it is not certain whether a work is in the public domain (e.g. whether it falls within the scope of the exclusion for official works), it shall be deemed as being copyright-protected.

[55] See

Overview of issues related to conflict of laws (which law to apply in cross-border situations?) for more information.

3.2 Related rights; the *sui generis* database right

Definition and sources. Traditionally, related rights protect performers (actors, singers, musicians...) and music and video producers who, despite not being authors, play an important role in the creative industry. The exact scope of related rights varies greatly between jurisdictions and may also include: editors of scientific and critical editions, makers of unoriginal photographs or typographical arrangements, or press publishers. Due to limited space, this report will concentrate on the *sui generis* database right, created by the Database Directive 96/9/EC and implemented (in a fairly uniform way) in all the EU Member States. Such a right does not exist in the United States, where other means (such as action in misappropriation or Digital Rights Management) are used to protect investment in databases. It should be taken into account that other related rights can also be relevant to some specific web crawling operations, particularly those concerning audiovisual materials.

3.2.1 Scope of protection by the *sui generis* database right

Database. A database is defined as *"a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means"*[56]. In the light of the above definition, most websites shall be classified as databases (regardless of whether they use a Content Management System or not[57]). Moreover, websites may also contain databases (listings, catalogues, maps etc.).

Condition: substantial investment. A database is protected by the *sui generis* right if there has been a substantial investment in the obtaining[58], verification or presentation of the contents. It shall be noted that the investment in the creation of the contents is irrelevant from the point of view of the *sui generis* right[59]. Nevertheless, the practice shows that this threshold is easily met and it can be assumed that in many cases the investment required to set up, maintain and occasionally update a website will be qualified as substantial.

Exclusions: public funds. From the formal point of view, the Database Directive does not contain any exclusion from the *sui generis* right related to the nature of a database. Therefore, in most EU countries investment of public funds into a database gives rise to an exclusive right. Nevertheless, some exceptions exist: most notably, in the Netherlands (where the investment of public funds cannot give rise to the *sui generis* right) or -- seemingly -- in France (where the *sui generis* right of public bodies is largely neutralized by rules on the re-use of Public Sector Information[60]).

Exclusions: territorial scope. In principle, the *sui generis* database right protects only databases whose makers or rightholders (see below on ownership) are nationals of an EU Member States or have their habitual residence in an EU Member State. This excludes in

[56] Art. 1(2) of the Database Directive.

[57] We would like to stress the fact that the use of Content Management Systems is irrelevant from the point of view of the definition of a database; both websites that are built using CMS and those that are not may be databases according to the Database Directive. Arguably, however, the use of a CMS application may be seen as reducing the investment in the presentation of the contents (see below), but it can hardly be argued that it suffices to establish that the investment is not substantial.

[58] even if the data come from the public domain (CJEU, case C-545/07 (Apis-Hristovich)).

[59] CJEU, cases C-203/02 (The British Horseracing Board Ltd and Others v William Hill Organisation Ltd), C-46/02 (Fixtures Marketing Ltd v Oy Veikkaus Ab), C-338/02 (Fixtures Marketing Ltd v Svenska Spel AB), C-444/02 (Fixtures Marketing Ltd v Organismos prognostikon agonon podosfairou (OPAP)).

[60] Cf. Art. L. 321-3 Code des Relations entre le Public et l'Administration.

particular the databases (e.g. websites) owned by companies based in the US, which can obviously still be protected by copyright (see above).

3.2.2 Term of protection

Fifteen years renewable. A database is protected by the *sui generis* right for fifteen calendar years[61] following its completion. However, this term is renewed after any substantial change to the contents of a database (including accumulation of successive additions, deletions or alterations), which is a result of a substantial investment. Therefore in practice a database can be protected for an unlimited period of time.

Significance for web crawling. It shall be assumed that most websites are still under the *sui generis* right. This does not concern websites that have not been updated or otherwise maintained for over fifteen years, which is relatively rare.

3.2.3 Ownership

Maker. The *sui generis* right is held by the maker of the database, i.e. the person or company that takes the initiative and the risk of investing in the creation of the database[62]. This excludes subcontractors (e.g. freelance web designers) and employees from the benefit of the *sui generis* right. Therefore, from the point of view of the *sui generis* right it is irrelevant who actually performed the acts leading to the creation of the database.

3.2.4 Exclusive rights

Introduction. The *sui generis* right consists of two exclusive rights: the right of extraction and the right of re-utilization. However, these acts require permission from the rightholder only if they concern a substantial part of a database.

3.2.4.1 Extraction

Extraction is defined as *"permanent or temporary transfer of all or a substantial part of the contents of a database to another medium by any means or in any form"*[63]. Therefore, it is essentially identical to the right of reproduction (see above). Since web crawling consists of transferring the contents of a website to another medium, it shall be qualified as an act of extraction.

3.2.4.2 Re-utilization

Re-utilization is defined as *"any form of making available to the public all or a substantial part of the contents of a database"*. It is therefore similar to the right of communication to the public (see above). In our view, "a public" in this context is the same notion as in the context of copyright (see above); purely internal use may therefore not be qualified as re-utilization. Nevertheless, some scenarios (4. Sharing and 5. Distribution) defined in above of this report involve re-utilization of data.

Substantial part. A part of a database can be quantitatively or qualitatively substantial[64]:

- in order to assess quantitative substantiality of a part of a database, the volume of the part shall be compared to the volume of the whole database. It seems that a part is

[61] I.e. the protection expires fifteen years from the first of January of the year following the date of completion

[62] Recital 41 of the Database Directive

[63] Art. 7.2 (a) of the Database Directive

[64] CJEU, case C-203/02 (British Horseracing Board v William Hill)

quantitatively substantial if its volume exceeds 10% of the whole database; however, the notion lacks clarity. For example, in 2011 an English court ruled that a part representing 11% of the volume of a database was *"at the lower end of what could be regarded as quantitatively substantial"*[65]; on the other hand, a part representing 12% of a database was ruled non-substantial by a French court[66];

- a part is qualitatively substantial if there has been a substantial investment in obtaining, verifying, or presenting the part of the database. As a result, a part that is quantitatively non-substantial (e.g. 1% of a whole database) can be qualitatively substantial if it meets the criterion of substantial investment. Therefore, even one paragraph of text can theoretically be a substantial part of the whole website, if its obtaining, verification or presentation was particularly costly (e.g. high quality human translation into or from a rare language, OCR of a rare manuscript...).

