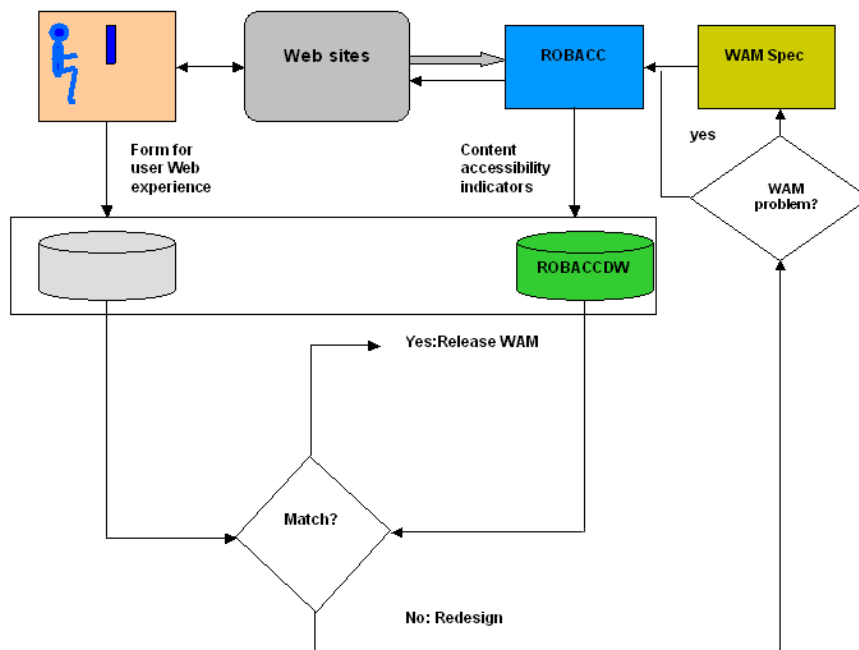




## European Internet Accessibility Observatory – some legal issues

The Norwegian Research Center for Computers and Law has contracted with the EIAO to develop a report discussing some legal aspects of the service suggested by the EIAO. This service is measuring accessibility of websites using an approach illustrated by the diagram below:<sup>1</sup>



**Figure 1: Elements of the European Internet Accessibility Observatory**  
(The relations among the elements indicate how user testing is planned to improve the automatic evaluation.)

<sup>1</sup> Jenny Craven and Mikael Snaprud “Involving Users in the Development of a Web Accessibility Tool”, <http://www.ariadne.ac.uk/issue44/craven/intro.html>.

The report is based on a *pro memoria* in which the EIAO has indicated the issues which are seen as relevant, and which has been discussed with the NRCCCL to arrive at a common accepted description of the issues.

This *pro memoria* has been the basis for the report.

Obviously, in developing the report, we have had to make several compromises. We have tried to keep the project of EIAO present, and not get lost in legal arguments which may be of interest, but of little relevance to the project. Often we have chosen to include a paragraph summing up the discussion to make the result of the legal argument more available.

It will be appreciated that the report does not expand on the legal issues. The discussion of copyright, the legal protection of geographical data, and the liability issues associated with the e-commerce directive, the application of the directive on re-use of public sector data all could be expanded into rather extensive discussions. We have tried to rim the discussion to fit the objective, and – of course – the resources made available for our small contribution to the project.

The report is written on behalf of the NRCCCL by Professor Jon Bing.

## 1 Lists of hyperlinks

*EIAO needs access to high-quality lists of URLs to public services. URL databases are probably copyrightable according to the EC database directive, so the EIAO may not be able to redistribute URL lists that were purchased and it is not clear if it would be legal to publish results based on these URLs, unless we have collected the URL lists ourselves. That will have to be investigated/negotiated for each individual case.*

### 1.1 Database protection

According to the Database Directive<sup>2</sup>, a certain legal protection of databases is established. Part of the directive deals with *copyright protection* of databases, which presumes that the database qualifies as a copyrighted work according to the traditional criteria, especially the criterion of originality. We will not deal with this protection. A hyperlink in the form of an URL will by definition not be an original work for several reasons, the most obvious being that if the URL was “original” in copyright terms, the probability that is at the same time would correspond to a live URL and therefore represent a function link, would be very small. Our concern is therefore with the *sui generis* protection of the directive Chapter III (art 7 ff).

Prior to the enactment of the directive, some of the member states did offer protection of databases, *ie* Denmark, Finland and Sweden. When the Nordic countries revised their copyright legislation at the end of the 1950s and beginning of 1960s, this was done on the basis of a report developed in cooperation between the countries. In this was proposed a rule for protection of some non-copyrighted material where a need for protection had been identified, examples being time-tables, radio programs, telephone directories and other types of catalogues – the provision was known as the “Nordic catalogue rule”, and became part of the Norwegian copyright law as section 43.

When databases become of commercial interest, it was rather trivial to observe that these were a type of material which typically fitted the protection of the catalogue rules. Within the EU, there therefore were some jurisdictions offering legal protection to databases though these did not meet the criteria for copyrighted work.<sup>3</sup> Also, there were differences in the interpretation between the member countries to what extent material needed to be “original” in order to be qualified as a copyrighted work.

In such a situation the law created differences inappropriate for the single market. But the major reason for introduction a database protection, was to promote investment in the emerging market for information services.

### 1.2 Matter of protection – hyperlinks

```
<A href="http://www.eiao.legal.no">EIAO LEGAL</A>
```

This is a typical hyperlink. If a user doubleclicks the text “EIAO Legal” as it occurs in the web page, the user will access another page. The html-formalism relies on tags (< and >) to indicate which part of the text is included in the formalism, and which therefore should be processed by a program made to present such pages’ in this HTML corresponds roughly to a programming language.<sup>4</sup>

The link has to ends called “anchors”, and a direction. The link starts at the source anchor (often just called the link), and points towards the destination anchor or link target. The text indicated on the screen, in this case “EIAO Legal”, is called the link label.

<sup>2</sup> Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases.

<sup>3</sup> Cf Database directive Preamble item 4.

<sup>4</sup> But is not, cf Urheberrechtsschutz von HTML-Quelltext, Oberlandesgericht), Frankfurt, 22 March 2005.

Generally, the target for a link is a Unique Resource Locator, as this is used within the Internet system for names and addresses. The first part (in the example “www”) indicates the protocol to be used with transferring data to or from the resource. This is followed by the IP-address or the domain name of the resource, in the example “eiao.legal.no”, the top level domain in this case being the country code domain of Norway (“no”). The registrant for this top level domain will only accept one registration for “eiao”, which therefore will be unique within the domain.

Our first concern is whether collections of links can be databases within the terms of the directive. The terms of the directive are rather inclusive. In the definition of art 1(2), a database is defined:

“For the purposes of this Directive, 'database' shall mean a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means ...”

In the preamble,<sup>5</sup> it is mentioned that the elements of the database may be:

“... literary, artistic, musical or other collections of works or collections of other material such as texts, sound, images, numbers, facts, and data; ...”

As far as we know, there are few decisions directly addressing hyperlinks. We have surveyed the decisions relating to the directive, using the database of the Institute for Information Law at the University of Amsterdam,<sup>6</sup> and found the following examples:

**Amtsgericht Rostock** 20 February 2001<sup>7</sup> decided a case concerned the copying of a website containing a collection of links which were organized by category. The court quickly reached the conclusion that the collection of links was sufficiently organized and individually accessible to constitute a database. The court elaborated on the substantial investment criterion. An investment was held to be substantial if it has substantial weight (“substantielles Gewicht”). Explicit reference was made to the English rule of thumb “What is worth copying is worth protecting”. In order to achieve the aims of the Directive, a low standard of substantiality should be applied, and small databases should be protected as well. A line should be drawn only at very simple databases. Although defendant’s database was not 100 per cent copied, a substantial part had indeed been copied.<sup>8</sup>

**Landgericht Köln.** Beschluß vom 12. Mai 1998 - 28 O 216/98 - “*Linksammlung als Datenbank*”.<sup>9</sup>

These two decisions are by themselves not sufficient to establish the principle that collections of hyperlinks may be protected databases under the database directive. It is rather a confirmation of what must be the straightforward interpretation of the directive – hyperlinks are examples of data that may be organised in a database and be protected by the *sui generis* database right. There is no reason to dwell upon this aspect – as the cases demonstrate, this is not really doubtful. But, of course, such a collection of hyperlinks must meet the usual criteria to qualify for protection.

<sup>5</sup> Item 17.

<sup>6</sup> Cf <http://www.ivir.nl/files/database/index.html>.

<sup>7</sup> Cf <http://www.jurpc.de/rechtspr/20020082.htm>.

<sup>8</sup> *Multimedia und Recht* 9/2001:631-632.

<sup>9</sup> Jfr <http://www.online-recht.de/vorent.html?LGKoeIn980512+ref=Urheberrecht>.

### 1.3 The two alternative qualification criteria

#### 1.3.1 The cost criterion

According to the directive art 7(1), the *sui generis* database right is awarded to

“... a database which shows that there has been qualitatively and/or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents to prevent extraction and/or re-utilization of the whole or of a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database ...”

The catchphrase for the criterion is “substantial investment”. The provision has a rather complex structure, and we will refrain from a fuller analysis in this context. In practice, what has emerged as an important relation is that between the investment and the reasons for this investment, which should be the disjunctive objectives of “obtaining”, “Verification” or “presentation”. A database is often part of a service.

Take, for the sake of argument, an archive for unique documents. The collection of these documents has required large investments, and for the research of the documents, they are carefully examined and classified. The result of this classification is then made available in a database which can be accessed over the web. Is this a protected database? It is obviously related to substantive investment, but the acquisition of the documents, their preservation *etc* is not really part of establishing the database, and it may be argued that these costs should not be taken into consideration. Also the classification has not as its main objective to make a database, but to make a scientific sound description of the document for further study and research, the classification would have been carried out even when no database was planned. Therefore, it may be argued, this is a cost not to be considered relevant. One is then left with the cost of collecting the registered data from a word processing system and inputting them into an appropriate database system for publishing the material on the web. This is a marginal cost, and does not qualify as a substantial investment.

The criterion of investment is therefore rather slippery. One should be careful in the analysis to assign the cost to the “obtaining, verification or presentation” of the data. There are examples of courts having deemed databases “spinoffs” from major efforts, and therefore not awarded protection to the databases. In my opinion this hardly can be justified unless the database is a separate by-product of the major project. In most cases, the database is an integrated part of the project, but not the only objective. One will therefore not be able to avoid case-by-case evaluation of databases.

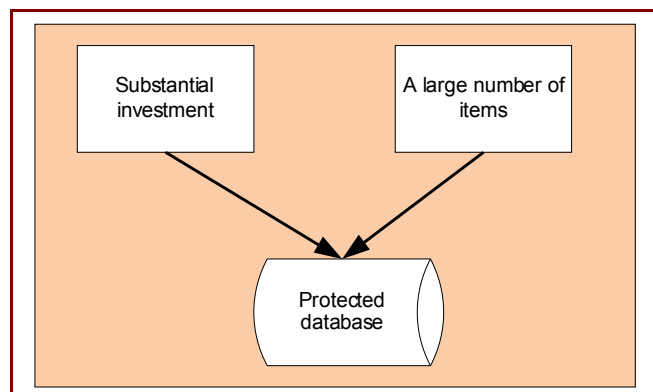
In my opinion, one should be guided by the principle that there should not be very difficult to obtain protection of a database. In projects where the database is just one of several objectives, it may nevertheless be that the database – which provides access to the material – is a necessary element, without which the other objectives cannot be realised.

This is just an introduction. The complex provision has several other criteria which deserve an analysis. But it would seem out of context to conduct a full legal analysis in general and in the abstract in this context. It must be sufficient to conclude that the cost criterion may present problems which have to be addressed in the concrete context in which they appear

#### 1.3.2 The quantitative criterion

As mentioned above, the directive recognises that there are in existence other regimes for the protection of databases among member countries. The directive does not require these countries to terminate their current regime, but permit this to be continued as the country also includes the *sui generis*-protection of the directive. Norway has through the EEA-agreement included the

database directive.<sup>10</sup> Consequently, there are two different disjunctive criteria for obtaining protection as a database. And the object of the EU to smoothen out the differences between jurisdictions has not been fully realised.



Above, we briefly discussed the criterion of substantial investment. The criterion used under Norwegian law, has been the number of “data” included. In principle, it may be queried whether the nature of data<sup>11</sup> permit them to be counted, but this largely remains a theoretical issue. In a pragmatic context, this does not cause much doubt.

The Nordic catalogue rule dates from before computerised systems become generally available. The material protected were conventional paper-based examples like a seed catalogue, a time-table for when planes take off from a certain airport, the publication of radio programmes, certain indexes or tables *etc.* There has not been much discussion of what the criterion really was to be interpreted, for instance how many items had to be included for the criterion to be satisfied.

The criteria are disjunctive, but they will in practice obviously have a large intersection – most databases containing a large number of items will also require a substantial investment. The legislative history in Norway<sup>12</sup> presumes that there may be examples where the number of items is rather small, but has required high investment. As practical would be that there is a large number of items, but that these have been harvested as a by-product of some other process, and therefore are *not* subject to a large investment. There are several examples of cases at European courts demonstrating this situation.

