Integrated legal information retrieval: new developments and educational challenges

Kees (C.) van Noortwijk [1]

Cite as Noortwijk, K. van, "Integrated legal information retrieval: new developments and educational challenges", in European Journal of Law and Technology, Vol 8, No 1, 2017.

ABSTRACT

The amount of legal information, available digitally, has increased gradually in the past three decades. We are now approaching a situation in which practically all legal information a lawyer needs on a daily basis can be obtained from digital sources. At the same time, powerful retrieval systems capable of integrating these sources and performing more effective search operations have become available. In this paper, new possibilities are outlined that have emerged now that such a large proportion of legal resources have been combined in unified collections. Also, the need to incorporate more advanced 'legal information skills' in the legal curriculum will be discussed. It will be argued that these skills are required to ensure that all newly educated lawyers will be able to use digital legal information optimally, now and in the future.

Keywords: Legal information retrieval; content integration; semantic searching; search intelligence; relevance ranking; legal information skills

1. INTRODUCTION

In the last decade, lawyers have come to rely on digital information sources in almost every aspect of their work. Traditional information sources such as books and journals have largely been replaced by their digital counterparts. Many law firms have already responded to this development and have abandoned their paper libraries in whole or in part (Dunlap 2014).

This transfer from paper to digital legal information has made it necessary also to adapt the way in which legal research is conducted. Not only because digital resources are often organised differently and make use of various specific mechanisms to access the required data, but also because the increasing 'completeness' of the digital collection (the great majority of new or re-issued publication being available digitally) opens up for entirely new ways to conduct legal research.

The latter is specifically true if efforts are made to combine as many relevant sources as possible, not only 'open access' ones but also, for instance, periodicals and books from commercial publishers. This objective, sometimes referred to by the term 'content integration' or 'content aggregation', not only simplifies searching (in one large collection instead of several smaller ones) but also makes it possible to cross-link information in several ways and even to implement certain forms of 'automatic classification'. An example of the latter will be given in section 5 of this paper. Furthermore, the filing of search results and the inclusion of signalling mechanisms (which point out new additions to the content to users, for instance based on previous queries) can be brought to a new level in systems like these.

Including more options, such as an option to show specific types of documents linked to, or linking to, the current document, or the option to set automatic notifications based on a search query, can make the user interface of a retrieval system considerably more complicated. Because of that, specific skills might be necessary with users, to ensure that they are not only aware of the existence of such added possibilities, but are actually capable of applying these in practice. It is a fact that many law schools already offer 'information skills' courses to their students (Margolis & Murray 2012). These usually cover the basics - which data collections are available, how does keyword search work, how can results be refined, how can a retrieved document be saved or printed - but often skip the more advanced functions (such as using linked information, implementing notifications, filing and re-using results, etc.). In itself that is understandable, as many of those functions require a certain level of familiarity with the wide range of different sources available within modern integrated retrieval systems and with their respective contents and relative importance, which undergraduate students might not yet possess. However, it is not just that, even experienced lawyers sometimes have trouble using all available options in modern search applications. They know exactly what they are looking for, but lack knowledge about certain technical aspects of the searching and processing of content, and therefore get suboptimal results.

Given all this, it is essential to improve education with respect to the use of the current generation of retrieval systems for digital legal content. At the same time we should continue efforts to make these systems - not only the basic functions, but also options that have become available since content collections are integrated and crosslinked, as well as options to use

search results more productively - easier and more straightforward to use, which will benefit students and legal professionals alike. Some examples of such improvements will be given in this paper. Their importance for the current period will be explained, as will be the benefits of future developments, such as the addition of 'conceptual' information retrieval methods.

2. DIGITAL LEGAL SOURCES

Although lawyers have often been said to work with legal sources in very traditional ways, preferring to use 'paper' books and journals in the way they have known since law school, they actually have been using digital sources for over three decades already. Legal professionals and legal researchers already used online databases with full text retrieval systems in the 1970s. The Lexis system, originally developed as part of a research project of the Ohio Bar Association in 1968, was an early example of a system capable of full text storage and retrieval of legal documents (Leith & Hoey 1998, p. 73). Case reports and legislation were the types of legal information available in the highest quantities digitally, at least in those days, whereas legal comments and literature followed somewhat later. This means that digital legal information, although sometimes considered a relatively new phenomenon, has already been available to a whole generation of lawyers. It is now extensively used for performing legal research, not only by academics writing extensive comments on legislation or case law, but also by practising lawyers applying those comments and searching for cases similar to those of their clients.