Neither qualitatively nor quantitatively substantial parts of a database can be extracted or re-utilized without the permission from the holder of the **sui generis** ***database right.***

Repeated extraction and/or re-utilization of non-substantial parts. Moreover, repeated and systematic extraction and/or re-utilization of insubstantial parts of the contents of the database implying acts which conflict with a normal exploitation of that database or which unreasonably prejudice the legitimate interests of the maker of the database is expressly not allowed (art. 7(5) of the Database Directive). Therefore, a crawling operation in which the crawler repeatedly and systematically visits a website only to copy a non-substantial part of its contents (e.g. 5%) at each visit is unlawful and cannot be seen as a mean to circumvent the obligation to obtain the rightholder's permission.

3.2.5 Exceptions to the *sui generis* right

Introduction. Just like in case of copyright, there are statutory exceptions to the *sui generis* database right. They are provided for in art. 9 of the Database Directive. Only one[67] of them -- the research exception -- seems to be relevant from the point of view of web crawling activities. Just like most copyright exceptions, the exceptions to the *sui generis* database right are optional, which means that Member States are free not to transpose them, or only transpose a limited version of them (see above about copyright exceptions).

Lawful user. Only a "lawful user" of a database can benefit from the exceptions to the *sui generis* right. In the context of web crawling it means that the crawled content has to be available on the open Internet, and not in a password- or paywall-protected environment, unless this access barriers have been lawfully surpassed (e.g. a password was lawfully obtained, or the necessary payment made).

Research exception. Art. 9(b) of the Database Directive allows the Member States to adopt statutory exceptions *"in the case of extraction for the purposes of (...) scientific research as long as the source is indicated and to the extent justified by the non-commercial purpose to be achieved"*. It shall be noted that the exception does not cover any form of re-utilization of the extracted data (communication to the public); nevertheless, it should be kept in mind that

[65] Beechwood House Publishing v Guardian Products Ltd [2011] EWPCC 22

[66] TGI Paris, 5 sept. 2001, Cadremploi / Keljob

[67] The equivalent of the 'private copy' exception in the Database Directive is only limited to non-electronic databases, and therefore of no use for web crawling activities; it shall also be noted that (unlike in the Copyright Directive) there is no exception for temporary acts of extraction (although transmission of databases in a network is probably covered by the notion of 'normal exploitation' of a database -- see above).

non-substantial parts of a database can always be re-used (i.e. communicated to the public) by a lawful user (see above).

Just like the corresponding copyright exception, the research exception to the *sui generis* right is limited to non-commercial purposes (see above). Moreover, it also requires indication of the source, which in the context of databases is somewhat unclear (although it seems that indication of the URL is enough to satisfy the requirement).

Due to the optional character of the exception, its national transpositions may vary (although it seems that most -- if not all -- of the EU Member States have implemented it). For example, in Germany the exception is as large as allowed by the Directive (i.e. there are no additional requirements), while in France faulty transposition reduced the exception to dead letter[68].

New Text and Data Mining exceptions. As mentioned above in the part concerning copyright, in the past few years new exceptions concerning specifically Text and Data Mining (TDM) were adopted in some EU Member States, including the UK, France and Germany. In the UK, the exception does not concern the *sui generis* database right. In Germany and in France, the TDM exceptions cover also the *sui generis* database right.

In Germany, the newly introduced art. 60d of the German Copyright Act expressly states that the extraction of non-substantial parts of a database (even repeated and systematic) for Text and Data Mining for non-commercial research purposes constitutes normal use of a database and as such cannot be prohibited by the holder of the *sui generis* right. It shall be kept in mind that the exception requires for equitable remuneration to be paid to a collecting society (see above).

In France, the scope of the exception seems to be limited to data “included in or associated with scientific writings”, which makes it of very little use for web crawling activities.

In addition to that, the mandatory TDM exception included in the proposal of the new Directive on Copyright in the Digital Single Market covers the *sui generis* database right (both extraction and reuse). As mentioned above, for now the content of the exception is impossible to predict.

3.2.6 Conclusion

Many websites can be qualified as databases and protected by the *sui generis* database right, which is independent from copyright. Therefore, in principle extraction and reuse of substantial parts of such websites needs to be authorised by the rightholder (the maker of the database).

However, non-substantial parts of such databases (typically less than 10%, unless the part is qualitatively substantial) can be freely extracted (copied) and re-utilised (shared) for any purpose. On the other hand, repeated and systematic extraction of such non-substantial parts is prohibited.

The existing research exception may allow extraction (reproduction) of substantial parts of protected websites for non-commercial research purposes (providing that the source is indicated). However, it needs to be checked whether (and how) the exception has been transposed in the applicable national law. Reuse (sharing) of substantial parts of databases for non-commercial research purposes is not allowed.

[68] See art. L. 342-3, 4 of the French Intellectual Property Code (note that the exception also allows for reuse, which manifestly goes against the Directive and will likely not be upheld in court).

It is useful to keep in mind that websites produced by US-based companies are not covered by the *sui generis* database right.

3.3 Digital Rights Management

Definition. In the legal context, the term "Digital Rights Management" (DRM, or technological protection measures) covers a range of technological measures designed to prevent or restrict acts not authorised by the copyright holder, rights related to copyright or the *sui generis* right in databases. Examples of DRMs include password protection, paywalls or captcha challenges. Efficient DRMs are legally protected against circumvention.

Source of legal protection. First mentioned in art. 11 of the World Intellectual Property Copyright Treaty (1999), DRMs are now protected in the EU by art. 6 of the InfoSoc Directive, and in the US by the Digital Millenium Copyright Act (section 1201, 17 USC).

> ***The result of legal protection of DRMs is that rightholders can control access to their content. Mere circumvention of DRMs is subject to sanctions, even if it is not followed by acts of copying and/or communicating to the public. It is therefore important to keep in mind that for crawling activities to be lawful, the bots should not attempt to circumvent DRMs.***

3.4 Personal data

Sources. The rules concerning processing of personal data in EU Member States were first harmonized by the Personal Data Directive of 1995, and soon will be unified by the General Data Protection Regulation (GDPR) which will enter into force on 25 May 2018. Unlike a directive, a regulation applies directly in the legal systems of all the Member States. This means that the same body of rules will apply across all the EU Member States. In the US, there is no general framework protecting personal information (instead, specific rules regulate e.g. health or credit information).

3.4.1 The concept of personal data

Definition. Personal data is defined very broadly as *"any information relating to an identified or identifiable natural person"*[69]. This definition can be analyzed into four elements[70]:

- **any information** regardless of its form (text, image, audiovisual recording) and content (facts, opinions, true or false, related to the public or private sphere);
- **...relating to [a person]…;** an information relates to a person by its content (it says something about a particular person), by its purpose (i.e. when it is likely to be used to evaluate the situation of a person, e.g. call log of a telephone inside a company office) or by its result (i.e. it can have an impact on the person's rights and interests, e.g. geolocation of a car);
- **...identified or identifiable [person]...;** a person is identified if he or she is singled out from a group; a person is identifiable if he or she can be singled out directly (e.g. by name and surname) or indirectly (e.g. by phone number, IP address, fingerprint…), by *any means reasonably likely to be used* (i.e. also by cross-referencing information from various available sources).