The best known case may be the European Court of Justice in its decision<sup>13</sup> on the British Horseracing Board. In this case, the Court discussed the “investment” criterion and its relation to other elements in the service offered by BHB:

“The expression ‘investment in...the obtaining...of the contents’ of a database in Article 7(1) of the directive must be understood to refer to the resources used to seek out existing independent materials and collect them in the database. It does not cover the resources used for the creation of materials which make up the contents of a database.

- The expression ‘investment in...the...verification...of the contents’ of a database in Article 7(1) of the directive must be understood to refer to the resources used, with a view to ensuring the reliability of the information contained in that database, to monitor the accuracy of the materials collected when the database was created and during its operation. The

<sup>10</sup> European Economic Area Agreement, Appendix XVII on intellectual property right, by approval of the Parliament 15 May 1997 of the decision by the EEA-committee no 59/96 (cf [St prp nr 36 \(1996-1997\)](#), [Innst S nr 182 \(1996-1997\)](#)).

<sup>11</sup> The Norwegian corresponding term is ”opplysninger”.

<sup>12</sup> *Rettslig vern av databaser*, Ot prp nr 85 (1997-1998).

<sup>13</sup> [The British Horseracing Board Ltd and Others v. William Hill Organization Ltd](#), ECJ, nr. C-203/02, 9 November 2004

resources used for verification during the stage of creation of materials which are subsequently collected in a database do not fall within that definition.

- The resources used to draw up a list of horses in a race and to carry out checks in that connection do not constitute investment in the obtaining and verification of the contents of the database in which that list appears.”

It is easy to argue that though failing to qualify as protected according to the investment criterion, there were sufficient items collected on the horses and races for there to be a large number, and that the database therefore would be protected under Norwegian law.

### 1.3.3 Summing up

This introduction of the qualification criteria of the database protection is incomplete. It would not be appropriate to extend the discussion on a theoretical basis. It is sufficient to demonstrate the core of the protected material. For EIAO the more important database protection is the general European protection relying on the directive, and being based on the investment criterion.

## 1.4 Collections of hyperlinks

### 1.4.1 Explicit collections

Hyperlinks are basic parts of the architecture and functionality of the web. They have a double nature. First, they are an explicit reference to a certain resource, typically an URL, within the web. As such, they may be written in any way which convey the necessary information to a reader, for instance the “@” may be replaced by “at”. Second, they offer certain functionality. When activated – a web browser will permit the user to activate a link by clicking the cursor on the label, and will often change the outline of the cursor when placing it over a label – a request will be made to have the resource identified by the link, transferred to the user. Therefore, a hyperlink has both an informative and functional aspect.

Hyperlinks are often collected for certain purposes. An organisation may present a web page of hyperlinks as an gateway to different services, publications *etc.* This may be *internal references* – then the collection is a navigation tool for the site or sites of the organisation itself. An example is shown below; this is an extract of the references offered by Eurostat to news releases:

<b>Euro area inflation estimated at 3.3%</b> Release Date: 30-APR-2008 11:00 AM <a href="#">Economy and finance</a>
<b>Euro area unemployment stable at 7.1%</b> Release Date: 30-APR-2008 11:00 AM <a href="#">Population and social conditions</a>
<b>Industrial new orders up by 0.6% in euro area</b> Release Date: 23-APR-2008 11:00 AM <a href="#">Industry, trade and services</a>
<b>EU27 current account deficit 2.8 bn euro</b> Release Date: 22-APR-2008 11:00 AM <a href="#">Economy and finance</a>
<b>A EU27 external trade deficit of 34 bn euro with Japan in 2007</b> Release Date: 21-APR-2008 11:00 AM <a href="#">External trade</a>
<b>Euro area and EU27 government deficit at 0.6% and 0.9% of GDP respectively</b> Release Date: 18-APR-2008 11:00 AM <a href="#">Economy and finance</a>

Figure 1 - EUROPA - Eurostat - Page News release

They may also be *external references*, pointing to resources with which the current organisation is co-operating or otherwise sees as useful for those visiting the site. An example may be the external links of the European Data Protection Supervisor



Figure 2 - EDPS external links

In both cases, the collections of hyperlinks are presented as just that, *collections*. They are small “databases”. Whether protected under the database directive, will mainly rely on whether they are the result of a substantial investment, under the catalogue rule would also the question of whether they represented a sufficient high number of items to be protected.

These are really not example of different types of collections, though it is characteristic that such lists contain either external or internal links. Obviously, the lists may contain both internal and external links; this does not in our context have any relevance. The important aspect is that they are developed (and maintained) as explicit collections of hyperlinks. There is no question that such collections are to be qualified as databases under the European database directive. However, it is very doubtful that these will be awarded a separate legal protection. In that case, it would have to be demonstrated that the links themselves have required substantial investment.

Regrettably, one cannot make any conclusion in principle. There is obvious the possibility that such lists are the result of a substantial investment. One would then look for lists of a much larger scope than these snippets (also the originals are rather brief lists).

One would also look for lists which are not generated as a side effect of other activities – for instance, it may be argued that the list of Eurostat to press releases is generated as a side effect of releasing the story, the database builds as historically the lists gets longer. It may be costly for a third party to collect the data for establishing an identical list, but for Eurostat it has not represented any substantive investment to establish the list over time.

As one cannot make any conclusion in principle, it may be precarious to make a general assumption. Nevertheless, it would seem that to find a list of explicit hyperlinks protected by the *sui generis* database right would not be typical.

#### 1.4.2 *Implicit collections*

The hyperlinks are part of the basic architecture of the web. There will be hyperlinks to and from any page on the web. Search engines actually exploits these when indexing the web – having access to a page, the engine will identify to which pages this page is linking (“to-links”), and use this data to access also these pages for expanding the indexing.

The indexes from a page to other pages may be seen as *implicit* collections. They have not been constructed to be collections of hyperlinks. Rather, it is the content and function of the page which is the reason for the author of that page to introduce hyperlinks. But a third person may observe that this page is about “European data protection”, assume that the pages linked to

from this page is related to the subject of the page, and harvest the hyperlinks as a collection relevant to the subject of the page.

Here it may be relevant to question whether this implicit collection qualifies as a database under the directive. The definition calls for a database to be “arranged in a systematic or methodical way”. In the implicit collections, the arrangement is by relation to the content, not in obedience to any syntactic or schematic way. The links may be identified using algorithms which recognise the formalism of a hyperlink in HTML, but are not organised to be retrieved as elements.

It is doubtful whether an implicit collection of hyperlinks will qualify as a protected database. But there are probably borderline cases between the two typical examples discussed in this report for there to be ample room for doubt in the concrete context.

### **1.5 Re-utilisation as infringement**

The database directive art 7(1) specifies the exclusive right of the maker of the, the maker has the right

“...to prevent extraction and/or re-utilization of the whole or of a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database ...”

This may be compared to art 8, where it is stated that a lawful user will be allowed to extract or re-use the material of the database in whole or of a substantial part. This amounts to an exclusive right of the maker which cannot be exploited without the consent of the maker, a consent which typically will be available on the condition of payment. This is the major exclusive right of the maker, and it is directed towards the whole or a substantial part of the database. A user will normally access a database to find the answer to some query, for this it is completely unnecessary to extract the database as whole or in substantial part – it suffices with a response containing one or a few items. This exclusive right of the maker does not, therefore, address the normal use of a database. If a third party, without consent of the maker of the database, was given access to the database and used it in a normal fashion, this would not be an infringement of the initial exclusive right.

However, the directive art 7(5) extends the right of the maker of the database:

“The 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 shall not be permitted.”

This makes it an infringement to extract and re-use “insubstantial parts” of the database if done repeatedly and systematically. This secondary right of the maker protects against what would be a normal use of the database. It would not be relevant for a third party who occasionally accessed the database; the action has to be “repeated and systematic”. There is also the further requirement that the access either shall conflict with “normal exploitation” or “prejudice the legitimate interests” of the maker of the database. An example of the first would be if the action made it more difficult for lawful users to access the database, for instance because capacity was reduced. An example of the second requirement would be that the action reduced the revenue of the service in which the database is integrated.

A possible use of a database by a third party would be to set up a system which regularly would access the database, extract some of the hyperlinks and use these to gain access to the resources to which they are anchored. For the question of infringement to arise, one has to presume that the collection of hyperlinks qualify as a protected database under the directive, see sect 1.3 above. A hyperlink would then be an item of the database, and alone or with a few other

items, it would represent an “insubstantial part” of the database. If the third party set up a systematic scheme for utilising such hyperlinks, for instance extracting and checking them at regular intervals, it would qualify as a “repeated and systematic” extraction. Finally, one would have to determine whether this conflicted with the normal exploitation of the database, or prejudiced the legitimate interests of the maker of the database. In most cases, such a scheme would have no impact on the utilisation by other users or the operation of the database. But if the maker of the database offered the services for payment, setting up a way to utilise the services without payment would prejudice the interest of the maker in generating revenue. In such a case the scheme by the third party would be an infringement.

It is difficult drawing a conclusion with respect to EIAO. An assessment must be based on a concrete example. If there is a website posting a database of hyperlinks to sites which all are to be evaluated by EIAO, the exclusive right of the maker of the database may require consent (or contract) for the extraction in whole or substantial part. If EIAO establishes schemes which access websites, check the collection of hyperlinks for changes, and extract such changed hyperlinks, this will represent a repeated and systematic extraction. Whether it is an infringement of the secondary right of the maker of the database, will rely upon the final requirements that such extraction has to conflict with normal use of the database or unreasonably prejudice the interests of the maker.

The conclusion therefore is less certain than desired.

### **1.6 Implicit licence**

As mentioned above, consent from the maker of the database will make it lawful to utilise the database (or the service in which the database is integrated). The obvious example of such consent would be a subscription contract. To make a subscription contract operational, the maker of the database has to be able to distinguish between those registered as lawful users, and those without any prior contract. The typical way to implement this is by assigning the user a name and a corresponding password.

Database protection is not dependent upon such arrangement. It is quite usual that databases in the form of collection of hyperlinks are openly accessible on the web without there being any form for protection, any system for the registration of users *etc.* The examples in sect 1.4 above are such collections.<sup>14</sup>

The question arises how to interpret this lack of access control on behalf of the maker of the database. For a user, the website appears to invite third parties to make use of the website, including the integrated databases of hyperlinks. The maker of the database will know that in making the database available on the web, users will access and exploit the database. It may be argued that the failure to integrate any form of protection or barrier to such exploitation should be interpreted as consent.

Consent is a one sided disposition of a legal nature. In this case, it is related to a gift – the recipient accepts no obligations, but the maker of the database makes a “gift” of the rights to utilise the database. Typically, such a disposition would be drafted in the form of a document, or will be spelled out explicitly on the website. This rarely is done. But a legal disposition may also take the form of an act, the handshake being a well-known example. It may be argued that if making a protected database openly available on a website, with the knowledge of how material is utilised on the web, this itself is an act which may be interpreted as a unilateral disposition inviting anyone to use the material in the way they find useful.

This may be even clearer where the maker of the database have no revenue interest in the material. The maker of the database has not made any arrangement for its licensing, and there is no payment schemes associated with the database. In such a case, the secondary protection hardly will take effect; a systematic re-utilisation of the database will hardly prejudice any of the legitimate interests of the maker, at least not the interest in revenue. There may still be the

---

<sup>14</sup> Though for different reasons, these examples may not represent protected databases.

primary protection, which does not depend on this requirement, ensuring that the maker has the exclusive right to the database as a whole or in substantial part.

An important sub-category of makers of databases are public agencies. It is presumed<sup>15</sup> that also a public agency may be maker of a database and benefit from the protection. If such an agency make a database of hyperlinks available, it would be rather obvious that this was an invitation to any third party to make use of this as the third party saw fit. The argument of implicit licensing becomes very strong. There would be difficult to find any legitimate interest of a public agency in controlling third party use. If the public agency for some reason sees the need to reserve the database for a limited number of users, or only for some reuse, the obvious solution would be to introduce some subscription scheme or access control system.

It may be argued that the theory of implicit license has some basis in European court decisions, cf the Austrain case cited below:

***ADV-Firmenbuch II*** (Oberste Gerichtshof, 28 May 2002).<sup>16</sup> Plaintiffs operate the Republic of Austria's online database containing the official Austrian company register. The defendants had used data from this database to update their own databases (note defendants are the same as in the ***ADV-Firmenbuch II*** case decided by the Supreme Court on 9 April 2002, see below). The court held that although the defendants had indeed infringed the Republic's database right, this did not mean that they had acted unlawfully (*ie* competed unfairly) against the plaintiffs.

### ***1.7 Limited protection of public sector material***

According to the Bernconvention art 2(4)

It shall be a matter for legislation in the countries of the Union to determine the protection to be granted to official texts of a legislative, administrative and legal nature, and to official translations of such texts

This provision applies to material which otherwise would be works protected by copyright, and does not directly apply in the case of databases of hyperlinks. It is cited only to indicate the possibility that there may be a similar provision for databases. A collection of hyperlinks at the site of a public agency may certainly be related to legislative, administrative or other legal functions, though not the “text” of an instrument of such a nature.