Given those facts, one would expect that using digital legal information would be part of every practising lawyer's daily routine nowadays and would definitely be a skill required for, and taught to, all law students. Many lawyers will admit that their abilities on this could be improved, however, and the amount of time dedicated to this subject in legal as well as in professional education is often surprisingly low. It is almost as if skills to deal with digital information are considered something that everyone develops 'naturally' these days. We all use the internet, don't we?

The point is of course, that the basic functionality of most information retrieval systems hardly presents problems to most users. But more advanced functions - such as using automatically generated crosslinks to find relevant legal comments for certain legislation, or to use a notification function in such a way that only relevant new documents will be shown, or to add (parts of) retrieved documents to a digital dossier shared with colleagues - require additional study and practice, the time needed for which is often not invested. The question is then if that is really a problem. Should modern computer software not be user friendly enough to be used without prior training? Indeed, almost every user may succeed in performing basic search and browsing operations in one of the major legal retrieval systems, by typing a few words in a single-line search field ('Search all content') and clicking the 'Search' button. And lo and behold, indeed lots of case reports and other documents then pop up in a list of search results, some of which are even relevant to the query! That is the moment many users (lawyers, too) feel they do not really need any special information skills. Anyone could do this!

Given the fact that many retrieval systems, including specialised ones concentrating on a particular domain, have access to several millions of documents, it should come as no surprise that even rudimentary queries will deliver a few relevant results. But is that enough? Specifically professionals need *complete* information, in order to be able to assess the subject properly. After all, the term 'information' (in the sense used here) is generally defined as 'the capacity to reduce uncertainty' (Klir 2005). When can we consider our information complete and therefore our uncertainty minimalised? How many 'hits' are necessary for that?

Unfortunately, it is impossible to give a general indication for that. Not only will the relevancy and with that the 'informational value' differ greatly between all documents retrieved. Apart from that, even the (relative) amount of documents retrieved by a query is difficult to judge. Users normally have no idea about the actual number of documents on a particular subject that are present in the database. Therefore, the 'recall' factor - the ratio between the number of relevant documents found and the number of those relevant documents actually present in the database - is normally difficult if not impossible to calculate or even estimate. For the 'precision' factor - the ratio between the number of relevant 'hits' and the total number of hits from a certain query (further explained in Meadow, Boyce & Kraft 2000, p. 321-328) - that is usually easier, but the effect of that could be adverse. Users often have the idea their search actions are successful when most of the presented hits prove to be relevant. Nevertheless, that might only concern a very small proportion - maybe only a few percent - of all the relevant documents present in the database, most of which were missed by the - possibly far too strict - parameters of the user's query. In that case, a lot of the available information remains hidden in the database.

As follows from the previous paragraph, the recall factor of a legal information system - indicative of the amount to which the system is capable of delivering all available, relevant information to the user - is more important from the *legal* perspective, whereas precision is a factor that should mainly be addressed from the *technical* perspective. Although it might be difficult to specify and assess the exact amount of information needed for a particular task, for instance for supporting a lawyer's argument, documents that are present but not retrieved and therefore not consulted definitely would present the hazard of missing something important. That is an essential point law students - who are for instance gathering information to complete an assignment - should be made aware of. Legal practitioners should also consider the fact that information *they* miss could be found and used by the other party's lawyer.

3. ADVANCED RETRIEVAL SYSTEMS - CONTENT INTEGRATION AND CONTENT AGGREGATION

The high number of available digital legal resources often complicates their practical use. Part of these resources consist of publicly accessible materials, such as legislation and case reports that can be retrieved from public websites. Another, major part consists of commercial publications from legal publishers, available through proprietary retrieval systems. And last but not least, lawyers and law firms usually compile extensive collections of documents themselves, often referred to as 'knowledge' or 'know how' documents, which they wish to

include in their research. It is not uncommon, therefore, to use five or more different databases, each with its own retrieval system, to perform a single research task.

Enter the so-called Content integration (CI) systems (Stonebraker & Hellerstein 2001, p. 552; Van Noortwijk 2011, p. 185). These are retrieval systems that are, in essence, operating independently of content to be retrieved, but capable of integrating multiple existing databases and retrieving content from these from one central console. To achieve that, the content integration system scans the separate, existing datasets and indexes every document it finds in them. To the user it presents itself by means of a more or less standard database retrieval interface, offering options for *searching* (usually by means of full text queries) and for *browsing* the content that was indexed. To the user, all content seems to be in one huge database (which is in fact true as far as the index is concerned) whereas the original documents are still in their respective, original databases. At the moment the user opens a particular document - from a list of retrieved documents or while browsing - the CI system can obtain that from the original database and display it in a new browser window, or 'framed in' in its own user interface. Overall, working with a CI system is like working with a Google variant that has access to all resources that are relevant to a lawyer.