[69] Art. 4 point 1 of the GDPR.

[70] See: Article 29 Data Protection Working Party, Opinion 4/2007 on the concept of personal data.

- **...natural person;** this excludes the deceased[71] and legal entities; however, information about these categories of entities can also relate to natural persons and therefore be personal data (e.g. information about the employer's financial condition, or about one's ancestors' cause of death).

The concept of personal data is defined in a very broad way; it covers not only information about persons that are directly identified (i.e. names etc.), but also about persons who can be indirectly identified by anyone, and by any means reasonably likely to be used. Moreover, it covers both the private and the public (e.g. professional) sphere of the person's activity.

For example, it has been demonstrated that 87% of the US population can be identified by three pieces of information: gender, ZIP code and date of birth[72]. An apparently random set of information (an IP lawyer from Paris who collects stamps and drives a Mitsubishi) can uniquely identify a person. Many websites are likely to contain such information in textual form (social media, blogs, Internet fora, discussion groups, news websites enabling users to write comments under articles…). Images and audiovisual data are even more likely to be personal data (as a person's face or voice are highly identifying). Therefore, the rules regarding the processing of personal data are relevant from the point of view of web crawling activities.

3.4.2 Rules governing processing of personal data

Definition of processing. Processing is defined as *"any operation or set of operations which is performed on personal data (...) such as collection (...) structuring, storage, (...) retrieval (...) dissemination or otherwise making available, alignment, (...) erasure or destruction"*[73]. Therefore, all the web crawling scenarios defined in above of this report, if they involve personal data, shall be qualified as processing.

Basic terminology. The natural person that personal data relate to is called a **data subject**. The natural or legal person which (alone or jointly with others) determines the purposes and means of the processing of personal data is called a **data controller**[74]. The natural or legal person which processes personal data on behalf of the controller is called a **processor**.

It is important to understand that if personal data are processed in a project (e.g. collected via web crawling), the entity (or entities) that define how and why the data are processed are data controllers. There can be several controllers for every processing operation. Therefore, if a participant in a project independently decides to crawl the Internet in order to compile a dataset, he or she is the controller of any personal data processed (collected, copied, stored, deleted…) in the process.

The person who acts on behalf of the controller (e.g. is asked by his or her employer to crawl the web) is a data processor; as such, he or she also needs to respect certain obligations distinct from those that fall upon the controller.

Main principles. The main principles relating to processing of personal data are listed in art. 5 of the GDPR. They include:

[71] Some EU Member States (such as France) have special rules regulating post-mortem processing of personal data.

[72] L. Sweeney, Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000.

[73] Art. 4 point 2 of the GDPR

[74] There can be several data controllers for one processing ("joint controllers").

3.4.2.1 Lawfulness, fairness and transparency

Introduction. *"Fairness"* is not defined in the GDPR and seems to refer to the common understanding of the word. *"Transparency"* refers mostly to the quality of information to be provided to the data subject[75]. *"Lawfulness"* is further explained in art. 6 of the GDPR, according to which processing is lawful only if (at least) one of the conditions (called *"grounds for lawfulness"*) listed in its first paragraph is met. The most important of these grounds is consent.

Consent. Consent does not necessarily need to be written or even explicit (although the controller has to be able to prove its existence), but it needs to be freely given, specific, informed and non-ambiguous[76]. This means that a certain number of information about the processing has to be provided to the data subject so that he can validly consent; moreover, consent cannot be blank, but it has to be limited to a specific (narrowly defined) purpose. Moreover, when it comes to processing of sensitive data (i.e. data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, data concerning health, sex life or sexual orientation), consent needs to be explicit.

Uploading as implied consent? One can argue that if a data subject publishes personal information about himself on a publicly available website, he gives his implied consent for this data to be processed by anyone with access to the Internet[77]. It should be kept in mind, however, that consent needs to be specific; i.e. placing information on a website shall not be interpreted as blank consent for anyone to re-use this information for any purpose, but rather as a consent for the information to be used for purposes related to the website. Reasonable expectations of an average user should be taken into account: e.g. when one posts something on Twitter, does he reasonably expect that the information will be used to develop language models? In our opinion, the answer to this question is in the negative.

Withdrawal of consent. It shall be kept in mind that consent may be withdrawn at any time; such a withdrawal, however, is not retroactive (i.e. it only concerns future processing, and does not make processing prior to the withdrawal unlawful).

Alternatives to consent. Data subject's consent is not always required for processing to be lawful, but alternative conditions are hard to meet and should only be relied upon in special situations. The most important of these alternative grounds is detailed in art. 6 (1) (f) of the GDPR, according to which processing is lawful when it is *"necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject (...)"*. Such a *"balance of interests"* test[78] would have to be carried out on a case-by-case basis, depending on the specific purpose of processing, the data that are being processed and the context in which they were obtained.

> ***Arguably, some crawling activities, especially those carried out for non-commercial research purposes, can pass the "balance of interests" test. In our opinion, however, art. 6(1)(f) cannot be a reliable basis for a sustainable***

[75] *"concise, transparent, intelligible and easily accessible (...) using clear and plain language"* (art. 12 GDPR).

[76] Art. 4 point 11 of the GDPR

[77] This interpretation is further supported by art. 9(2)(e) of the GDPR, according to which the fact that the data were "manifestly made public by the data subject" is a possible ground for lawful processing of sensitive data.

[78] For further information see: Article 29 Data Protection Working Party, Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC.

> ***business model. The key obstacle here is that the processing of non-anonymized personal data should be necessary to achieve the purpose, i.e. the same purpose cannot be achieved while processing anonymized data, which is rarely the case in text processing technologies.***

3.4.2.2 Purpose limitation

Principle. According to art. 5(1)(b) of the GDPR personal data shall *be "collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes"*. This means that in principle data collected for one purpose cannot be re-used for another purpose, unless this other purpose is compatible with the original one. In assessing compatibility of purposes account should be taken of the data subject's reasonable expectations (would the data subject be surprised to discover the use that is being made of his data?)[79].

3.4.2.3 Data minimisation

Principle. According to art. 5(1)(c) of the GDPR personal data shall be *"adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed"*. This means that it is forbidden to process more data than what is necessary to achieve the purposes of processing; the principle is therefore incompatible with big data technologies and is probably the biggest obstacle for crawling activities.