The provision is found in the Berne Convention, the major copyright convention. It has not been co-ordinated by any EU directives; consequently one will find that countries have taken advantage of this possibility of excluding some public documents from copyright protection to different degrees. Some countries let the public hold full copyright, with the English doctrine of Crown Copyright as the prime example. Some exclude such material according to different criteria.

Under Norwegian law, the exception is implemented as Copyright Act sect 9, and extends to any document concerning the exercise of public authority. The provision exclude the application of “this statute”, therefore also the protection of databases is excluded. And it would be rather easy to find examples where collection of hyperlinks were related to the exercise of public authority – the links may for instance have decisions by a public agency as their target.<sup>17</sup>

<sup>15</sup> Below under sect 1.7 some aspects of this are briefly discussed. The relation to Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information will not be discussed.

<sup>16</sup> Cf <http://rechtsprobleme.at/doks/urteile/firmenbuchII.html>.

<sup>17</sup> Cf for instance the web site of the Norwegian Data Protection Tribunal, <http://www.personvernemnda.no/vedtak/index.htm>.

However, the database directive art 9 enumerate an exhaustive list of the exceptions to the exclusive right for re-utilisation of the whole or a substantial part of the database:

- a) in the case of extraction for private purposes of the contents of a non-electronic database;
- b) in the case of extraction for the purposes of illustration for teaching or scientific research, as long as the source is indicated and to the extent justified by the non-commercial purpose to be achieved
- c) in the case of extraction and/or re-utilization for the purposes of public security or an administrative or judicial procedure

In this list, there is no exception for public sector material. Therefore one cannot presume that there is with respect to the database right a similar limitation of the exclusive right as permitted for copyright by the Berne Convention art 2(4).<sup>18</sup> This has also been confirmed by court decisions in Europe. The first of these relates to the Austrian company registry, and may therefore be of special interest to EIAO.

***ADV-Firmenbuch I*** (Oberste Gerichtshof), 9 April 2002).<sup>19</sup> Plaintiff, the Republic of Austria, offered an online version of the official Austrian company registry. The defendants used data from the registry to update their own databases. According to the court, the defendants had extracted a substantial part of the database without paying an appropriate remuneration. The Court rejected application by analogy of the copyright exemption which allows free use of government information. According to the Court, this would not have been allowed under the Database Directive's exhaustive set of permitted limitations. The Court concludes that defendants have infringed the Republic's *sui generis* database right. Defendants are enjoined from extracting data without paying adequate remuneration. However, the Republic may not demand inappropriate remuneration for the extraction. In the light of the "essential facilities" doctrine, this would otherwise amount to an abuse of a dominating position on the market, in view of the defendants' complete dependence on the data of the Republic's website.

This is also held in a German decision. It is of some interest because it relates to statutory material, which is the central example of what may be excepted according to the Berne convention art 2(4).

***Compilation of laws protected as database.*** (Landgericht Munich, 8 August 2002).<sup>20</sup> Plaintiff published legal texts (laws) in print, on CD-ROM and on-line, some of which were reproduced without permission on defendant's web site. The court considered this collection to be a database. The substantial investment was found in many years of gathering and updating the collected legislation. Considering the fact that a large number of grammatical errors on the defendant's website were identical to the ones on plaintiff's website, the court held that a substantial part of the contents of the database had been extracted and that there had been an infringement of the *sui generis* database right. Although laws and statutes are exempted from copyright protection, the court refused to apply this exemption by analogy to the plaintiff's *sui generis* database right, since the exhaustive list of exemptions in Article 9 of the Database Directive would not allow this.

<sup>18</sup> This would imply that the provision of the Norwegian Copyright Act sect 9 may be in violation of the database directive, or should be interpreted to maintain conformity, but see argument at the end of this section.

<sup>19</sup> Cf <http://rechtsprobleme.at/doks/urteile/firmenbuchI.html>.

<sup>20</sup> Cf <http://www.jurpc.de/rechtspr/20020369.htm>.

An Icelandic decision may be of special interest to the EIAO. It also addresses the exception of copyright by some state-created works, in this case map, or more generally, geographical information systems.

***Sui generis right to database of geographical maps.*** (Høyesterett 19 September 2002).<sup>21</sup>  
 In 1997 M purchased from the State Geographical Institution three themes (aerial lines, water and roads) from a computerised database in the scale 1:250.000. According to the conditions of purchase signed by M, permission from the State Geographical Institution was necessary for reproduction or making documents, based on the purchased information, available to the public. In 2000 M published without permission from the State Geographical Institution some maps in the scale 1:600.000, 1:300.000 and 1:100.000. The State Geographical Institution took legal action against M and asked for a fee according to the tariffs of the Institution. M claimed that Article 10 of the Statute on the State Geographical Institution (Statute 95/1997) did not give a legal basis for claiming payment for the use of data from the computerised database. M argued that the Icelandic state could not claim copyright in the digital material, and that it therefore could not ask for a fee. In its decision of 19 September 2002 the Supreme Court of Iceland confirms the arguments of the Institution on the rights of the state to the computerised material. This constitutes a database in digital form where the content consists of a derivation of a map to which the Icelandic state owned the copyright according to Article 1 of the Copyright Act. The establishment of the database had required considerable investment. The state therefore enjoyed the rights under Article 50 (*sui generis* protection) of the Copyright Act to this material. The Supreme Court found that the Statute 95/1997 gave sufficient basis to claim a fee for the sale and use of geographical information. M could not demonstrate that the State Geographical Institution in its tariffs, which had been confirmed by the Ministry, had asked for a disproportionate fee, or that the fee in any way was unreasonable. M was therefore sentenced to pay the State Geographical Institution the requested fee

Seeing these cases, it would seem that the only conclusion is that the database directive excludes the possibility of reduced protection for some governmental material. The argument is that the database directive art 9, which enumerates the exceptions that can be exploited by member states, does not include this alternative. However, there is another possibility:

***Vermande v Bojkovski*** (Arrondissementsrechtbank, The Hague 20 March 1998).<sup>22</sup>  
 The Vermande/Bojkovski case, which was decided prior to implementation, concerned the unauthorized publication on a web site of laws and regulations copied from a commercially published CD-ROM. According to the District Court of The Hague, the Database Directive did not permit a statutory limitation of the *sui generis* right in respect of such compilations. Under the Directive the CD-ROM publisher would, therefore, have been protected. However, since implementation had not yet been completed, and Article 11 of the Dutch Copyright Act clearly places laws and regulations in the public domain, no injunction was granted.

The case confirms the view stated above. However, it also discloses the legal policies relating to the intellectual property protection of legislative text, and did therefore initiate a development of Dutch law. In the database directive one finds art 13, which relates mainly to other areas of law:

<sup>21</sup> Abstract by Erla S Árnadóttir. Cf <http://www.haestirettur.is/domar?nr=1558>.

<sup>22</sup> Cf <http://www.ivir.nl/rechtspraak/vermande-en.html> (English version).

This Directive shall be without prejudice to provisions concerning in particular copyright, rights related to copyright or any other rights or obligations subsisting in the data, works or other materials incorporated into a database, patent rights, trade marks, design rights, the protection of national treasures, laws on restrictive practices and unfair competition, trade secrets, security, confidentiality, data protection and privacy, access to public documents, and the law of contract.

On this basis, the Dutch database legislation was enriched by a provision stating that “laws, decrees, ordinances, as well as court and administrative decisions” are not protected by the *sui generis* database right, relying on the reference to “access to public documents”, which is a reference mainly to the freedom of information legislation. This would seem to put the argument back to the point of departure. The member countries may decide to implement this general exception in the tradition of the provision in the Berne convention art 2(4).<sup>23</sup> One is therefore left with having to consult the national legislation.

But in this argument, we have had examples of the protection of databases containing company names and geographical features, in both cases owned by the state, and in both cases upheld as protected against the reproduction of substantial parts, and the secondary infringement of systematic re-utilisation.

### 1.8 Summing up

In this section we have examined the database protection of collections of hyperlinks. We have determined that collections of hyperlinks can be a database as the term is used in the database directive. We have also determined that such collections may qualify for protection as a database, and found at least one example in European caselaw for this to be the case. Re-utilisation of such a collection of hyperlinks relies on the consent of the database maker.

However, we have looked at three possible limitations in the protection.

- The secondary protection, the re-utilisation of a non-substantial part of the database, must by either impair the operation of the database, or prejudice the legitimate interest of the maker of the database, to be unlawful. It is suggested that this often will not be the case when the database of hyperlinks is available free of charge on the web. There is at least one case which seems to be based on the same argument.
- Second, if a database of hyperlinks is available on the web without protection mechanisms, it may be argued that the maker of the database in uploading the database to the web has implicitly consented in the use of the database. Such implicit licence may justify the re-utilisation of hyperlinks from non-commercial databases.
- Third, there is in copyright law a tradition for excluding some public sector instruments from protection. This would seem not to apply to the *sui generis* database protection, as the directive does not have a corresponding exception. However, at least one member country has exploited a general provision indicating the interface between database protection and other legal schemes, including freedom of information, to implement such an exception in their national law. Therefore, one may have to look at the national law in question to determine whether it has such an exception.

---

<sup>23</sup> This argument can also be used for the Norwegian provision in the Copyright Act sect 9, briefly discussed above.

## 2 Categorising according to NUTS and NACE

EIAO needs to categorise accessibility measurements according to standardised geographical regions (NUTS) and standardised sector codes (NACE) used by Eurostat. Authoritative NACE and NUTS categorisation of companies is managed by the European Business Register (EBR) and underlying national business registers. Information in the EBR is copyrighted to the national authorities and the data is available under a restrictive licence, so it is not clear that authoritative sources is available to EIAO.

### 2.1 Nomenclature of Territorial Units for Statistics (NUTS)

The somewhat surprising acronym for these units is originally derived from French, “*nomenclature des unités territoriales statistiques*”. They are the standard codes for administrative units, used in the European statistics since 1980. The standard has been developed by the European Union, and therefore only covers the member countries in great detail. Due to changes in the borders for administrative units, there may be deviations between the defined areas and the actual administrative units.

There are three levels:

- NUTS 0: Two letters identifying the country, corresponding to ISO 3166-1.
- NUTS 1: One number or letter identifying sub-state region, covering the same units as ISO 3166-2, but with different codes.
- NUTS 2 or 3: One further level indicated by numbers, the assigned numbers will often not refer to existing administrative units.

NUTS is available on the web,<sup>24</sup> and is associated with what is called a “copyright notice”:

© European Communities, 1995-2008

Reproduction is authorised, provided the source is acknowledged, save where otherwise stated.

Where prior permission must be obtained for the reproduction or use of textual and multimedia information (sound, images, software, etc.), such permission shall cancel the above-mentioned general permission and shall clearly indicate any restrictions on use.

This clearly makes the nomenclature itself available for purposes like the EIAO.

### 2.2 Statistical classification of economic activities (NACE)

The EU established through Council Regulation (EEC) No 3037/90 the statistical classification of economic activities in the European Community, referred to as NACE Rev 1. In order to reflect the technological development and structural change of the economy, an up to date classification has been established by Regulation (EC) No 1893/2006 of the European Parliament and Council of 20 December 2006. The objective is to have uniform interpreted categories for classifying activities in the Community.

According to the regulation art 2(1), a NACE Rev 2 has four levels:

- A first level consisting of headings identified by an alphabetical code (sections)
- A second level consisting of headings identified by a two-digit numerical code (divisions)
- A third level identified by a three-digit numerical code (groups)

<sup>24</sup> Cf [http://ec.europa.eu/comm/eurostat/ramon/nuts/home\\_regions\\_en.html](http://ec.europa.eu/comm/eurostat/ramon/nuts/home_regions_en.html).

- A third level identified by a four-digit numerical code (classes)

The NACE is set out in Annex 1 to the regulation. An example may look like what is reproduced below:

▢	<b>A</b>	Agriculture, forestry and fishing
▢	<b>01</b>	Crop and animal production, hunting and related service activities
▢	<b>01.1</b>	Growing of non-perennial crops
▢	<b>01.11</b>	Growing of cereals (except rice), leguminous crops and oil seeds
	<b>01.110</b>	Growing of cereals (except rice), leguminous crops and oil seeds
▢	<b>01.12</b>	Growing of rice
	<b>01.120</b>	Growing of rice
▢	<b>01.13</b>	Growing of vegetables and melons, roots and tubers
	<b>01.130</b>	Growing of vegetables and melons, roots and tubers
▢	<b>01.14</b>	Growing of sugar cane
	<b>01.140</b>	Growing of sugar cane
▢	<b>01.15</b>	Growing of tobacco
	<b>01.150</b>	Growing of tobacco
▢	<b>01.16</b>	Growing of fibre crops
	<b>01.160</b>	Growing of fibre crops
▢	<b>01.19</b>	Growing of other non-perennial crops
	<b>01.190</b>	Growing of other non-perennial crops

The economic activities in member states are produced using NACE Rev 2 or a national classification derived from this. The classification as such is not protected; anyone may use the codes as defined in the classification. This also is declared in the reference associated in the Norwegian national statistics.