Content Integration, as described here, has to be distinguished from Content Aggregation (or Resource Aggregation) (Selberg & Etzioni 1997; Sreekumar & Sunitha 2005) . That term is usually reserved for services that do not actually integrate document collections, but are capable of 'commanding' separate searches in multiple existing document collections, from one central interface. The original database search engines perform the actual searching and results are combined afterwards. For browsing purposes, aggregator sites often download brief descriptions (for instance titles and abstracts) from the separate document collections. When a user then selects one of these, or clicks on a 'hit' presented by the search function, the corresponding document is retrieved from the database where it resides, and is shown from there. Aggregation systems are relatively easy to implement, as the majority of professional databases not only provide user interfaces that give us the possibility to search and browse their contents, but also so-called web services that can be consulted by automatic processes (such as the search algorithm of a content aggregator's retrieval system). That means that no special software needs to be developed to perform these 'distributed search operations'.

There are also drawbacks to content aggregation, however. Performance of the search system can be problematic, as it is dependent on the response time of the separate database search engines. More importantly, the actual level of integration of the complete collection usually remains limited, because the documents themselves cannot be analysed and - whenever relevant - linked to each other across the borders of the separate databases before the search operation takes place. That makes it much more difficult, if not impossible, to show related documents or documents with similar contents together with a single document retrieved by the user.

Content integration, on the other hand, makes all that possible in an integrated retrieval system that is just as fast as each of the separate retrieval functions of the databases from which the content is obtained. This content is read and indexed beforehand, making subsequent search operations in the separate databases unnecessary. The system can show an

integrated list of results quickly, without the need to consult any external data collections at that time. Links between documents can be established at indexing time, with no restrictions as to the origins of these documents. Such links can be added to the indexed content in the form of extra metadata, producing a collection that is homogeneous with respect to the parameters that can be used for retrieval. Because of these characteristics, CI systems can save time when performing legal research, while at the same time making it possible to increase the quality of the output, for instance because of improved retrieval of linked information.

The result of this is not only an increase in the quantity of documents retrieved. Because the whole collection is prepared and used as a single set, the ranking of all search results can be optimised. Furthermore, documents referring to for instance the same case but from different sources (for instance, from the case law collections of different publishers) can be clustered together and shown as a single entity (with links to different expressions of that entity). This then leads to a more compact and at the same time more transparent list of results, in which documents that are probably most relevant appear on top. Together with the fact that all sources can be cross-linked, which enables the user to 'follow' references to related documents immediately, this not only results in higher amounts of documents being retrieved and presented, but can also increase the quality of information delivery considerably.

4. CONTENT INTEGRATION - OTHER ADVANTAGES AND COMMERCIAL APPLICATIONS

The application of CI can have additional advantages, specifically in professional environments. Because of the fact that such a wide selection of resources is effectively joined together to form one single collection, the system can become the focus point for gathering and storing information for a whole organisation. That can be achieved by adding to the CI system a possibility to group retrieved documents (or, even better, links to these documents) in custom dossiers (or files), together with an option to add extra information to such dossiers. Within organisations, the dossiers could be shared with colleagues, making this a very effective way of managing knowledge and know how.

At first sight, this joining and connecting of legal sources might look similar to what can be found in well-known textbooks or comment editions, such as Chitty on Contracts [2] or Allen's Textbook on Criminal Law. [3] Of course these are authoritative sources, but in essence they are static (at least, until the moment a new edition appears). CI systems provide a 'live' and therefore dynamic combination of all available legal documents (texts, legislation, case law, official governmental publications, etc.). These are not just added to one large database, but are actively connected together. This makes it possible to find and open related information from any document consulted. Furthermore, users can add to these functionalities themselves, by forming private or shared collections of links to specific (sets of) documents, which can be completed by uploading extra documents, for instance from a recent case or from one's private know how collection.

Another option is the inclusion of notification services, which can be tuned to deliver certain content that is newly added to one of the sources (databases) covered by the CI system. This could either be based on a particular source itself (if any new content appears in it, for instance

in the form of a new edition of a journal, the user is notified) or on a previous query that a user has stored. In the latter case, the query is in fact repeated periodically by the system, and any new content that is found is included in the notification.