3.4.2.4 Accuracy

Principle. According to art. 5(1)(d) of the GDPR personal data shall be *"accurate and, where necessary, kept up to date"*. Personal data that are inaccurate shall be erased or rectified without delay.

3.4.2.5 Storage limitation

Principle. According to art. 5(1)(e) of the GDPR personal data shall be *"kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed"*. Therefore in principle long-term storage of non-anonymized personal data is impossible from the legal point of view (but see below about exceptions concerning archiving in public interest).

3.4.2.6 Integrity and confidentiality

Principle. According to art. 5(1)(f) of the GDPR personal data shall *be "processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures"*. Arguably, if personal data are publicly available on the Internet, the threshold of "appropriate security" required in their processing is rather low.

3.4.2.7 Accountability

Principle. According to art. 5(2) the controller shall be responsible for, and be able to demonstrate compliance with the abovementioned principles. This means that in case of

[79] For further information see: Article 29 Data Protection Working Party, Opinion 03/2013 on purpose limitation.

conflict between a data subject and a data controller, the burden of proof is on the latter; in other words, it’s the controller that has to prove that he respected the law.

3.4.3 Rights of data subjects

Introduction. Regardless of the right to give consent, to refuse consent or to withdraw consent at any moment, data subjects have other rights with regards to their data (as long as they are not anonymized). Some of these rights include:

Right to information. Art. 14 of the GDPR provides a detailed list of information that should be provided to the data subject when the data is not collected directly from him (e.g. obtained via web crawling); these include (but are not limited to): the identity of the controller, the purposes of the processing, the categories of personal data concerned, the recipients of envisaged transfers, the source from which the personal data were obtained, the rights of the data subject and the right to lodge a complaint at the data protection authority. This information shall be “*concise, transparent, intelligible and easily accessible (...) using clear and plain language”.* The use of a URL seems allowed, if it meets all the above criteria and does not make the process unnecessarily complex for the user.

Right of access. According to art. 15 of the GDPR the data subject shall have the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the personal data and the information related to the processing.

Right to rectification. According to art. 16 of the GDPR *“[t]he data subject shall have the right to obtain from the controller without undue delay the rectification of inaccurate personal data concerning him or her”*.

Right of erasure (right to be forgotten). In certain circumstances (defined in art. 17 of the GDPR), the data subject has the right to obtain from the controller the erasure of personal data concerning him or her without undue delay. This is the case e.g. when the data subject withdraws his consent, when the data were unlawfully processed or when they are no longer necessary to fulfill the purpose for which they were collected.

3.4.4 Obligations of data controllers and processors

Obligations of data controllers. In correlation with the rights of data subjects, certain obligations fall on data controllers and processors. These include (but are not limited to):

Data protection by design and by default. According to art. 25 of the GDPR, data controller shall, already at the time of determination of means of processing, implement technological and organizational measures (such as pseudonymization, approval by an ethics committee, access restrictions…) to implement data protection principles (such as minimization or storage limitation) and to protect the rights of data subjects. Therefore, when web crawling may involve collection of personal data, the operation should be designed in such a way as to take into account the principles of the GDPR.

Obligation to keep records of processing activities. Art. 30 of the GDPR obliges data controllers and processors to keep detailed records of their processing activities. The obligation does not apply to organizations employing fewer than 250 persons unless the processing:

- is likely to result in a risk to the rights and freedoms of data subjects, OR
- is not occasional, OR
- includes special categories of data (as defined in art. 9 and 10 of the GDPR).

Notification and communication of data breaches. If a data breach occurs, the processor shall notify the data controller without undue delay. The controller shall then (again without undue delay) notify the national data protection authority and, if the breach is likely to result in a high risk to the rights and freedoms of natural persons, communicate it to the data subjects concerned.

Obligation to carry out a data protection impact assessment. Art. 35 of the GDPR stipulates that where a type of processing (in particular using new technologies) is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall, prior to the processing, carry out an assessment of the impact of the envisaged processing operations on the protection of personal data.

> Whenever crawling activities may involve processing of personal data, a heavy burden lies on the data controllers and processors. In particular, the crawling process shall be designed in such a way as to take into account the principles of the GDPR already at the inception phase (and, according to the accountability principle, it is data controller's obligation to demonstrate that he met this obligation). Quite often, the load may be disproportional to the outcomes of the crawling.

3.4.5 Transfer of personal data

Principle: free transfer within the EU. One of the main goals of harmonization of EU law was to allow free transfer of personal data within the European Union. As a result, such data can be freely transferred within the EU (of course providing that all the requirements of the GDPR, including principles of lawfulness and purpose limitation, are met).

Transfer to non-EU countries. When it comes to transfer to third countries, it is possible if one of the three conditions is met:

- the European Commission decided that the third country ensures an adequate level of protection[80] OR
- the transfer is subject to appropriate safeguards (such as the Privacy Shield Framework[81] concerning transfers of personal data to the United States[82]) OR
- the data subject has expressly consented to the transfer, after having been informed about potential risks.

3.4.6 Anonymization

Introduction. Given the multitude of legal constraints that apply to the processing of personal data, anonymization of crawled data is a practical necessity. Indeed, the GDPR does not apply to data that are anonymous or have been anonymized.

Definition of anonymized data. Anonymized data can be defined as data that can no longer (taking into account all the means reasonably likely to be used) be related to an identified or identifiable natural person. Therefore, anonymization is a process of breaking the link between the information and a natural person. In doing that, account should be taken of the fact that data that refer to persons that can only be identified indirectly (e.g. by reference to additional

[80] List of decisions concerning adequate level of protection can be found at: http://ec.europa.eu/justice/data-protection/international-transfers/adequacy/index_en.htm

[81] https://www.privacyshield.gov/welcome

[82] *Editor's note, on July 16, 2020, The European Court of Justice ruled that the Privacy Shield Network did not provide sufficient guarantees to data subjects and was therefore invalid (cf. C-311/18 Data Protection Commissioner v Schrems)*

information available on the Internet) are still personal data. As a consequence, simply removing named entities from a text is not sufficient to anonymize it and additional steps need to be taken for data to be properly anonymized.

High standard. Since anonymization is a technical issue, it will not be further discussed in this report. However, it should be noted that the standard set for anonymization is high; more about various anonymization techniques can be read in an Opinion of the Article 29 Data Protection Working Party[83], which discusses such techniques as k-anonymity, l-diversity and t-closeness.

3.4.7 Special rules concerning research and archiving in the public interest

Introduction. In order not to paralyze research activities (in particular in the domains of statistics and history) and archiving in the public interest, the GDPR introduces a number of flexibilities and “safety valves”. In the context of data protection “research” is to be interpreted broadly, including e.g. technological development and demonstration, fundamental research, applied research and privately funded research[84].

