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 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 infringement, 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 way in which documents were indexed according to for instance Dewey’s classification system used by libraries around the world.

1 Other situations are proxy servers, using a web browser, hyperlinking, and backup procedures.
2 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).
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.  

The first such system was successfully demonstrated by Professor John F Hority 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.

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.  

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.

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 ability to analyse the ‘back links’ pointing to a given website through hyperlinks. Use of citation frequency as a ranking criterion was well known for retrieval purposes, but the integration of

---

3 The history and theory of text retrieval is discussed in some detail by Jon Bing _Handbook of Legal Information Retrieval_, North-Holland, Amsterdam 1984.
4 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.
6 One of the more interesting developments is the Babel Fish, the first automatic translation service on the Internet.
7 Cf [http://www.google.com/corporate/history.html](http://www.google.com/corporate/history.html) [7 Aug 2006].
8 Cf Eugene Garfield _Citation Indexing_ (1979).
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 trade name having graduated to an accepted, English language word – to google.

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, 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.

[10] 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.
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 search index 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;\textsuperscript{11} will present the “hits”. The sequence will be sorted or 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.

\textbf{Oslo - Norway - travel}
Interactive maps and virtual guided tours through Oslo. Information about hotels, transport, activities, events and shopping.
www.virtualoslo.com/ - 164k - Cached - Similar pages

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.\textsuperscript{12} 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.

\section*{2.3 Issues related to copyright law}

\subsection*{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 is 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

\textsuperscript{11} There will be rules to allow for less that an perfect match, for instance compensating for plurals and inflections, perhaps allowing right hand truncation \textit{etc.}

\textsuperscript{12} There is also the option “similar pages”, which will generate a new search request based on the snippet.
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.

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, 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. One of the most common is a robots.txt file. 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 “macrorobot”, 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.

---

14 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.
15 The scheme was originally launched in 1994. For more information, see http://www.robots.txt.org/ [12 March 2007].
16 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 rouge robot. Such more technical consequences will not be pursued here.
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 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.
In addition, not the whole original site is indexed. Serge Bondar\(^\text{17}\) 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,\(^\text{18}\) 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.

---


\(^{18}\) One byte corresponds roughly to one alphanumerical 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.
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 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\(^{19}\) 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.\(^{20}\)

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.\(^{21}\)

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\(^{22}\) 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\(^{23}\)

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

\(^{19}\) This has not been confirmed by literature or correspondence with the operators of any of the search engines.

\(^{20}\) 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.

\(^{21}\) This is based on information made available from representatives of one search engine.

\(^{22}\) Case 06/10.928/C.

\(^{23}\) Jeg er veldig lite fornøyd med de to siste punktene, både systematisk og innholdsmessig. Vi må se nærmere på hvorvidt det skal med, og eventuelt under hvilket perspektiv det skal diskuteres.
it is presumed that the database is protected, a presumption which may not be trivial. 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 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 a re-utilisation of these parts. To be unlawful as such 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 database 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 database.

(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) 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 engines. 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, i.e. 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 be 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 solutions – 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 where the court held that the defendant had “established a likelihood of proving that Google’s creation and public display of thumbnails” directly infringed the copyright of the defendant. The case had other elements which are relevant in the decision of the court. 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.

2.5 The sui generis database protection

Referring to the Belgian decision above, reference is made to the sui generis database protection under European law.

\[26\] United States District Court, Central District Court of California, Perfect 10 v Google, Inc et al (CV 04-9484 AHM (AHx)).

\[27\] United States District Court, Central District Court of California, Perfect 10 v Google, Inc et al (CV 04-9484 AHM (AHx)) page 33.

\[28\] For instance that Perfect 10 marketed scaled down images for mobile telephones, the thumbnails directly competing which this service.

Databases may be protected compilations under copyright law, 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 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.