CI is in fact not a new technology, many publishers use it - to some extent - in their digital portals that can be used to retrieve content from all publications for which the user holds a subscription. What is new here, is that sources from *different* publishers are combined, together with publicly available sources (legislation, case law) and optionally private sources from a particular user or organisation (only available to themselves, not to other organisations). The reason why this has received a lot of attention in The Netherlands, in the past decade, is that here, legal data have always been relatively scattered, with over 10 legal publishers and numerous important public sources. Given that, there was a lot to gain for, for instance, law firms if all content relevant to them could be retrieved through one portal. Some of these firms even took the step of developing CI technology themselves; just to optimise access to legal data for their employees. Because these law firms were important customers, the publishers - in some cases maybe reluctantly - chose to cooperate and to make their content available to several specialised organisations that offered CI technology for the legal market commercially. After a few years, two of these organisations remained: Legal Intelligence [4] and Rechtsorde. [5] Although in the meantime, these two companies have been the subject of takeovers, and are in fact owned by two of the largest publishers now, this has not altered the fact that they are licensed to integrate the content of all legal publishers in their systems. There seems to be a win-win situation, publishers can sell more content when that content can be retrieved and used effectively.

5. NEW WAYS TO RETRIEVE LEGAL INFORMATION

CI technology is not only important because of the integration of sources, it also opens the possibility to search and retrieve information from these sources in new and more effective ways. I will give three examples of that in this section.

5.1 SEARCH INTELLIGENCE

The first example focuses on the initial searching of content. Most legal information retrieval systems, for instance those supplied by publishers together with particular content sets, focus on full text retrieval. The content is divided in manageable 'documents', which can be searched and retrieved by specifying a *search query*, one or more words the user expects to be present in the documents that he or she is interested in. These documents are then shown in a 'hit list', often ranked according to a calculated relevance factor or to the publication date of the documents. This is in itself an effective way of working with collections of text based data [6], it is in fact the same way we have become used to search the vast contents of the World Wide Web by means of retrieval systems like Google and Bing. Nevertheless, this way of searching definitely has its flaws when optimal *recall* is required, which is usually the case for legal professionals, as was argued in section 2 of this paper.

Optimal recall can only be achieved if we make sure that with an *initial* query, as many documents that could possibly be relevant are put in the initial list of hits as possible. This list of hits can then be refined step by step, by means of 'facets' (such as the type of document, the

source it was published in, the area of law, etc.) while carefully assessing the results of each step. The essential point is: any relevant document missed (not retrieved) by the original query, will stay out of the set and will diminish the recall during all subsequent steps. That is why it pays, specifically in legal information systems, to optimise the results of the initial query. Several ways exist to do that, the common element in which is that they try to look beyond the specific form in which the user has typed the query. Instead of just taking the terms in that query for granted, algorithms are used to find out what they could *mean*, what the user's intention might be to enter these terms, in this order. For instance, if the user has typed a number, the name or abbreviation for a certain piece of legislation, and the word 'comments', it is probably not very useful to retrieve just documents that contain these three elements. Instead, the system should look for documents from 'legal comments' editions, using the article of a law that can be derived from the number and the law name (or abbreviation) as a criterion to search those documents, be it in their 'body text' or in the metadata they contain. The latter is of particular importance for publisher's content, as relevant law articles are commonly added as metadata by the editorial staff of these publishers. Another example might be the automatic addition of synonyms to a search query and the recognition of well-known legal terms to add corresponding articles of law or even certain case law identifiers to the query. All such additions to basic full text searching can lead to improvement of the legal quality of retrieval results, and with that usually also of the recall that is achieved.

5.2 LINKED CONTENT

The second example of more effective access to relevant documents is to use crosslinks in document collections. This enables users to find relevant documents that are not part of the set that is retrieved by an initial query and subsequent refinements. Using crosslinks, recall can be improved further. The links between documents that are established by the CI system itself or by, for instance, a publisher play a prominent role here. Such links can be direct: one document refers to another and this is implemented as a functioning hyperlink to open the second document from the first. Or they can be indirect: two documents both refer to the same article of law, or to the same precedent case, which makes that they can both be retrieved via that third, linking document. These powerful possibilities, implemented in CI systems, require additional skills with the user, because they usually work best when a search operation is conducted in a particular order (for instance, search for an article of law first, then find related content using links) and because they require knowledge about specific options in the CI system.

5.3 SELECTING RELEVANT SUBSETS

Finally, the third example I would like to give describes the importance of uniform metadata by means of which the data can be divided in relevant subsets. Defining such subsets, to which documents can belong, in fact entails the addition of extra metadata to these documents. This makes it possible to retrieve them (or filter for them) more flexibly, which not only improves search precision, but can also be beneficial for recall. A requirement for the latter is that it is ascertained that relevant subset metadata (or, in other words, classification data) are present in - and if necessary added to - all document that can be retrieved by the CI system.