<b>Derived from:</b>	
NACE Rev.2	
<b>Dissemination allowed:</b>	Yes
<b>Updates possible:</b>	No
<b>Copyright:</b>	Not relevant
<b>Predecessor version:</b>	SN2002
<b>Successor version:</b>	Not relevant
<b>Legal base:</b>	
Council Regulation (EEC) No. 1893/2006	
<b>Change(s):</b>	
<b>Publications:</b>	
Standard for næringsgruppering - NOS D383	

Figure 3 - Norwegian reference to NACE Rev 2 for national statistics

### 2.3 European Business Register

The European Business Register is a European Economic Interest Group registered in Belgium and owned by its members. An EEIG is a legal person established in accordance with the Council Regulation (EEC) No 2137/85 of 25 July 1985 on the European Economic Interest Grouping.

The members of the EBR are company registers from 19 countries (including Norway). Accessing the EBR one can search for data on a certain company or a company officer. There are different types of reports available depending on country and company. EBR can always provide you with a Company Profile that contains some of the most important information on a business, for example company name, registration number, address, country of registration, date of registration, registration authority, legal form, current status, type of business activities, share capital, date of the latest annual account.

As we understand it, the EBR is based on the national company registers, and operate as a network of these registers. Therefore the EBR is not a physically maintained register or data base as such, but a virtual database established through the co-operation and networking of the national systems. Reports may be generated by data collected from several registers and organised in a common presentation form by the EBR.

The business registers are protected, but *not* by copyright law. There will be national solutions, but the Norwegian may serve as an illustration. There is national legislation creating a central register for businesses (Act 1985-06). This mainly governs who has to report to the register, in which form *etc.* It clearly states that anyone may have access to the register (sect8-1), there may be associated documentation in addition to the data registered, and access to this is governed by the freedom of information act. The same provision also authorises to levy a fee for access to the register. Data in different formats are available, there is a price list associated with this.

Access to EBR is integrated in the national service.

One may also access the EBR. To receive data, one will have to enter into a subscription agreement; this is to be concluded with the national representative of EBR. In Norway, this is the national register.

As mentioned, the national Norwegian legislation states that the data is accessible to all. When received, the data may be used freely, also communicated to third parties as part of a business transaction.

## **2.4 Summing up**

The issue is somewhat clarified. The classification codes as such are not protected, and can be used by EIAO without any restrictions. Data on the businesses themselves are also available, but for a fee. The ways in which one may re-utilise the data, will rely on the contract made through the national representative with the EBR. But on the basis of the Norwegian example, it will seem that there is no restrictions on re-utilisations (apart from establishing a competing register, which would be contrary to the act establishing the register).

There would seem to be no alternative for the EIAO but to establish a contractual relationship to EIAO for re-utilisation of the company data.

▪

### 3 Publication of barriers on a web site

*EIAO needs to investigate if there are any legal limitations for publishing detailed information about encountered barriers on a particular web site.*

As we understand this issue, EIAO plans to publish detailed data on encountered in accessing a web site, or other problems as qualified by the W3C/WAI Web Content Accessibility Guidelines (WCAG) for public web sites in the EU. A collection of web accessibility metrics (WAMs), based on the checkpoints developed by World Wide Web Consortium for the Web Content Accessibility Guidelines (WCAG) version 1.0 (with potential migration to 2.0), and the tools for automated data collection and dissemination are developed within the project, and will be used to measure the accessibility of a certain web site. Such data is planned available online from a data warehouse of collected accessibility data (ROBACC DW).

Obviously, negative data would also be negative for the reputation of site in question. If the site is offering a commercial service, this may have an impact on the revenue of the site. The issue of liability may be considered.

There is freedom of expression in Europe; the European Convention of Human Rights art 10 is a basis for national law. In order to be held liable, one will have to find a provision in national law which meets the test of the ECHR.

It is obvious that the freedom of expression includes negative criticism of business and public services.

There may be two limitations in national law.

- One would be for untrue allegations, which might be construed as libel – this would mainly relate to criminal liability.
- The other is for data to be published which by negligence of the EIAO was not correct. This would have a negative effect on the business generated by the service – this would mainly relate to civil liability, compensation for the economic loss (or the reduced profit) caused by the misleading information.

These two possibilities do not present any real risk for the EIAO. The data published will be the result of a test to measure compliance with a respected set of guidelines. The test will be carried out observing the applicable standards, and with care.

There are no barriers to publishing detailed data about identified barriers on a particular web site.

#### **4 Publication of barriers generated by authoring tool**

*Are there any legal limitations for publishing detailed information about encountered barriers generated by a particular authoring tool?*

This issue is closely related to the issue discussed in sect 3 above. There the issue is related to a web site, which was tested according to a certain procedure and against a set of published guidelines. The result would be an identification of barriers for accessing the site.

A web site may be generated using certain programs or other “tools”. The EIAO may disclose that using certain authoring tools the resulting website will have certain undesired properties; there is a functional relation between the properties of the authoring tool and the resulting barriers of the website.

Again, it would fall clearly within the freedom of expression to discuss such properties of the authoring tools. The negative characteristic obviously may cause reduced revenue by the service offering the tools. This may even be amplified; the disclosure of such negative characteristics has as its objective either to cause the tools to be improved, or to warn potential customers to find alternatives which give better results.

The risks are identical to those briefly mentioned under sect 3 above – and again these do not present any real problems for EIAO.

## 5 Legal effect of trespass clauses etc

*Are there any legal problems by not following robots.txt, to get as good mapping as possible for a web site?*

### 5.1 Introduction: Search engines and copyright infringement

To collect data of web sites, EIAO will use a web robot (ROBACC - ROBot assessing web ACCessibility) for automatic and frequent collection of data on web accessibility and deviations from web standards like the Web Accessibility Initiative (WAI)<sup>25</sup> guidelines. This will work like a search engine – it will follow links from one page to another, copy the data of the site to the EIAO site where it will be analysed, the analysis resulting in the publication of web accessibility metrics.

Descriptions of a robot or a crawler will often seem to indicate that the robot or the crawler – which in reality is a computer program residing on the web site of EIAO – analyse the web site “at” the site itself. But to make the processing necessary for the analysis, the accessed web site has to be copied, at least partially, to the system of EIAO. This presumes a reproduction of the web site in copyright terms, and reproduction is the exclusive right of the copyright holder.

This is by no means unique for the ROBACC, all search engines will have to reproduce the target site in order to index the text. Therefore there is an unsolved issue related to search engines, as they do not work on the basis of consent from the rightholder. In principle the reproduction is an infringement of copyright. But without search engines, the function of the web would be greatly impaired. To argue that this copyright issue should cause search engines to close down, is not a viable legal policy.

Today, we have to accept that there is tension between copyright law and the general functions of a search engine. A partial solution is to rely upon the fiction of an implied license, as briefly discussed above under sect 1.4.2.

It falls outside the scope of this report to discuss this issue in any detail. But as the NRCCCL has available a draft reports on the matter, we enclose this as an appendix. This discussion is of a general nature, and takes search engines similar to Google as its example. ROBACC does not have the same purpose or functionality as Google, therefore the appended report does not in detail correspond to an analysis of ROBACC.

### 5.2 Robot exclusion standard

The robot exclusion standard, also known as the Robots Exclusion Protocol or robots.txt protocol, is a convention to prevent co-operating robots, [web spiders](#) and other web crawlers from accessing all or part of a [website](#) which is otherwise publicly available. Using metatags in the HTML formalism may have a similar effect.

The instructions are directed towards the programs. They are in principle requests from the operator of the web site for a program accessing the site to refrain from certain operations. The request has by itself no legal effect. A legal effect requires there to be a rule which triggers some consequence by not observing the request. In our experience, there are at least two such set of rules.

---

<sup>25</sup> Cf <http://www.w3.org/WAI/>.

### 5.2.1 Copyright infringement

The first are based on copyright law. As briefly sketched above in sect 5.1, and discussed in some more detail in the appended paper, the reproduction by a search engine of a web site is relevant for the exclusive right of the rightholder. In order to avoid the formal consequence of this requiring the consent of the rightholder, the doctrine of implied license is introduced. This observes that the act of uploading material to the web is done with the knowledge of search engines and the way they operate. As the rightholder has not made any reservations, one must be permitted to interpret the uploading as an acceptance of the consequences of making material available on the web. One of these consequences is the indexing by search engines, and in the vast majority of cases, the rightholder also would want this, as users will be able to access the web site by finding a hyperlink to the site in the index produced by the engine.

But this argument is vulnerable when it can be demonstrated that the rightholder *did* make a reservation. The robot exclusion command may be interpreted as such a reservation, indicating to a computer program that the reproduction of the web site for analytical purposes is not permitted. As there is no consent, the reproduction will be unlawful and an infringement of the exclusive right of the rightholder. This exclusive right may be copyright to the material on the site, but also the database right if the site contains a database.

According to this argument, the robot exclusion command would make it a copyright infringement to reproduce the content of the site for analytical purposes. EIAO should therefore construct ROBACC to respect such commands.

### 5.2.2 Liability for caching – directive on e-commerce

The directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (“Directive on electronic commerce”) includes in Section 4 provisions for the liability of three types of operators on the web. One of these types are the operators that permit caching on the web, typically on proxy servers to enhance the capacity for communicating data through the Internet, cf the directive art 13. The provision has the form of a preclusion of liability, this is made conditional. Two of these conditions are

- c) the provider complies with rules regarding the updating of the information, specified in a manner widely recognised and used by industry;
- d) the provider does not interfere with the lawful use of technology, widely recognised and used by industry, to obtain data on the use of the information;

The interesting element in these conditions is the reference to rules “specified in a manner widely recognised and used by the industry”. Robot exclusion commands are such rules. The e-commerce directive therefore should be interpreted to make such commands relevant. If a site has embedded a rule using robot exclusion commands, and the operator of the caching service disregard such rules – or, to be more specific, the computer programs used by the operator of the caching service is not designed to take such rules into account – the result is that the operator may not claim the preclusion for liability offered by the e-commerce directive.

This provision does not directly apply to EIAO, which will not operate a caching service. The importance is more indirect, the e-commerce directive art 13 seems to be the first international instrument to explicitly recognise the legal relevance of “program-to-program” instructions: The robot exclusion command is not directed from a person to another person, but from a program or computerised representation of a command to a program, the robot, but it will have consequences for the liability of the person (legal or physical) operating the robot. This recognition of robot exclusion commands may be seen as a casuistic example of a more general principle of their legal relevance. This relevance may be evident also outside the area of application of the e-commerce directive.

### 5.2.3 *Trespass and taking unfair advantage of a competing service*

One possible such area is the relation between two competing services of the web. This has especially been discussed with respect to metasearch services that is a service which itself makes available other services.

An illustration may be the relation between Bidder's Edge and eBay. eBay is the major auction service on the web. Bidder's Edge was a metasearch service – a user would specify for Bidder's Edge the object he or she was looking for. Bidder's Edge would then access a number of auction services, it would identify those objects offered which satisfied the criteria of the user, and present these in a format which made it easier for the user to compare the offers. eBay objected to this service, which circumvented the usual path through its site. In the decision by the District Court for the Northern District of California for preliminary injunction,<sup>26</sup> the issue of "trespass" was one of the points discussed:

"Trespass to chattels 'lies where an intentional interference with the possession of personal property has proximately caused injury.' *Thrifty-Tel v. Beznik*, 46 Cal. App. 4th 1559, 1566 (1996). Trespass to chattels 'although seldom employed as a tort theory in California' was recently applied to cover the unauthorized use of long distance telephone lines. *Id.* Specifically, the court noted 'the electronic signals generated by the [defendants'] activities were sufficiently tangible to support a trespass cause of action.' *Id.* at n.6. Thus, it appears likely that the electronic signals sent by BE<sup>27</sup> to retrieve information from eBay's computer system are also sufficiently tangible to support a trespass cause of action.

In order to prevail on a claim for trespass based on accessing a computer system, the plaintiff must establish: (1) defendant intentionally and without authorization interfered with plaintiff's possessory interest in the computer system; and (2) defendant's unauthorized use proximately resulted in damage to plaintiff. See *Thrifty-Tel*, 46 Cal. App. 4th at 1566; see also *Itano v. Colonial Yacht Anchorage*, 267 Cal. App. 2d 84, 90 (1968) ('When conduct complained of consists of intermeddling with personal property 'the owner has a cause of action for trespass or case, and may recover only the actual damages suffered by reason of the impairment of the property or the loss of its use.') (quoting *Zaslow v. Kroenert*, 29 Cal. 2d 541, 550 (1946)). Here, eBay has presented evidence sufficient to establish a strong likelihood of proving both prongs and ultimately prevailing on the merits of its trespass claim."

Arguing that the trespass is unlawful, eBay refers to its use of robot exclusion orders:

"eBay argues that BE's use was unauthorized and intentional. eBay is correct. BE does not dispute that it employed an automated computer program to connect with and search eBay's electronic database. BE admits that, because other auction aggregators were including eBay's auctions in their listing, it continued to 'crawl' eBay's web site even after eBay demanded BE terminate such activity.

BE argues that it cannot trespass eBay's web site because the site is publicly accessible. BE's argument is unconvincing. eBay's servers are private property, conditional access to which eBay grants the public. eBay does not generally permit the type of automated access made by BE. In fact, eBay explicitly notifies automated visitors that their access is not permitted. 'In general, California does recognize a trespass claim where the defendant exceeds the scope of the consent.' *Baugh v. CBS, Inc.*, 828 F.Supp. 745, 756 (N.D. Cal. 1993).