Derogations from the principle of specificity of consent. While in principle consent has to be specific (i.e. limited to a specific purpose), some flexibility is allowed in the context of research. According to recital 33 of the GDPR *“data subjects should be allowed to give their consent to certain areas of scientific research when in keeping with recognised ethical standards for scientific research”*. This means that consent, rather than concerning a specific project, may concern a whole area of scientific research (e.g. linguistics and language technology). In such case, the processing has to respect *“recognised ethical standards for scientific research”*, which may be interpreted as a prohibition of commercial distribution of data processed on the basis of such extended consent.

Derogations from the principle of purpose limitation. According to the principle of purpose limitation (see above), personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes. However, this expressly allows further processing for purposes that are compatible with the initial purpose. By express derogation of art. 5(1)(b) of the GDPR, further processing for research purposes (and for purposes of archiving in the public interest) is always to be regarded as compatible with the initial purpose. This means that data lawfully collected (i.e. in principle with the data subject’s consent) for a different purpose (e.g. marketing, registration etc.) can always be lawfully re-used for scientific research. This significantly broadens the possibilities to re-use crawled data, providing that crawling respected all the principles of the GDPR (which, as stated above, is very difficult in practice).

Derogations from the principle of storage limitation. According to the principle of storage limitation (see above), personal data shall be stored for no longer than necessary for the purposes for which they are processed. However, the GDPR provides for an express derogation from this principle for scientific research and archiving in the public interest. This means that in this context personal data can be stored for longer periods (e.g. after the completion of a research project), even without being anonymized.

Appropriate safeguards. In order to take advantage of the abovementioned derogations from the principles of purpose and storage limitation, processing shall be subject to *“appropriate safeguards for the rights and freedoms of data subjects”*[85]. According to art. 89(1) of the

[83] Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques.
[84] Recital 159 of the GDPR.
[85] Art. 89(1) of the GDPR.

GDPR *"[t]hose safeguards shall ensure that technical and organisational measures are in place in particular in order to ensure respect for the principle of data minimisation"*. They may include pseudonymization[86] (as expressly recognized by the GDPR) and possibly also approval by an ethics committee (at universities), encryption or other advanced technological and organizational measures.

Possible derogations from certain rights of data subjects. Art. 89(2) of the GDPR allows Member States to provide for derogations from certain rights of data subjects (those defined in art. 15 (right of access), 16 (right to rectification), 18 (right to restriction of processing) and 21 (right to object)). It seems that as of now, few countries have adopted such derogations[87].

Necessary application of the principle of data minimization. Formally, the GDPR does not allow any derogations from the principle of data minimization for processing for the purposes of research or archiving. This means that, also in the context of research, only the personal data that are *"adequate, relevant and limited to what is necessary in relation to the purposes [of processing]"* can be collected. This is particularly problematic from the point of view of web crawling activities, where it is difficult to respect this principle.

3.4.8 Conclusion

"Personal data" is a broad concept that covers not only information that is directly identifying, but also information that can be used to indirectly identify a natural person. Therefore, when particular kinds of websites (such as discussion fora, social media or even online shops where users can post reviews) are crawled, personal data are likely to be collected in the process. Likewise, when audiovisual material is collected, it should often be treated as personal data (as voice and a person's appearance are highly identifying).

Processing of personal data is regulated by the GDPR, which imposes strict rules with regards to data minimization (only necessary, adequate and relevant data can be processed), storage limitation (data cannot be stored for longer than necessary) and lawfulness of processing (which in principle requires consent of the data subject).

In order to comply with these principles, crawling operations would have to be designed in such a way as to collect only data that are necessary from the point of view of their purposes (which seems completely incompatible with data-intensive technologies).

Moreover, even after consenting to the processing, data subjects (i.e. the natural persons that data relate to) have non-waivable rights in relation to their data (such as information, access and rectification), and data controllers (i.e. persons or entities that define the purposes of processing) and processors (i.e. persons on entities who process data on behalf of controllers) need to comply with complex obligations regarding organizational and technical aspects of processing (to implement data protection by design and by default, to carry out an impact assessment, to keep a register of processing operations, to notify breaches...). All these requirements are indeed difficult and costly to comply with.

[86] 'Pseudonymisation' means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person (art. 4 point 5 of the GDPR)

[87] Germany being one of them (at the level of federal law, which does not concern universities): cf. art. 27 BDSG(neu).

It seems necessary, therefore, to resort to anonymization techniques. Ideally, data should be automatically anonymized already at the stage of their collection. The crawler should either omit personal data, or automatically anonymize them.

Under the GDPR, some leeway is allowed for research activities, providing that *"appropriate safeguards"* are implemented. However, even in this case the principle of data minimization needs to be observed, and the data need to be anonymized as soon as possible taking into account the purposes of processing. This provides little relief for web crawling activities, even carried out for research purposes.

3.5 Contracts (Terms of Use, licenses, notices, waivers)

Introduction. Internet content often comes with contracts (Terms of Use, licenses) and notices (such as "all rights reserved") that aim to regulate the conditions of its re-use. It is therefore important to examine the validity of such conditions and their influence on web crawling activities. The following analysis will be divided into three parts: first, the general enforceability of such terms will be discussed; then, various instruments aimed at preventing and allowing web crawling will be presented.

3.5.1 Enforceability of standard form contracts

Introductory remark on freedom of contract. It is important to keep in mind that freedom of contract is a fundamental principle of modern economies. As a result, in principle contracts are always enforceable and can even prevail over (some) statutory rules (in short, in the domain of contract law what is not expressly prohibited is allowed). For example, contracts override statutory copyright exceptions -- this means that even if a certain use (e.g. for research and teaching purposes) is allowed by a statutory exception, it can still be efficiently prohibited by a contract (e.g. if Terms of Use forbid any reproduction and communication to the public of the website's contents).

The concept of a standard form contract. While the Internet is still relatively new, the question of enforceability of standard form contracts is a fairly old one. A standard form contract (also known as *adhesion contract*) is a contract in which all the terms are set by one party and the co-contractor has no possibility to negotiate (it's a take-it-or-leave-it situation). Such contracts, especially when proposed to non-professionals, are rarely read (as they are often -- on purpose -- unnecessarily long and written in small print and/or overly technical language). At the same time, in the current economy it would be unworkable to individually negotiate every contract. Therefore, there is a need to protect the weaker party (e.g. the one unable to negotiate) against unfair terms in such contracts. Terms of Use or public licenses are indeed a type of standard form contracts (which are typically implicitly accepted simply by accessing the content, cf. "shrink-wrap license"). In order to be enforceable, such contracts need to satisfy certain criteria.