Subsets that can be distinguished easily, and essentially for every document, are based on such characteristics as the source it was taken from (journal, book, web site), the location within that source (edition, volume, chapter, section) and often also the 'information type' they belong to (case law, commentary, journal article, news item, model document). Metadata describing these characteristics can be added to practically every document, which makes it possible to direct a search towards the parts of the content that are specifically relevant to it. Defining subsets based on for instance the area of law a document belongs to however, is usually more complex. One reason for that is that there are many of such areas and there is only limited uniformity in the way they are named. That could make it necessary to, for instance, 'map' area names from publisher 1 to those of publishers 2 and 3. Otherwise, we could easily end up with an integrated search system that contains overlapping classes such as 'civil law', 'civil and trade law' and 'trade and insurance law'. Not very useful to pinpoint the exact category of documents we are interested in. If no action would be taken, a user could unintentionally refine a search query just by one of these categories, which would result in only a subset of potentially relevant documents - maybe just the ones from one particular publisher that added the metadata for that category to its content - to be selected, influencing recall negatively. Therefore, creating uniformity in subsets (such as the area of law featured here) is essential for an effective CI system. Mapping of the subset information found in certain parts of the data (for instance, all content from a particular publisher or organisation) is usually a good way to achieve that.

Unfortunately, quite a number of documents usually lack the information that is necessary to classify them for relevant subsets. That is for instance true for a lot of case law, for instance from the European Court of Justice (published at the Curia and Eur-Lex web sites). Of course we could attribute an area of law to these like 'EU Law', based on their 'origins', but that would ignore their actual subject area (for instance: trade law or intellectual property law). A solution that has been tried for that particular problem, in the Rechtsorde system mentioned earlier, is to use automatic classification technology, a technology that is part of the field of computer science that is known as 'machine learning' (Mitchel 1997; Van Noortwijk, Visser & De Mulder 2006). Documents that lack the necessary metadata to decide about the subsets they should belong to, are classified automatically by comparing them to sets of example documents, one set for each 'class' or subset. They are then attributed to the class they share the highest number of characteristics with. Using advanced technology such as this helps to create more uniform classifications within large data collections with dissimilar roots. This, in turn, makes it possible to retrieve documents from such collections (like those that are combined in CI systems) more effectively.

5.4 RELEVANCE RANKING

Optimising the recall of search operations, using methods like those described above, is vital but of course could lead to higher numbers of retrieved, potentially relevant, documents. To make sure that users can still find the most relevant documents within this larger 'base set', retrieval systems need to be equipped with additional selection functions. Options to refine search results are of course part of that, but the most important feature in this category these days is *relevance ranking*.

The basis of most ranking mechanisms is usually formed by 'full text' query matching. A document receives 'ranking points' based on the presence of query terms. If terms are present close to each other, extra point can be awarded for that. Furthermore, the amount of points per term is usually related to the document size, to avoid disproportionate ranking advantages for large documents (which, just by their size, have higher probabilities to contain a ranking term a number of times). [7] In modern retrieval systems, this full text base layer is often enhanced with multiple extras. A document can be granted additional ranking points - and therefore end up higher in the list of search results - for various other characteristics, such as the source it comes from (more points for important, primary sources), its topicality, a match with specific metadata such as the author's name or an article of law, or for the number of other documents that refer to it. The ratio for the latter is of course that important documents (cases, journal articles) often are referred to in many other documents, which increases the probability the user might be interested in its contents too. The position in the list of search results for each document is determined by the sum of all ranking points it receives, and therefore by a sometimes extended mix of factors.

The objective of all such efforts for achieving optimal ranking is to ensure that, even when the number of search results is high, users can still find those 'hits' that are most relevant to their search query, at the top of the list of results. With a properly functioning relevance ranking mechanism, users no longer need to keep on refining queries until a small number of results remains. Such refining always bears the risk of cutting off parts of the content that, although not in a certain refine category, are still important for the user's query. With proper ranking, users can still find what is important for them in a 'hit list' containing thousands of documents. This has proven to work in internet search engines like Google and Yahoo, but can work even better when it is carefully tuned and applied in a system that contains information for a single, specific domain, such as the legal one. [8]

6. THE CHALLENGES OF SEMANTIC SEARCHING

The 'search intelligence' and relevance ranking, described in the previous section, can improve retrieval results considerably, especially in a domain-specific retrieval system. It cannot be denied, however, that the way in which search intelligence is usually implemented, namely by checking for common patterns and well-known legal terms in queries, is still far from optimal. The number of different patterns that can be recognised reliably can be rather limited in practice, whereas the recognition of legal terms is usually dependent on precompiled lists of terms together with searchable elements (such as articles from legislation, case law report numbers or other identifiers) that need to be maintained manually.