Even if BE's web crawlers were authorized to make individual queries of eBay's system, BE's web crawlers exceeded the scope of any such consent when they began acting like

<sup>26</sup> NO C-99-21200 RMW, eBay, Inc v Bidder's Edge Inc, 14 April 2000.

<sup>27</sup> The court's abbreviation for Bidder's Edge.

robots by making repeated queries. See *City of Amsterdam v. Daniel Goldreyer, Ltd.*, 882 F. Supp. 1273, 1281 (E.D.N.Y. 1995) ('One who uses a chattel with the consent of another is subject to liability in trespass for any harm to the chattel which is caused by or occurs in the course of any use exceeding the consent, even though such use is not a conversion.'). Moreover, eBay repeatedly and explicitly notified BE that its use of eBay's computer system was unauthorized. The entire reason BE directed its queries through proxy servers was to evade eBay's attempts to stop this unauthorized access. The court concludes that BE's activity is sufficiently outside of the scope of the use permitted by eBay that it is unauthorized for the purposes of establishing a trespass. See *Civic Western Corp. v. Zila Industries, Inc.*, 66 Cal. App. 3d 1, 17 (1977) ('It seems clear, however, that a trespass may occur if the party, entering pursuant to a limited consent, . . . proceeds to exceed those limits . . .') (discussing trespass to real property)."

This case is well known, but it is generally held that its reliance on the trespass doctrine is more inventive than convincing. One should also note that the case is based on the state law of California, which certainly in many respects will be different from the law in European jurisdiction. One should therefore not be too concerned with the detailed reference to precedents and doctrine, as to the general thrust of the argument: The robot exclusion orders were ignored, and this should be relevant for deciding whether the action is lawful.

There is one first instance decision from Norway, *Finn v Supersøk*, Trondheim first instance decision of 17 March 2006.<sup>28</sup> The facts were rather similar to the facts in *eBay v Bidder's Edge*. *Supersøk* was a metasevice which gave the user the possibility to compare offers from homes from four different estate agents, one of which was the major operator Finn. There were several questions in the case; one of them was whether a metasevice as such was unlawful. *Supersøk* had ignored a robot exclusion order which Finn had explicitly directed to its robot. Finn claimed that this was unlawful according to the general clause of the Norwegian marketing act sect 1, which makes an action unlawful if disloyal in the market, a rather broad assessment. The court held that a metasevice by itself was not disloyal, but that it had to be careful to represent its offer to the public. However, other elements in the case lead the court to find *Supersøk* unlawful. Personally I would argue that ignoring Finn's robot exclusion orders directed as *Supersøk* was a strong argument for finding the metasevice in such a case unlawful according to the provision of the marketing act.<sup>29</sup>

The conclusion for EIAO is less certain. The trespass doctrine will hardly apply directly in many jurisdictions, and the Norwegian marketing act does not apply unless the case is governed by Norwegian law. But both cases may be seen as examples of a standard qualifying disregard of a robot exclusion order as contrary to good practises or business standards. There would seem to be little co-ordination of law in this respect, it may be indicate to consider the difference between the Californian state law and the Norwegian marketing act. But diverse as the applicable law may be, there may nevertheless be some convergence upon a core standard which requires respect of robot exclusion orders.

Obviously, the reason for excluding robots may be diverse. It may be a perfectly understandable action taken to protect some business practise or values. But it may also be to avoid being measured by EIAO and to have the quality of the service compared with others. This may be something that EIAO feel should not be respected. But EIAO does not represent any public authority, and it is our recommendation that one should abide by the requests of the operators of the site. EIAO may find some method for encouraging sites to accept their ROBACC, for instance clearly publish which sites request to be excluded from the evaluation.

<sup>28</sup> Jfr TTRON-2004-85946, *Lov&Data* 86/2006:4.

<sup>29</sup> I appeared as an expert witness before the court on behalf of *Supersøk*.

## 6 Protection of map data

EIAO has the possibility to present accessibility data on maps down to NUTS level 3. However updated maps with NUTS regions listed are not available freely. We would have to negotiate terms and conditions to include recent map data in EIAO.

### 6.1 Introduction: Geographical information systems and maps

The basis for geographical data is positions. Positions may form structures – the curve of a road or the area of a lake. The structures are associated with properties, like “road” or “lake”, to which are added further properties – the road will have a number, a maximum axel load, *etc.* There may be properties which the result of statistical processing and political processes. Some properties are assigned based on observation and expert judgement, like the gradual transition of coniferous to deciduous forest.

The traditional representation of geographical data is a map. The property data is here associated with lists of co-ordinates defining the geometric positions. There has been a gradual development of the possibilities to represent geographical data. Today the data is stored in computerised form and not bound to a predefined way of presentation. A map may be seen as a form of report based on this database.

The scale of a map may vary from a very large scale used for maps of regions or nations, to small scale used for instance for development of an area. NUTS 0 to 3 are the units applied by the Nomenclature of Territorial Units for Statistics, and will often not refer to existing administrative units, at least for the more detailed units. They will (probably) be areas defined by lines of co-ordinates with associated properties.

In legal theory, the concept of a “map” is the conventional reference rather than geographical databases, though this is under development.

### 6.2 Legal protection of maps

Sometime in the past, maps were drawn by the hand of experts, and ancient maps may be beautiful, illuminated documents. Today, much of the data on which maps are based, is collected by computerised systems – for instance the geographical position system, aerial or satellite data, *etc.* While it is obvious that a map drawn in the old way was a type of technical drawing qualifying as a copyrighted work, it has been contested that modern maps meet the criterion of originality decisive for qualifying material as a copyrighted work.<sup>30</sup>

It is obvious that the geographical databases are protected by the *sui generis* database protection according to the database directive, cf above at sect 1.3: They are the result of substantial investment, and they organise a very large number of facts. This is sufficient obvious that a discussion can be excluded.

A database may at the same time be protected as a copyrighted work, cf database directive art 7(4). There are elements of geographical data which are based on expert judgement, and which may be argued meet the criterion of originality. And even when the database as such may not be subject to copyright protection, a map produced on this basis may require intervention of a creative nature which will make the result a copyrighted work. Indeed, the Norwegian copyright act gives “maps” as an example of copyrighted works in sect 1(1)(11), which is an indication of the possibility of maps meeting the criterion of originality.

Whether maps are copyrighted works, is contested, and the evaluation will also be related to the technical process for developing the maps, a process which gradually are becoming more automatic. In our context it would be futile to attempt resolving this issue on a general level. Also it is quite unnecessary. Though there are important differences between copyright and

<sup>30</sup> For Norwegian law, this is discussed in some detail by Steinar Taubøll *Rettigheter til geograifisk informasjon: Opphavsrett, databaser og avtalepraksis*, CompLex 2/2005, Norwegian Research Center for Computers and Law, Oslo 2005.

database protection, they both give the creator or maker an exclusive right to reproduce the material in whole or in substantial part.<sup>31</sup>

Maps are used extensively by public agencies. The case may often be that one agency holds a national responsibility for maintaining official maps. A municipality may purchase a map for its area, and use this to keep track of local infrastructure like pipes or cables in the ground, political decisions on development of the area for homes, business or recreational area, *etc.* The municipality in this way produces an “overlay” to the map, which may be fed back to the national agency for maintaining a detailed database. In this way, there may be more than one rightholder to the composite resulting database.

### **6.3 Geographical information systems – a commercial service**

The Norwegian Research Center for Computers and Law conducted approximately ten years ago a major study for EUROGI (European Umbrella Organisation for Geographic Information)<sup>32</sup> of several European countries and their policy with respect to the commercialisation of geographical information. In the conclusion, it was emphasised that geographical information is becoming increasingly valuable as it is being utilised in a widening circle of services and instruments. GPS has become commonplace for cars and mobile phones, office systems may have geographical interface to access different user groups, *etc.* In the country studies, there are numerous examples of the government establishing organisations for providing information services, which are different from conventional agencies or authorities. This is typical for the Trading Funds in United Kingdom, but also in other countries there is a tendency to create institutions under government control, but of a commercial nature (see for instance the Norwegian Land Information or the Land Registry in the Netherlands).

This development has several policy aspects. With respect to the freedom of information legislation, this will typically *not* apply to such organisations, which in this respect are not seen as public authorities. Also, the organisation enters into the market place, but will often have advantages due to its origin, background and operation which are different from other operators in the private sector. This obviously may have an impact on the competition for information services, giving these government related operators a lion's share of the market, and making market entry more difficult for operators that have to acquire the data (at least the raw data) from public sources.

There certainly is an issue here, as the government may see such a policy of developing self financed organisation as a strategy to ease public budget, and reduce the burden of the tax payer, making those requiring data also finance that it is made available. At the same time, this may reduce the possibility of growth in the private market for information service, which also is a policy priority within the European Union and most member countries.

### **6.4 Directive on the re-use of public sector information**

Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information is part of the effort to vitalise the sector of information services in Europe by making public sector information available for services in the private sector.

The re-use directive has as its objective to make data in the public sector available for re-use in the private sector for the development of information service in the innermarket. To achieve this, the directive gives some minimum rules to ensure “fair, proportionate and non-discriminatory conditions” for re-use.<sup>33</sup>

The directive does not directly introduce an *obligation* for the member states to permit re-use, the directive only applies when data is made available for re-use according to the national regime for access or by license, sale, distribution or exchange. To avoid cross-subsidies, it also

<sup>31</sup> This is discussed in another context above at sect 1.5.

<sup>32</sup> Cf <http://www.eurogi.org/>.

<sup>33</sup> Preamble item 8.

applies re-use within the public organisation if the re-use is outside public tasks.<sup>34</sup> Use outside what is necessary to fulfil the obligations as a public agency triggers the obligation to make the data available for re-use.

The directive permits that the data is made available for a price, but has provisions which have as their objective to make the price reasonable, and which encourages the member states to limit the price to the marginal costs.<sup>35</sup>

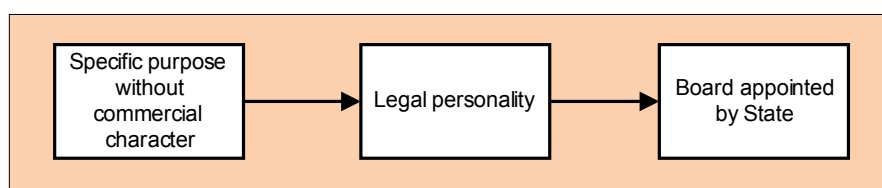
The core of the directive is art 4 "Requirements applicable to the processing of requests for re-use", and art 4(1) may be cited:

"Public sector bodies shall, through electronic means where possible and appropriate, process requests for re-use and shall make the documents available for re-use to the applicant or, if a licence is needed, finalise the license offer to the applicant within a reasonable time that is consistent with the time-frames laid down for the processing of requests for access to documents."

The obligation is limited to "public sector bodies". This is not directly defined in the re-use directive, but we presume it is to be interpreted in the same way as "body governed by public law", cf re-use directive art 2(2).

"body governed by public law' means any body:  
 (a) established for the specific purpose of meeting needs in the general interest, not having an industrial or commercial character; and  
 (b) having legal personality; and  
 (c) financed, for the most part by the State, or regional or local authorities, or other bodies governed by public law; or subject to management supervision by those bodies; or having an administrative, managerial or supervisory board, more than half of whose members are appointed by the State, regional or local authorities or by other bodies governed by public law; ...

There are three conjunctive criteria for a body to constitute a public sector body, summarised in this small diagram:



The definition is identical to that in the directives on public procurement.<sup>36</sup> This illustrates that what is qualified as a "public sector body" under the re-use directive may be somewhat different from what in common language is considered a public body. For instance, an organisation set up to furnish geographical information usually will have a specific purpose, a legal personality and often have a board appointed by the state. The decisive element may very well be whether the institution has a commercial character.

The application of the re-use directive probably has a limited interest for EIAO. But if there is a public sector body making available or licensing geographical information, it may be of interest that EIAO may require to have this information on equal terms.

<sup>34</sup> Preamble item 9.

<sup>35</sup> Preamble item 14 *in fine*.

<sup>36</sup> Cf directives 92/50/EEC(5), 93/36/EEC(6), 93/37/EEC(7) and 98/4/EC(8).

## 6.5 *Summing up*

The discussion in this section has briefly introduced the nature of geographical data and their possible legal protection by copyright or the *sui generis* database right. It has also briefly explained that such data increasingly are being made available by public sector through strategies for the commercialisation of public data, making access by freedom of information legislation (in those jurisdictions where this is relevant) less useful. It has introduced the directive on the re-use of public sector data, which in a very limited way may be relevant to EIAO.

The main conclusion is that geographical data will be protected by copyright or the *sui generis* database right. Gaining access to this data will typically presume the negotiation of a license from the agency controlling its commercialisation. This is in principle not difficult, but obviously comes at a price.

## 7 Liability for usage during crawls

*Would EIAO risk liability due to extensive resource usage during crawls?*

This issue has already been discussed above under sect 5.2.