Multitude of national solutions. The exact conditions for enforceability of standard form contracts vary from country to country[88]. Various elements may be taken into account in deciding whether a given clause of a standard form contract is enforceable in the particular circumstances: the characteristics of the contracting parties (e.g. whether the weaker party is acting within its professional capacity or not), the language of the contract (contracts in a foreign language may not be enforceable), the circumstances of acceptance (i.e. whether the weaker party was given a reasonable opportunity to read the contract) and of course the

[88] Cf. art. 1119 of the French Civil Code, ss. 305 and following of the German Civil Code.

content of the clauses (some clauses may be *ex officio* unenforceable, other may shift the burden of proof). As a general rule it seems that (at least among businesses[89]) the party who claims that a clause is unenforceable would have to prove that the conditions for enforceability were not met, or at least raise the argument in court. Since the stronger party is unlikely to claim that their standard contract terms are unenforceable, in practice only those terms of standard form contracts that are not in favor of the weaker party can see their enforceability questioned. Accordingly, those terms of standard form contracts that are in favor of the weaker party are much more likely to be enforced. Therefore, it makes sense to separately analyze clauses that allow crawling and those that prohibit it, the former being more likely to be enforced than the latter.

3.5.2 Clauses that allow crawling -- public licenses

> ***Web content may be available under public licenses (i.e. contracts by which the rightholder grants everyone permission to perform certain acts that are normally restricted under copyright) or notices that are intended to produce the same effect. By far the most common public licenses for web content are the Creative Commons (CC) licenses.***

3.5.2.1 Creative Commons licenses and tools

3.5.2.1.1 Presentation of the Creative Commons licenses

The Creative Commons organization. The Creative Commons is a non-profit organization founded in 2001 and based in California. It released a series of copyright licenses (“license suite”) that are characterized by their modularity and widely known iconography.

Creative Commons licenses. In principle, CC licenses grant a broad permission to accomplish various acts that could be restricted under copyright or related rights; they are built of four blocks which correspond to four requirements or restrictions on the freedom of use: Attribution (BY), Share-Alike (SA), Non-Commercial (NC) and No Derivatives (ND). The BY block is a mandatory part of every CC license. This amounts to a total of six CC licenses (since the SA and ND blocks are incompatible and cannot be combined in one license): CC BY, CC BY-SA, CC BY-NC (used by Wikipedia), CC BY-ND, CC BY-NC-ND and CC BY-NC-SA. The most recent version of CC licenses is labelled 4.0, although content licensed under earlier versions of CC licenses (which were often subject to “porting”, i.e. adaptation to one national legislation, e.g. CC BY 3.0 France) can still be found on the Internet. For version 4.0, porting (creation of national versions) is not authorized by the Creative Commons organization. In the following analysis, only 4.0 CC licenses will be used as examples.

Creative Commons licenses and statutory exceptions. It is important to keep in mind that unlike “ordinary” licenses, CC licenses expressly do *not* override statutory exceptions to exclusive rights. Therefore, if a use is allowed e.g. by an exception for text and data mining, it is never prohibited by a CC license, even the most restrictive one (CC BY-NC-ND).

Scope of CC licenses. It shall be kept in mind that CC licenses cover not only copyright, but also (expressly since version 4.0) the *sui generis* database right and other neighboring rights. Unfortunately, the *sui generis* database right is not expressly addressed in older versions of CC licenses and therefore additional permission may be necessary for crawling certain websites licensed under these licenses.

[89] Consumer contracts (between a professional and a consumer) are outside the scope of the report, as crawling activities are typically carried out by businesses within their professional capacity.

3.5.2.1.2 Presentation of the building blocks and their impact on crawling activities

Attribution (BY). The BY requirement obliges the user to retain certain information[90], including the author's name, in every copy of the licensed content, whenever this content is communicated to the public[91]. As long as this obligation is respected, there is nothing else in the BY requirement that would restrict any of the web crawling scenarios described in above of this report.

Share-alike (SA). The SA requirement is triggered when Adapted Material[92] based on the licensed content is communicated to the public: this Adapted Material must then be shared under a compatible license. This may create certain problems related to license interoperability; it also makes commercial distribution of adapted material (i.e. distribution under a proprietary license) impossible. However, the SA clause is not triggered when there is no communication to the public, so arguably scenarios 1-3 (archiving only, data analysis and exploitation) are not affected by this requirement, and scenario 4 (sharing) is affected only if there is actual communication to the public (and no purely internal use). In such a case, as well as in scenario 5 (distribution), the SA requirement needs to be respected. It is important to keep in mind that technical modifications necessary to use the material "in all media and formats" are always allowed, and that they do not trigger the SA requirement.

Non-commercial (NC). Web content licensed under a license concerning the NC requirement can only be crawled for purposes *"not primarily intended for or directed towards commercial advantage or monetary compensation"*. Arguably, this only allows crawling e.g. for academic research purposes, or other non-commercial purposes.

No Derivatives (ND). The ND requirement prohibits any sharing (i.e. communication to the public) of Adapted Material. Such Adapted Material can still be made (so: scenarios 2 and 3 are still allowed), but it cannot be distributed (scenario 5 may only concern the licensed content in unmodified form; in scenario 4, Adapted Material can only be used internally). It is important to keep in mind that technical modifications necessary to use the material *"in all media and formats"* are always allowed[93].

[90] According to section 3(a) of every CC license*: If You Share the Licensed Material (including in modified form), You must:*
retain the following if it is supplied by the Licensor with the Licensed Material:
identification of the creator(s) of the Licensed Material and any others designated to receive attribution, in any reasonable manner requested by the Licensor (including by pseudonym if designated);
a copyright notice;
a notice that refers to this Public License;
a notice that refers to the disclaimer of warranties;
a URI or hyperlink to the Licensed Material to the extent reasonably practicable;
indicate if You modified the Licensed Material and retain an indication of any previous modifications; and
indicate the Licensed Material is licensed under this Public License, and include the text of, or the URI or hyperlink to, this Public License.

[91] Arguably, scenarios 1-3 (archiving only, data analysis, exploitation) do not involve communication to the public; scenario 4 (sharing) does not involve communication to the public if the sharing is purely internal (within one entity); scenario 5 (distribution) necessarily involves communication to the public.

[92] Adapted Material is defined as "material subject to Copyright and Similar Rights that is derived from or based upon the Licensed Material and in which the Licensed Material is translated, altered, arranged, transformed, or otherwise modified in a manner requiring permission under the Copyright and Similar Rights held by the Licensor". Arguably, scenarios 2 (data analysis) and 3 (exploitation) involve making of Adapted Material, unless the use is limited to simply collecting statistics (e.g. word frequencies) about the licensed material.

[93] Section 2, a, 4 of all the CC 4.0 licenses.