Understanding the *meaning* of a user's search request and responding to that is still one of the most challenging aspects of information retrieval. Lawyers know what a document they are looking for should be *about*, what *concepts* it should deal with, but usually have difficulty in describing *what it should look like*. The latter is in fact the only thing that retrieval systems have direct access to, however. Full text search operations take care of selecting those documents that contain certain terms, which are in fact no more than rows of characters to them. Results can be enhanced somewhat by also looking for different word forms, synonyms or the relative position of the terms in the document. The fact remains, however, that the vital 'conversion'

of the subject area that the user is interested in, into the manifestations this takes in actual documents, in full text retrieval systems is essentially left to the user. Additions to the retrieval process like those described in the previous section may alleviate this limitation somewhat, but are usually incapable of solving it completely.

'Semantic searching', which aims at understanding the user's intent in performing search operations, in the last decade has become a popular proposition to overcome such problems. It is a broad term, which usually includes such things as context recognition, query generalisation by using various types of related terms, but also 'concept based searching' (Guha, McCool & Miller 2003, p. 702). Specifically the latter could, in my opinion, bring real improvements to legal information retrieval, compared to full text searching, as it enables users to search for 'concepts', instead of for manifestations of keywords in documents. Several methods to implement this 'searching for semantic concepts' have been described so far, for instance distinguishing approaches that use *explicit* concepts (for instance: legal concepts like 'tort' or 'fundamental breach') from those that use *implicit* ones. The latter can for instance be generated by extracting latent relations between the words (terms) in documents, or by calculating probabilities based on the co-occurrence of terms in documents (Egozi, Markovitch & Gabrilovich 2011, p. 8:2). Explicit concepts can be defined using comprehensive taxonomies of a domain, but could, as Gabrilovich & Markovitch (2006) have shown, also be based upon data from for instance Wikipedia (concept term and explanatory text).

Katz, Bommarito & Blackman (2014) describe a modelling approach, in which variables from a set of Supreme Court cases are used to construct prediction models for the outcome of the respective cases. This is in fact another way to establish a relationship between certain characteristics of documents and their meaning automatically. In that respect, this again might be a - potentially very powerful - technology that could be used for semantic searching.

An interesting approach of the use of implicit concepts for legal information retrieval can be found in De Mulder *et al.* 2010. The application described there is capable of selecting documents from a set of eyewitness reports (which are used to deliver legal proof). The selection is made automatically after the user has 'trained' the system by marking a few reports according to their truthfulness. This means that when such a training procedure is used, a retrieval systems could facilitate its users in finding documents, not based on the presence of specific keywords, but based on their similarity to (certain aspects of) the training documents. Although a retrieval system solely based on the principles of conceptual searching might be impractical (certainly if it would indeed implement example-based querying as described in the previous paragraph), this technology can be combined with more traditional retrieval methods, for instance to refine a preliminary set of results. This is generally known as 'relevance feedback' (Salton & Buckley 1997). It provides users with the option to mark, in a set of retrieved documents, one or more 'hits' that are relevant to them. Next, the system can look for similar content (within or outside of the current set). This technology, which has been tested in various forms in the past decade, can be expected to find its way into integrated legal information retrieval systems in the near future. When combined with existing retrieval functions like those described in section 5, that could result in systems capable of achieving both high recall rates and good search precision.

7. INFORMATION SKILLS IN LEGAL EDUCATION - FROM TRADITIONAL TO FUTURE PROOF

This brings me to the final point in this paper. As can be concluded from the preceding sections, systems capable of retrieving legal information from large, integrated collections are a reality these days. They make it possible to combine all digital sources a lawyer needs in one huge collection, which can be searched from one single user interface. By adding elements such as the filing of selections of documents in shareable dossiers, these information systems are on their way to become the central 'hub' for knowledge management in many legal organisations, including the major law firms. At the same time, examples were given of functionalities these systems contain - or could contain in the near future - which are important for their effective use. Together with the specific information needs that inhere in the legal profession, the conclusion must be that lawyers, and certainly also students learning to become a lawyer, should be trained to operate these systems properly. [9]

Of course, information skills have always been part of the law school's curriculum, to a certain extent. Students are told about traditional sources, to be obtained from the library, by most of their teachers. Apart from that, basic information skills courses provide them with the information necessary to use digital legal sources. For this, e-learning tools are often used, such as those described by Smith & Presser (2005).