We presume the situation is that the EIAO crawler accesses a site, and it is claimed that this access consumes sufficient resources for the site to experience this as a disadvantage, perhaps the result is reduced service quality, or even not being available. This argument may also be found in the decision discussed above under sect 5.2.3, *eBay v Bidder's Edge*:

“A trespasser is liable when the trespass diminishes the condition, quality or value of personal property. See *Compuserve, Inc. v. Cyber Promotions*, 962 F. Supp. 1015 (S.D. Ohio 1997). The quality or value of personal property maybe "diminished even though it is not physically damaged by defendant's conduct." *Id.* at 1022. The Restatement offers the following explanation for the harm requirement:

*The interest of a possessor of a chattel in its inviolability, unlike the similar interest of a possessor of land, is not given legal protection by an action for nominal damages for harmless intermeddlings with the chattel. In order that an actor who interferes with another's chattel may be liable, his conduct must affect some other and more important interest of the possessor. Therefore, one who intentionally intermeddles with another's chattel is subject to liability only if his intermeddling is harmful to the possessor's materially valuable interest in the physical condition, quality, or value of the chattel, or if the possessor is deprived of the use of the chattel for a substantial time, or some other legally protected interest of the possessor is affected. . . . Sufficient legal protection of the possessor's interest in the mere inviolability of his chattel is afforded by his privilege to use reasonable force to protect his possession against even harmless interference. Restatement (Second) of Torts § 218 cmt. e (1977).*

eBay is likely to be able to demonstrate that BE's activities have diminished the quality or value of eBay's computer systems. BE's activities consume at least a portion of plaintiff's bandwidth and server capacity. Although there is some dispute as to the percentage of queries on eBay's site for which BE is responsible, BE admits that it sends some 80,000 to 100,000 requests to plaintiff's computer systems per day. (Ritchey Decl. Ex. 3 at 391:11-12.) Although eBay does not claim that this consumption has led to any physical damage to eBay's computer system, nor does eBay provide any evidence to support the claim that it may have lost revenues or customers based on this use, eBay's claim is that BE's use is appropriating eBay's personal property by using valuable bandwidth and capacity, and necessarily compromising eBay's ability to use that capacity for its own purposes. See *CompuServe*, 962 F.Supp. at 1022 ('any value [plaintiff] realizes from its computer equipment is wholly derived from the extent to which that equipment can serve its subscriber base.').

BE argues that its searches represent a negligible load on plaintiff's computer systems, and do not rise to the level of impairment to the condition or value of eBay's computer system required to constitute a trespass. However, it is undisputed that eBay's server and its capacity are personal property, and that BE's searches use a portion of this property. Even if, as BE argues, its searches use only a small amount of eBay's computer system capacity, BE has nonetheless deprived eBay of the ability to use that portion of its personal property for its own purposes. The law recognizes no such right to use another's personal property. Accordingly, BE's actions appear to have caused injury to eBay and appear likely to continue to cause injury to eBay. If the court were to hold otherwise, it would likely

encourage other auction aggregators to crawl the eBay site, potentially to the point of denying effective access to eBay's customers. If preliminary injunctive relief were denied, and other aggregators began to crawl the eBay site, there appears to be little doubt that the load on eBay's computer system would qualify as a substantial impairment of condition or value. California law does not require eBay to wait for such a disaster before applying to this court for relief. The court concludes that eBay has made a strong showing that it is likely to prevail on the merits of its trespass claim, and that there is at least a possibility that it will suffer irreparable harm if preliminary injunctive relief is not granted. eBay is therefore entitled to preliminary injunctive relief.”

The decision is partly based on the assumption that “BE's activities consume at least a portion of plaintiff's bandwidth and server capacity.” It can hardly be contested that a “portion” is consumed, but to which extent this represents a problem for eBay, certainly is contested. On this point, the decision has been criticised, it being maintain that the extra load of the 100,000 extra requests per day was marginal.<sup>37</sup>

If the EIAO's crawler had a noticeable effect on resources at the web site, this could constitute an economic loss, and in principle the question of civil liability could be explored. Liability has to have some legal fundament, and it would seem to require that the activity of the EIAO could be construed as unlawful or negligent. A basis for this would be that the crawler disregarded a robot exclusion order, see the discussion above under 5.2.3. This is, as far as I can see, the only possibility – though there may be aspects of this not grasped by the author.

Apart from this possible exposure to liability, there would seem to be very low legal risks in launching the EIAO crawlers.

---

<sup>37</sup> This in conversation with the author, who have no citations for backing this opinion.

Norwegian Research Center for Computers and Law  
Department of private law  
Faculty of Law  
University of Oslo

Jon Bing  
08.09.08



Oslo 08.09.2008  
Version 2.1

## Copyright aspects of search engines

### 1 Introduction

This paper offers a discussion on certain aspects of copyright law related to the right of reproduction. It identifies one of several situations<sup>38</sup> in which conventional use of information technology requires reproduction, but where the author cannot find any authorisation for this in current Norwegian (or European) copyright law. This is not an argument for the case to be *an infringements*, without them, the use of Internet (or more to the point, World Wide Web) would be cumbersome and less efficient. The basic functionality of the service should be provided for, and the cases should not be allowed to bar this. But then a legal policy is called for. If the analysis is correct, the legal policy currently adopted is to look another way. That is not sufficient, one should scrutinise the example, point out the possible failures in the argument or issues on which there may be different opinions. If one in this way cannot be dismiss the case, one should address the it through a policy of legal reform.

### 2 Search engines

#### 2.1 A brief history of search engines

In the spring of 1995, the research laboratory of Digital Equipment Corporation in Palo Alto, California introduced a new system based on the Alpha chip which was able to operate databases much faster than competing systems. To demonstrate the technology, they decided to index the Web, which at this time was less than five years old.

The idea is superficially simple, and one well known to lawyers, as the method is basic for all legal information services using text retrieval. Traditional retrieval systems rely on an index of intellectually assigned terms – an indexer would consider a page, and decide which terms best characterised its content, often selecting terms from a pre-defined vocabulary. This is the

<sup>38</sup> Other situations are proxy servers, using a web browser, hyperlinking, and backup procedures.

way in which documents were indexed according to for instance Dewey's classification system<sup>39</sup> used by libraries around the world.

Computers made possible an alternative approach. Each word of the text would be sorted alphabetically, retaining a reference ('an address') to the location within the text. A user would then specify a search term, this would be matched to the index, and the pages containing those words could be displayed. Or rather, the preferred format would be KWIC – 'keyword in context' – showing the occurrences of the search term with its adjacent text, this made relevance assessment easier: Did the word occur in a context relevant for the problem of the user, if so the address made access to the source readily available. More sophisticated search strategies were developed, typically using Boolean logical operators – find a text in which both the word 'Digital' and 'Equipment' occurs, or augmented to a requirement for these words to occur in the same sentence or adjacent. Ranking of retrieved documents according to different principles were added.<sup>40</sup>

The first such system was successfully demonstrated by Professor John F Harty of the Health Law School, University of Pennsylvania at an American Bar Association conference in 1960. It is not by chance that lawyers were the first to use such systems professionally, the intimate relationship between exact wording of a statute or regulation and interpretation for use in legal argument, required access to the original, authentic text – solutions as abstract journals like 'Chemical Abstracts' were not viable. Though lawyers are not well known as the technology *avant garde*, they somewhat reluctantly pioneered text retrieval; major systems like Reid-Elsevier LEXIS-NEXIS are examples of what has come out of this development.<sup>41</sup>

A team led by Louis Monier at DEC's Western Research Laboratory, developed the search engine. By August 1995, this conducted its first full scale crawl of the Web. Using the domain name system, the crawler visited websites, it might utilise the HTML-coding to identify interesting bits of a page if not copying the site in total, and communicating the copies back to the home site for further processing. In its first trial, it brought back some ten million web pages.

The new service was called AltaVista – 'the view from above'. It became available to the public 15 December 1995 with an index for 16 million documents. It was an immediate success, with more than 300,000 searches the first day. At the end of 1996, it was handling 19 million requests daily.<sup>42</sup>

The miraculous aspect of AltaVista and other search engines is not the search logic. Compared to the sophistication of professional text retrieval systems, the search 'language' is not very advanced. The miracle is the vast number of websites indexed, and the maintenance of this index. AltaVista was sold off when Compaq acquired Digital Equipment Corporation, and has been further developed.<sup>43</sup>

Several other major search engines have been launched, like Yahoo!, MSN, Lycos *etc* – not all of them only using text retrieval methods, but augmented by other methods, like intellectual indexing.

Google was founded by [Larry Page](#) and [Sergey Brin](#), both graduates from Stanford, in 1998, literary out of a garage in Menlo Park, California. The story of this company is another tale of innovation and intuition. The first search engine they built was called BackRub, named for its

---

<sup>39</sup> Based on *A Classification and Subject Index for Cataloguing and Arranging the Books and Pamphlets of a Library*, first published 1876, the system is today maintained by the non-profit organisation Online Computer Library Center (OCLC).

<sup>40</sup> The history and theory of text retrieval is discussed in some detail by Jon Bing *Handbook of Legal Information Retrieval*, North-Holland, Amsterdam 1984.

<sup>41</sup> It must be permitted to indicate that the Norwegian Lovdata legal information service also is an interesting example where several innovating strategies for text retrieval has been implemented.

<sup>42</sup> Cf 'AltaVista: A brief history of the AltaVista search engine', [http://www.websearchworkshop.co.uk/altavista\\_history.php](http://www.websearchworkshop.co.uk/altavista_history.php) [7 Aug 2006].

<sup>43</sup> One of the more interesting developments is the Babel Fish, the first automatic translation service on the Internet.

ability to analyse the ‘back links’ pointing to a given website through hyperlinks.<sup>44</sup> Use of citation frequency as a ranking criterion was well known for retrieval purposes,<sup>45</sup> but the integration of such links in the Web made them different from the formal references to literature *etc* in conventional texts. The ranked results improved performance considerably, increasing the probability of the first document presented to the user being relevant. Today, Google is by far the most popular search engine, with numerous additional and often innovating services supplementing the basic search function. Its popularity is reflected in the tradename having graduated to an accepted, English language word – to google.<sup>46</sup>

## 2.2 Basic function of a search engine

The core of any search engine is the index. This has already been briefly introduced above; it is an alphabetically sorted list of the words occurring in a document or collection of documents. It presumes that the indexing program is able to identify a “document” and the “words” of the document. What is a “word” is not quite trivial, any string of characters may be qualified as a word, but will be stripped of some initial or terminal characters, like commas, parentheses *etc*. Also, the indexing program may be designed to give a somewhat more precise “address” to the words that just its association to a specified documents – the program may identify the paragraph and the sentence in which the word occurs, as well as the position within the sentence.

In order to index a document, the indexing program must have the document available in the memory of the central processing unit controlled by the program. This requires the material to be indexed to be communicated to this computer. To load the search engine, it must start by identifying the material to be indexed. This is achieved using the URLs. Some search engines allow an operator of a site to “register” their site for indexing, a request will then be made to the site, material will be copied to the site of the search machine, and indexing will take place. When identified a site,<sup>47</sup> the search engine will identify links on this site to other sites, these also will be followed, *etc*. In this way, the search engine will unravel the net, using the hyperlinks to include a growing fragment of the sites available.

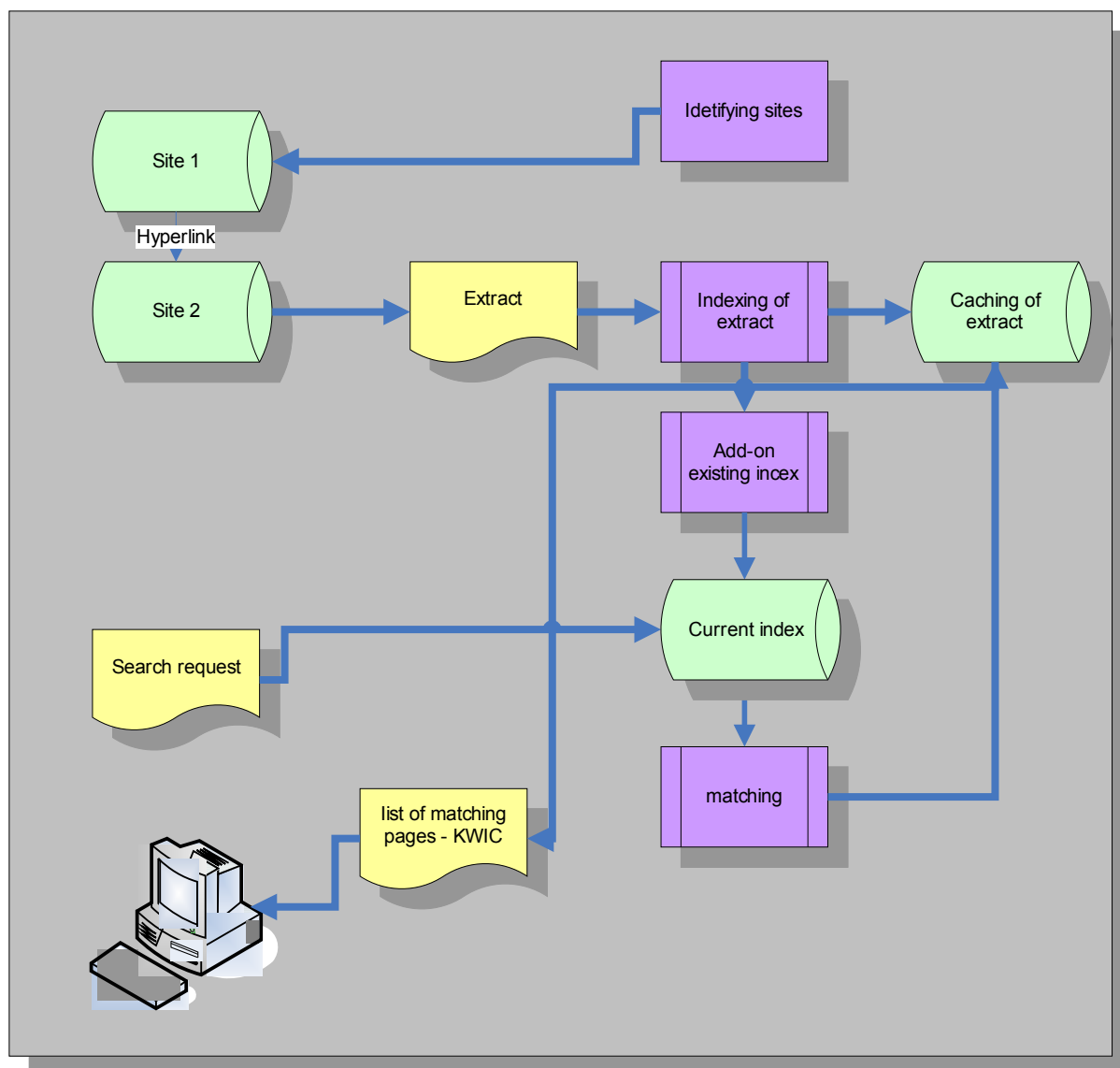
---

<sup>44</sup> Cf <http://www.google.com/corporate/history.html> [7 Aug 2006].