3.5.2.1.3 Other CC tools

CC0. CC0 is a waiver that allows rightholders to waive their exclusive rights and place their works (and databases) in the Public Domain. In some national laws (such as German law) copyright is not waivable; in order to preserve the validity of the tool under such laws, CC0 contains a "fallback clause" which stipulates that if the waiver is not valid, it shall be interpreted as the broadest possible license. In both cases, it seems that content available under CC0 can be freely crawled and re-used for any purpose (even if the waiver is not valid and some difficulties arise as to the scope of the fallback license, the risk of the user being sued by the rightholder seems extremely low).

Public Domain Mark (PDM). The PDM is a tool that allows to mark a work that has been identified as part of the public domain. It does not create any contractual relation, it is simply intended to avoid effort multiplication. Public Domain content can be freely crawled and re-used by anyone and for any purpose.

3.5.2.2 Other licenses and notices

Open Government licenses and PSI-related notices. In certain countries, public sector bodies make Public Sector Information (documents held by public sector bodies and not subject to any access restrictions) available for re-use under specific government licenses. These include (but are not limited to) the UK's Open Government License (OGL), France's *Licence Ouverte* and Germany's *Datenlizenz Deutschland*. These licenses are country-specific and their analysis fall beyond the scope of this report. In general, it can be said that they allow re-usability to a large degree, which covers most (if not all) identified web crawling scenarios (usually subject to the requirement to mention the source). In some other countries (such as Spain) Public Sector Information is often made available with a notice which is intended to have the same effect as a government license.

Possibility to request a license for re-use of Public Sector Information. According to the Directive 2003/98/EC on the re-use of Public Sector Information, all documents held by public sector bodies (including their websites) that are not excluded from access shall also in principle be available for re-use. This means that even if Public Sector Information is not made available under a public license or with an accompanying notice, users can request a license for their re-use for a non-excessive fee (in principle limited to marginal costs). Such requests shall be answered in principle in twenty days. National rules regarding re-use of Public Sector Information may vary[94], so it is important to consult the applicable national law on the matter.

Open Data Commons. Open Data Commons licenses were released by the Open Knowledge Foundation; they were intended for licensing databases in the EU at the time when CC licenses did not cover the *sui generis* database right (i.e. for versions older than 4.0). These licenses are rarely used nowadays.

3.5.3 Clauses that prohibit crawling

Examples. Terms of Service may generally prohibit crawling, e.g.:

Apart from legitimate search engine operators and use of the search facility provided on the Website for users, no person may use or attempt to use any technology or applications

[94] Most importantly, despite the fact that educational and research establishments (except university libraries), as well as cultural establishments other than libraries, museums and archives are excluded from the scope of the PSI Directive, they may be concerned by national rules on the re-use of PSI (e.g. in France, national PSI rules apply to documents held by universities).

(including web crawlers or web spiders) to search or copy content from the Website for any purpose without Our prior written consent.[95]

They can also limit the possibility to carry out crawling operations:

You may not copy, reproduce, republish, disassemble, decompile, reverse engineer, download, post, broadcast, transmit, make available to the public, or otherwise use tate.org.uk content in any way except for your own personal, non-commercial use[96]

Enforceability. As mentioned above, the enforceability of such clauses may often be quite debatable (albeit less so if express approval, such as ticking a box or clicking a button, is necessary to access the website) and may depend on the circumstances of each case. Nevertheless, generally such clauses, if enforceable, may indeed prohibit crawling, even if it is allowed by an applicable statutory copyright exception (e.g. for non-commercial research purposes). What cannot be prohibited by a contractual clause is the right of a lawful user of a database to extract a non-substantial part of the database (see above)[97]. Such a clause would not be enforceable. Also the new text and data mining exception may be immune to contractual clauses (it is in Germany and in the UK, but not in France; the soon-to-be-adopted mandatory TDM exception at the EU level is also expected not to be overridable by contracts).

Contractual limitations of access to public domain content. Recently, the CJEU ruled that nothing prevents the adoption of contractual clauses regulating the conditions of use of content that is not protected by any exclusive right (i.e. content that is in the public domain)[98]. This is potentially a very significant reduction of the public domain.

3.5.4 Conclusion

> ***Most websites are available under conditions specified in a contractual instrument "attached" to the website (Terms of Use, public license). In principle, this instrument becomes binding once the website is accessed. The clauses of such contracts can roughly be divided into those that allow and those that prohibit crawling.***

Regarding the first group of clauses, websites can lawfully be crawled if they are available under a public license (such as a CC license), providing that the conditions of the license are respected. Some notices may have effect similar to public licenses.

The second category of contractual clauses can effectively prohibit any crawling (even allowed under a statutory copyright exception, unless this exception is expressly non-overridable by contractual clauses). However, the enforceability of such clauses may be doubtful (depending on the applicable law), especially if express acceptance of the instrument (ticking a box or clicking on a button) is not necessary to access the website.

[95] Art. 6.4 of the Terms of Conditions available at: http://www.dmasa.org/terms-conditions.asp

[96] Art. 5 of the Terms of Use available at: http://www.tate.org.uk/about-us/policies-and-procedures/website-terms-use

[97] Art. 8 of the Database Directive.

[98] CJEU, case C-30/14 (Ryanair).

3.6 Overview of issues related to conflict of laws (which law to apply in cross-border situations?)

Introduction. In cross-border situations (e.g. when an US-based company infringes on copyright of a French author by making his work available to the public on a server based in Thailand), it is often unclear which national law should apply. Questions related to conflict of laws (also called international private law) can indeed be very complex; the following analysis will therefore be limited to presenting some general rules applicable in the EU.

Intellectual Property Rights. According to article 8 of the Rome II Regulation[99] *"[t]he law applicable to (...) infringement of an intellectual property right shall be the law of the country for which protection is claimed"*[100]. This is generally to be interpreted as the law of the country in which the alleged infringement was committed. In other words, if crawling consists of reproducing web content to a hard disk situated in a certain country, the copyright law of this country should apply. This means that e.g. while in France, the entity that crawls the Internet needs to respect French law, and may benefit from French statutory exceptions (so, for example, even if the crawled content is on a server situated in the United States, the French entity cannot invoke the *fair use* defense).

In practice, however, the question of applicable copyright law is not necessarily that simple. It seems that sometimes the mere fact that the activity of the alleged infringer is "directed" towards a given country is sufficient for a judge to apply the copyright law of that country[101]. For example, the German Federal Court of Justice applied German law to acts committed by Google (on US territory) because their services were available in German language and intended for the German public, the claimant was a German resident, and the consequences of infringement were mostly related to German territory[102].