Still, this often stops at a rather basic level. Students are taught about existing legal resources and their importance, about using these in developing a legal argument and referencing them correctly. The impression seems to exist that students, because of their use of the internet starting at the age of four or five, have more knowledge about data retrieval than the average university teacher and therefore do not require any training on the subject. This, however, is a misunderstanding (Peoples 2005, p. 678; Palfrey 2012, p. 120).

It is true that practically every (legal) information retrieval system these days has a user interface that is in itself simple and straightforward to operate. But research shows that this is no guarantee that students will be able to use it effectively (Andretti 2001, p. 261-262). Without sufficient knowledge about the (often extended) contents of these systems and the more advanced retrieval functions they contain, legal information retrieval could become some sort of a lottery: there will be an outcome, there might even be people who are pleased with it, but it is far from optimal.

In my experience, students are often surprised when the importance of a search operation's recall factor is discussed in class. They have become used to a situation in which the second page of results from a retrieval system (read: Google) is seldom inspected, as the first page usually already contains one or two useable hits, and who needs more than that? Therefore, when teaching legal information skills, some attention should be paid to theoretical aspects of retrieval processes as well.

The second element that should definitely be incorporated in every legal information skills course is the importance of using linked information. As illustrated in the previous sections, searching by means of full text queries, even if supported by intelligent features capable of

recognising patterns and adding synonyms to the query, always will have its shortcomings. When link information, present in the documents already retrieved, is used to find related documents, for instance based on metadata such as relevant law articles, items that would otherwise have been missed completely can be added to the collection that is retrieved. Specifically CI systems contain very powerful options to achieve that, and learning to use those will be a vital skill for every lawyer. This can be achieved by developing adequate training materials, for instance in the form of practical assignments with explanations, which effectively illustrate the use of the functionalities available. Such materials are also, to some extent, made available by suppliers of CI systems, which organise seminars and 'webinars' on working with their products on a regular basis.

CONCLUSIONS

Digital legal sources, although still seen by many as 'the new way to gather legal information' have in fact already been around for over thirty years. The last decade has seen a development towards integrated and far more intelligent retrieval systems, which are capable of supporting lawyers very effectively in conducting legal research. In this paper, some examples were given of new functionalities present in many of these systems. Possibilities occurring when 'semantic search' options would be incorporated were also described. These functionalities are of importance to legal professionals because they can improve the recall rate - and in some cases, the precision as well - of a search operation: a larger proportion of the relevant documents present in the huge, combined data collections can be retrieved with limited 'false hits', which can be vital for any lawyer involved in, for instance, a legal dispute or in litigation.

Because of this, operating advanced legal information systems should be taught to students of law schools, as part of their training in 'legal information skills'. The fact that most students already have experience in browsing the internet is not a sufficient guarantee they will also be capable of working effectively with such retrieval systems nor that they will be able to use whatever they find with these systems effectively.

REFERENCES

Andretta (2001)
Andretta, Suzie (2001),'Legal information literacy: a pilot study', *New Library World*, Vol. 102, Iss 7/8, p. 255-264.

Dunlap (2014)
Dunlap, David W. (2014). 'So Little Paper to Chase in a Law Firm's New Library', *New York Times*, Oct. 22, 2014, cited in O'Grady 2015.

Egozi, Markovitch & Gabrilovich (2011)
Egozi, O., Shaul Markovitch and Evgeniy Gabrilovich (2011). 'Concept-Based Information Retrieval using Explicit Semantic Analysis', in: ACM Transactions on Information Systems, Volume 29, Issue 2 (April 2011), (New York: ACM 2011, p. 8:1-8:34).

Gabrilovich & Markovitch (2006)
Gabrilovich, Evgeniy; Markovitch, Shaul (2006), 'Overcoming the brittleness bottleneck using Wikipedia: enhancing text categorization with encyclopedic knowledge', in: *Proceedings of the 21st National Conference on Artificial Intelligence* (AAAI 2006), p. 1301-1306.

Guha, McCool & Miller (2003)
Guha, R., Rob McCool and Eric Miller (2003). 'Semantic Search', in: *Proceedings of the 12th international conference on World Wide Web* (New York: ACM 2003), p. 700-709.

Hazelton (2012)
Hazelton, Penny A. (2012). 'Law Students and the New Law Library', in: Rubin, Edward (Ed.) (2012). *Legal Education in the Digital Age* (Cambridge: Cambridge University Press 2012), p. 158-182.