<sup>45</sup> Cf Eugene Garfield *Citation Indexing* (1979).

<sup>46</sup> Included in the 11th edition of the Merriam-Webster Collegiate Dictionary (2006).

<sup>47</sup> The owner of a site may ask a search engine to add its URL to the sites being indexed, cf <http://www.google.com/addurl/?continue=addurl> for an example for Google.



**Figure 1 - Simplified functions of a search engine**

The number of sites accessed and indexed in this way, is staggering, and it is a cause of wonder how the big search machines are able to access and index billions of documents. In spite of this, it is being maintained that the Web expands sufficiently fast for new links to be identified each new site are accessed, therefore only a fraction of the total material of the Web has actually been indexed.

The index for a new site is integrated in the general, current index which is the basis for the service offered to users. And when the indexing has been completed, the copied site is not deleted, but typically stored by the searchindex in auxiliary files often referred to as the “cache” of the search engines. One should note that these copies represent the indexed site at the time of indexation. They will not necessarily be replaced when the original site is being updated, the cache will only be updated when the original site is being re-indexed. The updating frequency of the original site may be hours or days; the frequency of re-indexing will depend on other considerations, for instance the popularity of the original site.

A user will access the search engine using a search request, which will be a single search term or a Boolean request. The search engine will match the words of the request with the words in the index, and finding a match;<sup>48</sup> will present the “hits”. The sequence will be sorted or

<sup>48</sup> There will be rules to allow for less than an perfect match, for instance compensating for plurals and inflections, perhaps allowing right hand truncation *etc.*

ranked, above we have briefly mentioned the citation ranking of Google, there will be more simple versions based on the frequency of search terms, and there will be combinations of different kinds.

The presentation typically has the traditional KWIC format, presenting the search term in the context of the text from which it has been indexed, this snippet also greatly assisting the relevance judgement of the user. The snippet is obviously not from the current text of the site, but from the site as indexed and stored in the cache of the search engine.

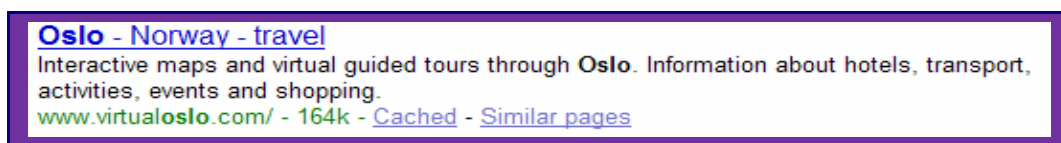


Figure 2 - KWIC extract for hit by a search engine on the term “Oslo”

The example of a snippet using the search term “Oslo”, illustrate the response from the search engine. First (in green) is the URL of the indexed site, and then there is a link to the version cached by the search engine.<sup>49</sup> The user may also click the “heading” (in which the search term occurs), which also is a link to the original site.

The user may now decide whether to access the material in the cache of the search engine, or the original site. In the latter case, there is a possibility for this site to have been updated after having been indexed, revealing a discrepancy to the snippet. And also the original site may be discontinued, in which case the web browser of the user will bring an error message.

In this process, and without taking the end user into consideration, there are at least three issues related to copyright law which should be considered.

## 2.3 Issues related to copyright law

### 2.3.1 The copying of sites

As mentioned, for indexing purposes, the site to be indexed has to be communicated to the search engine. The indexing presumes processing of the text and this obviously has to take place in a central processing unit controlled by the program. The material to be indexed, will be stored in such a way that it may be paged into the primary memory of the processor when required by the program. The program will basically sort the text occurring in alphabetical order, but may also do more: It may interpret the tags of the html-coding and use them for rendering data more accessible – for instance, if a term occurs in a field qualified as “heading”, it may be assigned relatively higher weight as probably more representative for the content than words in the body of the text.

From the early history of search engines, there are many anecdotes on how to “fool” the engines. For instance a word could be repeated many times in a non-printing field, the search engine would rank different documents on the basis of word frequency, the document having the highest frequency climbing to the top of the list. Or it could be the trade mark or name of a competitor which was stated in a non-printing field – using the name of your favourite brand as a search term, mysteriously would result in the site of the competitor appearing on top of the list. A favourite is the solicitor whose web page had a sober, black background – or so you thought, until you enlarged the background and realised that it was composed on thousand and thousands of repetitions of the word “solicitor” printed with very small font.

<sup>49</sup> There is also the option “similar pages”, which will generate a new search request based on the snippet.

The search engine also will have to interpret composed pages, identify images *etc.*

The point in this respect is that in order to do the indexing, the original site will have to be reproduced. This reproduction does not have to last longer than the process of indexing (but will, as already mentioned, in practice be retained). This reproduction is not transient or incidental,<sup>50</sup> and is therefore part of the exclusive right of the rightholder. Therefore, the operator of the search engine has to find authority for this reproduction either in a statutory license or a limitation of the exclusive right, or in a license from the rightholder. Again we will find that there is not provided for statutory licenses or limitations of the exclusive rights in European law. There will only in exceptional cases have been communication with the operator of the indexed site (who also may be the rightholder to the indexed material), one such exception is mentioned above when the operator of a new site register the site with the search engine with the objective to have it indexed. Such a registration must also be seen as a *license* to make the reproductions necessary for indexing, indeed for all the reproductions customary by that search engine (also the reproductions cached).

For the large majority of instances, there will not have been any contact between the operators of the search engine and the indexed sites. But in a typical case, the operator of the site will see the indexing as beneficiary through the search engine users may identify the site without knowing the URL in advance. It certainly will not be an exaggeration to maintain that the search engines are *presumptions* for successful services. It will therefore be easy to argue that the operator of the sites being indexed have implicitly given their license when making their site available on the web.

However, this is not always the case. One may take a web service which makes available veterinary advice on the web – a user is invited to specify an animal, symptoms or whatever, and retrieve advice. If a search engine indexed this whole site, the user could restrict the search to the engine – it would display the desired data and become a competitor to the web service.

To avoid this, solutions have been developed.<sup>51</sup> One of the most common is a *robots.txt* file.<sup>52</sup> This is mainly designed for search engines, which is a simple type of electronic agents. The program constituting the search engine will look into the root domain for a file named "robots.txt". The first part specifies the robot, while the second part consists of directive lines, disallowing the robot to index files or directories. A simple example would be

```
User-agent: macrobot  
Disallow: cat.htm
```

This is addressed to the robot "macrobot", and direct this not to index the file "cat.htm".

The robots.txt solution is a "polite" strategy. Whether the "robot" (the program of the search engine accessing the site to reproduce material) follows the advice, will depend on the program. But a loyal search engine would comply.<sup>53</sup>

In our context, the question is the consequence for the implied consent. We argued that the operator of a web site typically has given his or her implied license by making the site available on the Web. But in the case a robots.txt is included; this explicitly excludes defined elements of the site. It is explicit, but written in formalism different from natural language, directed at a

<sup>50</sup> Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society art 5(1).

<sup>51</sup> For Google, <http://www.google.com/support/webmasters/bin/answer.py?answer=35301&topic=8459> gives advice in how to avoid indexing (for instance using the tools briefly indicated below), remove cached copies *etc.*

<sup>52</sup> The scheme was originally launched in 1994. For more information, see <http://www.robotstxt.org/> [12 March 2007].

<sup>53</sup> There will be supervisory programs of the site noting if a robot does not comply, and this may initiate further action, like barring any request originating from the site of the rogue robot. Such more technical consequences will not be pursued here.

computer program rather than a human being. However, it is appropriate to do so, for the action the operator of the site want to exclude, is that of a computer program, not a human being.

If the robots.txt is seen as an explicit withdrawal of what otherwise would be an implicit license, we have to consider the consequences in copyright terms. If the robot program does not follow the polite request in the robots.txt, but goes ahead and copies the files, communicating them to the site of the search engine where they are reproduced for indexing purposes – is this to be construed as an illegal reproduction under copyright law?

This also will rely on whether formal statements addressed to computer programs are to be considered to have a legal binding effect. It is obvious that the operator of the search engine has not read the statements, and does not have any actual knowledge of the precise site using the statement. But the operator will have (or ought to have) knowledge of the way such statements are used to govern indexing on the Web. If the operator does not make sure that the program he or she deploy checks for such statements, is able to correctly interpret the statements and act on the statements according to the generally accepted semantics, then it may be argued that the operator does not act in good faith. It may be argued that not only may the operator then be liable for damages, but also held criminally liable for copyright infringement.

### 2.3.2 *The indexing*

The index is an alphabetic list of all the words occurring in the document. Its level of detail depends on the analysis of the document, and what data are included on the location of each word within the document.

If the index only include a document number, one may try to reconstruct any document by collecting all the words which in the index are specified as part of that document. The result will be a jumble of words without any internal sequence (apart from the alphabetical listing), and no indication of how frequently the word occurs in the document. It would be difficult to see this set of words as a “reproduction” of the original document.

The index may also contain a more detailed location of each word, specifying paragraph within the document, sentence within the paragraph and word number within the sentence. Using this data, a document may be fully reconstructed – and this reconstruction is identical to the original document, and would therefore be a reproduction of that document. For this reason, the index is also known as the “inverted file”.

To reconstruct the document, it is necessary also for all the words to be indexed. It is quite usual to exclude some common words like conjunctions, modal verbs, articles *etc* – these are words necessary to make the text flow, but often characterised as vehicular; taken by themselves and considered separately they will not convey any meaning. For this reason, and some additional considerations, one may exclude such words from the index. The reconstruction of the original document would therefore be incomplete.

Page size, Kb	MSN			Google			Yahoo!		
	MSN bot downloaded, Kb	MSN Indexed, Kb	% of text	Google bot downloaded, Kb	Google Indexed, Kb	% of text	Yahoo bot downloaded, Kb	Yahoo Indexed, Kb	% of text
45	44	40	100	44	40	100	44	40	100
68	67	60	100	67	60	100	67	60	100
90	90	90	100	90	90	100	90	90	100
98	96	90	100	96	90	100	97	90	100
109	107	100	100	107	100	100	107	100	100
118	115	100	100	115	100	100	115	110	100
129	126	120	100	126	120	100	125	120	100
137	133	130	100	133	130	100	136	130	100
147	145	140	100	145	140	100	145	140	100
174	171	170	100	171	170	100	173	170	100
221	218	210	100	220	220	100	220	210	95
262	258	250	100	260	260	100	218	210	80
366	359	350	100	360	360	100	224	210	57
401	391	390	100	400	400	100	220	210	52
547	535	530	100	535	520	95	211	210	38
788	773	770	100	633	520	66	233	210	27
1042	1025	1020	98	638	520	50	229	210	20
1228	1126	1030	84	610	520	42	222	210	17
1506	1125	1030	68	657	520	35	221	210	14
1990	1105	1030	52	617	520	26	240	210	11
2213	1107	1030	47	630	520	23	228	210	9
2644	1123	1030	39	633	520	20	213	210	8
3187	1117	1030	32	617	520	16	229	210	7
3497	1132	1030	29	624	520	15	245	210	6

Figure 4 - Indexing depth for three major search engines

In addition, not the whole original site is indexed. Serge Bondar<sup>54</sup> has through an experiment tried to determine how much of a site the three major search engines index, the results are given in the table below.

From this we see that MSN downloads 770 Kb,<sup>55</sup> while Yahoo! only downloads 210 Kb. The index for a voluminous site will therefore not be complete, but sufficiently will be indexed for the representation to constitute an infringement of the original work *if* one accepts the argument indicated above.