> ***It seems possible, therefore, that even if the entity respects the law of the country in which it is situated, it may be found liable for infringement of foreign copyright rules, if its activities are directed towards a foreign country (or countries).***

Contractual claims. When it comes to claims arising from a contract (e.g. liability for breach of Terms of Use), the applicable law can be chosen by the contract itself. If there is no such clause (called choice of law), or if the clause is unenforceable (which may be the case in standard form contracts), the law governing terms and conditions seems to be the law of the country where the owner of the website is normally situated[103] (e.g. German law governs Terms of Use of a website of a German company).

Personal data. The GDPR governs the processing of personal data by controllers and processors established in the EU (even if the processing actually takes place outside the EU, e.g. processing carried out in Morocco on behalf of a French company is governed by the GDPR). Moreover, GDPR also governs the processing by non-EU-based processors and controllers related to the offering of goods or services to EU citizens (e.g. an online store in

[99] Regulation 864/2007 of 11 July 2007 on the law applicable to non-contractual obligations (Rome II)

[100] Cf. art. 5(2) of the Berne Convention.

[101] CJEU, cases C-585/08 (Pammer v. Karl Schlütter GmbH & Co. KG) and C-144/09 (Hotel Alpenhof v. Mr. Heller)

[102] BGH, 29.4.2010, I ZR 69/08 (Vorschaubilder)

[103] Art. 4(2) of the Rome I Regulation contains a fallback rule according to which *"the contract shall be governed by the law of the country where the party required to effect the characteristic performance of the contract has his habitual residence"*.

which EU citizens can shop), or to monitoring the behavior of EU citizens on EU territory[104]. In short, the principles of the GDPR need to be observed by all those established in the EU, even if they process data of non-EU citizens.

[104] Art. 3 of the GDPR.

4 Sanctions

4.1 Copyright infringement

The sanctions for copyright infringement vary from country to country; they may include:

- injunction (court order to stop infringing acts);
- impounding of infringing copies;
- compensatory damages (compensating the rightholder for the suffered loss) and punitive damages;
- fines (in France up to 300 000 EUR, and up to 750 000 EUR if infringement is committed by an organized group);
- imprisonment (in France up to 3 years, and up to 7 years of infringement is committed by an organized group);
- costs of proceedings and the claimant's attorney's fees.

4.2 Infringement of the *sui generis* database right

The sanctions for infringement of the *sui generis* database right are essentially similar to those for copyright infringement.

4.3 Circumvention of Technological Protection Measures (DRMs)

Circumvention of a technological protection measure is also subject to a fine (in France up to 3 750 EUR), and so is the act providing others with means to do so (e.g. with an algorithm that can circumvent a specific category of DRMs -- in France sanctions for such act go up to 30 000 EUR and six months' imprisonment).

4.4 Breach of contract

Financial sanctions for breach of contract can be defined by a contractual clause; however, such clauses would not always be enforceable in standard form contracts. More likely, sanctions for breach of contract would be limited to compensatory damages. Moreover, violating Terms of Use of a website would normally lead to blocking access to the website, shutting down the user account etc.

4.5 Unlawful processing of personal data

> ***Under the GDPR, administrative sanctions for unlawful processing of personal data can go up to 20 000 000 EUR, or in the case of an undertaking, up to 4 % of the total worldwide annual turnover of the preceding financial year, whichever is higher. Member States may also lay down rules on other penalties (such as imprisonment) for unlawful processing.***

5 Conclusion

5.1 Main findings

As the above analysis has demonstrated, web crawling is subject to various legal constraints, including copyright, the *sui generis* database right and data protection.

When it comes to copyright, the unauthorized reproduction and communication to the public of copyright-protected contents can expose the user to substantial sanctions. Since crawling by definition consists of making reproductions of websites, it shall either be authorized by the rightholder, or carried out on a basis of a statutory exception. Unfortunately, in the current state of things in the European Union, statutory exceptions (such as private copy or temporary reproduction) have only very limited relevance for web crawling activities. This, however, may change soon, if a new statutory exception for Text and Data Mining (TDM) is adopted at the EU-level (as suggested by the European Commission in the proposal for the new Directive on copyright in Digital Single Market). Some EU-countries, such as Germany, have already provided for TDM exceptions in their national laws, but for now they can only be limited to non-commercial research activities. It shall be kept in mind, however, that Text and Data Mining exceptions can allow reproductions, but not large-scale communication to the public.

In the United States, copyright law seems more favorable for crawling activities, with doctrines such as fair use or implied license. The inconvenience of the American approach, however, is a certain lack of legal certainty.

As far as the *sui generis* database right is concerned, it is also relevant for web crawling activities, as many websites can be regarded as databases. Extraction and re-utilization of substantial parts (i.e. more than 10%) of such websites (as well as repeated and systematic extraction of their non-substantial parts) requires authorization from the rightholder. Statutory exceptions to this exclusive right provide even less relief for web crawling activities. The *sui generis* database right does not exist in the United States.

Crawled content can also contain personal data, the processing of which is highly regulated in EU law. In practice, crawled data shall undergo thorough anonymization, ideally already at the collection stage.

Moreover, web contents are often available under terms and conditions which may altogether prohibit any web crawling activities, even if they are allowed by statutory exceptions. The enforceability of such clauses, however, may sometimes be questioned. On the other hand, many public licenses (such as Creative Commons) entail broad permissions to copy and re-use the licensed content.

5.2 Roadmap for Web Crawlers

It seems that the most viable way of making sure that the crawling operations are lawful is to perform an *a priori* clearance of the sources that are to be crawled. It shall be checked whether the contents available via the list of URLs are:

- protected by copyright: this excludes public domain material such as official works (in some jurisdictions), works whose authors died more than 70 years ago or material that does not meet the threshold of originality;
- protected by the *sui generis* database right: this assessment is, arguably, very difficult to make.
- If both questions are answered in the negative, the content can be crawled;
- if the answer to at least one of these questions is in the positive, the content can be crawled only if it is available under a public license (such as Creative Commons) or with a notice that allows crawling (i.e. free reproduction and communication to the public). Of course, the conditions of the license or notice shall always be respected (see above).
- If the contents are held by a public sector body (with the exclusion of educational and research establishments as well as cultural establishments other than museums, libraries and archives), and they are not already available under a public license or with an appropriate notice, it is possible to contact the public sector body to request a license for the re-use of the contents of the website. Such requests should follow the procedure laid down in the national law on the re-use of public sector information of the country in which the public sector body is situated.

Only the sources that pass this validation procedure (e.g. non-protected by an exclusive right, available under a public license or with a notice, or constituting Public Sector Information, the request for re-use of which was answered favorably) can be lawfully crawled. Even in such cases, the data obtained shall be anonymized (taking into account any means likely reasonably to be used to re-identify the data subjects).