The argument is based on the suggestion that the address of the indexing terms has a sufficient detailed address to permit the index to be inverted and reconstruct the indexed documents. This is admittedly a rather theoretical argument, but indicates that also the index may be a reproduction of the work, and therefore has to find a legal basis for the reproduction it represents. Again, we will have to turn to implied consent for constructing such a basis. The argument will be parallel to that for the reproduction necessary to permit indexing – the index is the result of the processing of the data from the original sites, and is also the core of the service which the search engine offers the end users. The interest of the operators of the original sites is in enabling this service; therefore the implied license includes the index itself to the same extent as it permits the processing necessary to construct the license.

### 2.3.3 The cached documents

As mentioned in presenting how the search engines work, the search engine will store the original site (what has been downloaded from that site) in a cache. The cache is available for the end user.

This storage is not necessary for the basic indexing. When the index is established, the search engine could in principle discard the copy of the original sites, only displaying the URL

<sup>54</sup> "Search Engine Indexing Limits: Where Do the Bots Stop?", 28 April 2005.

<sup>55</sup> One byte corresponds roughly to one alphanumeric character, 770 Kb corresponds then roughly to 385 "normal" pages, a normal page being defined as 2,000 characters by customary standards in publishing. It would be sufficient for a normal seized novel.

for the user searching for material using key-words. But there are several reasons for the search engine to retain the original sites.

First, it is the KWIC format displayed to the user. It is presumed<sup>56</sup> that this context is retrieved from the stored material at the time the list of “hits” is constructed for the end user. Using the address of the indexed word, the system dips into the stored material and retrieves a pre-defined string of text which includes a certain number of words preceding and following the search term. As mentioned, this is very useful for the end user. Determination of the possible relevance of the source to the problem of the user may be done with high performance based on such snippets. Excluding the context would leave the user uncertain, and the functional performance of the search engine would clearly be impaired.<sup>57</sup>

There may also be other reasons. The maintenance of the vast system of the search engine presume a continuous refinement and re-tuning of various parameters, for instance for ranking retrieved documents. In this, data from the cached sources are useful.<sup>58</sup>

It may therefore be argued that the implied license also should extend to the caching of the downloaded, original sites. An appropriate function for the assessment of relevance is also in the interest of the operators of the site indexed, and the argument above could be repeated. The tuning of the system would seem to be just an additional facet of the establishment of the index and its related functionality, continuing after the site has been indexed for the first time.

However, the cached material is also available for the end user, as an *alternative* to the original site. And though the interests of the operators of the original sites and the operator of the search engine may coincide with respect to relevance assessment and tuning of the system, they *do not* coincide in this respect. The original operators would prefer the end user to access the original site. This becomes especially important where the site is updated after having been visited by the search engine. Then the original site and the cached reproduction will deviate. For instance, the site of a news service will not necessarily contain the latest updates, and the news service may argue that this has a negative impact on its reputation.

While the other aspects of tension between the search engine operations and copyright has not realised themselves as court cases, one will find cases on this issue.

The Court of First Instance of Brussels<sup>59</sup> in a decision of 13 February 2007 between Copiepresse (which manages copyrights for Belgium's French and German-language newspapers) against Google that the reproduction in the cache infringed both copyright and the *sui generis* database right.

There have been statements from other representatives of the news media arguing that the practise of search engines like Google is unlawful. As one may have expected, the representatives of search engines have argued that news services can have their sites excluded from indexing by a robots.txt or other formalism. But, as we see above, it is in the operators of the news services to have their sites indexed, but not to have the cached reproduction available for the end user.

## 2.4 Snippets and thumbnails<sup>60</sup>

The database directive also establishes a *sui generis* protection of databases, provided that the database meets the criteria for protection set out in the directive (art 7(1)). In this brief discussion it is presumed that the database is protected, a presumption which may not be trivial.<sup>61</sup> In indexing the database, a substantial part of the database is reproduced. It can hardly be otherwise, even if all the items are not indexed in full depth, it will be sufficient to establish

<sup>56</sup> This has not been confirmed by literature or correspondence with the operators of any of the search engines.

<sup>57</sup> This is based on both theoretical and empirical studies, for an overview, see Jon Bing *Handbook of Legal Information Retrieval, North-Holland, Amsterdam 1984:93-97.*

<sup>58</sup> This is based on information made available from representatives of one search engine.

<sup>59</sup> Case 06/10.928/C.

<sup>60</sup> Jeg er veldig lite fornøyd med de to siste punktene, bade systematisk og innholdsmessig. Vi må se nærmere på hvorvidt det skal med, og eventuelt under hvilket perspektiv det skal diskuteres.

that a reproduction of a substantial part of the database has been made. The argument would be parallel to the argument that the indexing a copyrighted database presumes a reproduction.

In addition, the repeated and systematic extraction or re-utilisation of insubstantial parts of a database which conflicts with the normal exploitation of the database, or unreasonably prejudices the legitimate interests of the maker of the database is not permitted (art 7(5)). A web site will typically be organised as a database (the term is defined in the directive art 1(2)). For indexing purposes, the search engine will in a systematic way extract material from these databases. Typically, it will not be limited to insubstantial parts; therefore the auxiliary protection in the directive art 7(5) may be of less practical importance than the primary protection of extraction or re-utilisation of the whole or substantial parts (directive art 7(1)), see above.

But it may be argued that the snippet of text in which a search term is embedded using the KWIC format, is an insubstantial part of the database. It certainly represents re-utilisation of these parts. To be unlawful as such<sup>62</sup> this use also has either to conflict with the normal exploitation of the database, or unreasonably prejudice the legitimate interests of the rightholder.

The use of the snippets does not seem to conflict with the normal exploitation of the database. Users accessing the database will not experience any problems of limited availability due to the snippets offered by the search engine. It may easily be turned the other way around, the data base *rely* on the search engine for its normal exploitation, the search engines provide a retrieval function which enhances the typical services offered on the web. The legitimate interests of the rightholder may take several forms. It may be an interest for remuneration, but would more typically be an interest in maximising the number of hits at the website – this may relate to the income from banner advertisements offered on the website, or more generally to promoting the operator of the website and the associated services, and strengthening the reputation of the operator. The snippets themselves are no replacement for the website itself, and it is difficult to see that in a typical example, the legitimate interests of the rightholder are prejudiced by this practice.

Therefore, it may be argued that viewed in isolation, the snippets do not infringe the *sui generis* database right. Obviously, one cannot argue this issue in isolation, and as has been demonstrated above, there are other aspects of a search engine that has to be justified with respect to copyright law.

The use of snippets and thumbnails may be directly related to the functions of a search engine as an information system:

As any information system, it must support three functions to be operative.

- (1) It must support a *search* function, which for search engines is realised by an automatic indexing based on an extract of the authentic text of the material included in the data base.
- (2) It must have a *source* function; this means that the system must give the user access to the material in a form which is appropriate for exploitation by the user. For instance, it is of little or no interest to a user to find that there are a number of documents satisfying his or her search request without there being any way of accessing the documents. This would be like finding an index card in a library for the book desired, only to discover that the shelf is empty and the book lost. This is a performance failure, and no information system will survive without generally satisfying the source function. For the search engine, the source function is satisfied in two ways: (i) the user is given a link to the original site, and can look it up, and (ii)

---

<sup>61</sup> When a database is established as a secondary service on the basis of a primary paper publication, it may not be obvious that the database as such meets the criterion of substantial investment in the directive.

<sup>62</sup> Disregarding that the process which make it possible to present such snippets itself has to be justified.

the user is offered access to the “cached” document copied for the purpose of indexing.

- (3) The *relevance* function – an information system will have a way of assisting the user to make a judgement of the relevance of a document before accessing it. This is typically achieved with an abstract or the title of the document. In search engines it is solved by displaying the search term in context, the context having been pulled out from the document copied for the purpose of indexing.

It may be argued that if one is to allow search engines, one must allow them to have a satisfactory functional performance as information systems. Therefore, *if* one allows a document to be processed for the retrieval by a search engine, *then* one must allow the use necessary to realise the retrieval, source and relevance function. Otherwise one will impair the functional performance and argue that the search engines should be permitted, but not permitted to work as information systems – which certainly will be to the disadvantage of users, and which can be proved by the massive literature on the performance of information systems.

Therefore it would seem that much is resting on the *if* which allows a document to be processed by a search engine. There are ways for the rightholder to specify that this is not permitted, examples have been given for robot.txt and metatags in HTML. A case can be made for strengthening these possibilities, especially giving them standard formats, and require that a specification for a prohibition to process a document uploaded to the Internet should have a format corresponding to industry standards, and permitting itself to be read and interpreted by an agent of a search engine.

In the absence of such explicit directives, we argue that there is an implicit license for a search engine to process the document. This permission would then cover the use of the document which otherwise would require the permission of the rightholder, and which is necessary for the functional performance of the search engine as an information system. This would include:

- (1) Copying of the document, or an extract of the document, for indexing purposes, the indexing process and the index itself. This would enable the search function.
- (2) Linking to the site from which the document is copied (original or rightholder’s site). This would solve the source function. Offering the document from a file maintained by the search engine, containing the document in the form used for indexing, is not necessary for the source function. If the original is not available, this will have a reason, typically that the site has been updated or changed by the rightholder. There may be desirable to work more sophistication into the use of metatags or other formalism to govern this aspect, but it should be maintained that the basic *implicit licence* cannot be interpreted to cover making accessible a copy of the document originally made for indexing.
- (3) Making available a snippet of the document to provide a keyword in context index. This is necessary for the user to make a relevance assessment, without which the system will be functionally impaired. The snippet of text itself is hardly a copy of a literary work, but it is extracted from the copy made for indexing, re-using this copy to provide the context. This should be permitted, and therefore also the storage of the copies by the search engine should be permitted to support the relevance function. While the continued storage of the copies cannot be permitted on the basis of supplying them to the users as alternatives for the original websites to satisfy the source function, they may be stored for the limited purpose of supporting the relevance function.

The example is based on literary works, ie textual documents. A parallel argument could be made for images.

However, images can only be made retrievable by associating them with some textual element (excluding advanced methods for pattern recognition which currently hardly are in use). Such texts could be captions associated with the image by the author, but there are other methods for determining that terms in a text associated to the image also “describes” the image. Obviously these are methods somewhat uncertain, a uncertainty anyone having used a standard search engine for retrieving images will have experienced.

The retrieval function clearly has a lower functional performance compared to textual documents. This will cause the retrieval to include a relatively lower ratio of relevant images (lower precision). This makes the user more dependent upon an efficient relevance function for discarding irrelevant matches. Thumbnails provide an excellent solution – small images with a low resolution, sufficient for the eye to judge its possible relevance, but insufficient for reproducing the image on paper or screen. But in difference to a snippet of text, a thumbnail represents the whole visual work, though the quality may be low. There is therefore no doubt that the thumbnail is a reproduction of the work, and that listing thumbnails will represent both reproduction (in the CPU of the work station of the user) and making the work available to the public.

This difference does not have a direct bearing on the argument that the rightholder has implicitly permitted search engines to make the necessary use of the uploaded material to function appropriately. The argument will be that the implicit licence in the case of images includes the display of thumbnails for relevance assessment.

But as the whole image is displayed, the situation will in practice be somewhat different. The situation may more frequently contain elements which will make the implicit licence fail – the pragmatics are such that it cannot be argued that uploading implies consent by the rightholder for displaying thumbnails. An example may be a 2006 Californian case<sup>63</sup> where the court held that the defendant had “established a likelihood of proving that Google’s creation and public display of thumbnails”<sup>64</sup> directly infringed the copyright of the defendant. The case had other elements which are relevant in the decision of the court.<sup>65</sup> But it is of interest that the court did not discuss the function performance of the search engine, and the role of thumbnails for relevance assessment in this perspective. It would seem that such a perspective would not be irrelevant for a discussion of the application of the fair use doctrine.

In this way, the ambit of the implicit license may be based on the well explored functions of an information system. If implicit license permits a search engine to make the document available to the users, then the implicit license is interpreted to permit such use which make it possible for the search engine to operate with satisfactory performance.

### 3 The *sui generis* data base protection

Referring to the Belgian decision above, reference is made to the *sui generis* database protection under European law.<sup>66</sup>

Databases may be protected compilations under copyright law,<sup>67</sup> and it is generally accepted that an edition of a newspaper is a compilation by the editor, who will be the original copyright holder of the compilation as such, in practice the rights of the editor will be transferred to the owner of the newspaper by the employment contract. There is no need in this context to dwell

<sup>63</sup> United States District Court, Central District Court of California, Perfect 10 v Google, Inc *et al* (CV 04-9484 AHM (AHx)).

<sup>64</sup> United States District Court, Central District Court of California, Perfect 10 v Google, Inc *et al* (CV 04-9484 AHM (AHx)) page 33.

<sup>65</sup> For instance that Perfect 10 marketed scaled down images for mobile telephones, the thumbnails directly competing with this service.

<sup>66</sup> Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases Chapter III.

<sup>67</sup> Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases Chapter II.

on this aspect, it may be sufficient in passing to note that not only is the search engine infringing the rights of reproduction of the authors of the copyrighted items being indexed, in some cases the infringement will also be of the compilation of items.

In many of the services made available on the Web, the items are not by themselves copyrighted works. Also, taken as a whole, the site does not qualify as a copyrighted work. Typical examples would be telephone directories and other similar collections of fact – useful, but not subject to copyright.