Katz, Bommarito & Blackman (2014)
Katz, Daniel Martin, Michael James Bommarito, and Josh Blackman (2014), 'Predicting the behavior of the Supreme Court of the United States: A general approach', available at SSRN 2463244 (2014).

Klir (2005)
Klir, George J. (2005), *Uncertainty and Information: Foundations of Generalized Information Theory* (Hoboken, NJ, USA: John Wiley & Sons 2005).

Leith & Hoey (1998)
Leith, Philip and Amanda Hoey (1998), *The Computerised Lawyer* (London: Springer 1998).

Long & Chang (2014)
Long, Bo, and Yi Chang (Eds.) (2014). *Relevance Ranking for Vertical Search Engines* (Waltham, MA, USA: Morgan Kaufmann / Elsevier 2014, ISBN 978-0-12-407171-1).

Margolis & Murray (2012)
Margolis, Ellie, and Kristen E. Murray (2012). 'Say goodbye to the books: Information literacy as the new legal research paradigm', *Univ. Dayton Law Rev.* 38 (2012), p. 117.

Meadow, Boyce & Kraft (2000)
Meadow, Charles T, Bert R. Boyce and Donald H. Kraft (2000). *Text Information Retrieval Systems* (San Diego: Academic Press / Elsevier 2000).

Mitchel (1997)
Mitchel, T. (1997). *Machine Learning* (McGraw Hill 1997).

De Mulder *et al.* (2010)
Mulder, R.V. De, Noortwijk, C. van, Goldsmith, M., Pansky, A., Koriat, A. & Labin, S.K. (2010). 'CORMAS: A Computerized Tool for the Analysis of Eyewitness Memory Correspondence', *European Journal of Law and Technology (EJLT)*, 2010 (3), p. 1-18.

Van Noortwijk, Visser & De Mulder (2006)
Noortwijk, C. van, Visser, J.A. & Mulder, R.V. De (2006). 'Ranking and Classifying Legal Documents using Conceptual Information', *Journal of Information, Law & Technology (JILT)*, 2006 (1), p. 1-15.

Van Noortwijk (2011)
Noortwijk, C. van (2011), 'Computers and Law - The Central Role of Legal Knowledge', in P. Kleve & C. van Noortwijk (Eds.), *Something Bigger than Yourself* () (Rotterdam: Erasmus School of Law 2011), p. 179-187.

O'Grady (2015)
O'Grady, Jean P. (2015). *12 Building Blocks Of A Digital Law Library* (New York: Law360 2015, http://www.law360.com/articles/607548/12-building-blocks-of-a-digital-law-library, consulted January 6, 2015)

Palfrey (2012)
Palfrey, John, 'Smarter Law School Casebooks', in: Rubin, Edward (Ed.) (2012). *Legal Education in the Digital Age* (Cambridge: Cambridge University Press 2012), p. 106-129.

Peoples (2005)
Peoples, Lee F., 'The Death of the Digest and the Pitfalls of Electronic Research: What is the Modern Legal Researcher to Do?'. *Law Library Journal*, Vol. 97, p. 661, 2005. Available at SSRN: https://ssrn.com/abstract=767124

Rubin (2012)
Rubin, Edward (Ed.) (2012). *Legal Education in the Digital Age* (Cambridge: Cambridge University Press 2012).

Salton & Buckley (1997)
Salton, G. and Chris Buckley (1997). 'Improving retrieval performance by relevance feedback', in: *Readings in information retrieval* 1997 24(5), p. 355.

Schweighofer, Rauber & Dittenbach (2001)
Schweighofer, Erich, Andreas Rauber, and Michael Dittenbach (2001). 'Automatic text representation, classification and labelling in European law', *Proceedings of the 8th international conference on Artificial intelligence and law* (ACM 2001).

Selberg & Etzioni (1997)
Selberg, Erik, and Oren Etzioni. 'The MetaCrawler architecture for resource aggregation on the Web', *IEEE expert* 12.1 (1997), p. 11-14.

Smith & Presser (2005)
Smith, Nicki Mclaurin, and Prue Presser (2005), 'Embed with the faculty: legal information skills online', *The Journal of academic librarianship* 31.3 (2005), p. 247-262.

Spärck Jones, Walker & Robertson (2000)
Spärck Jones, K.; Walker, S.; Robertson, S. E. (2000). 'A probabilistic model of information retrieval: Development and comparative experiments: Part 1 & 2'. *Information Processing & Management* 36 (6), p. 779-840.

Sreekumar & Sunitha (2005)
Sreekumar, M. G., and T. Sunitha (2005), 'Seamless aggregation and integration of diverse datastreams: essential strategies for building practical digital libraries and electronic information systems', *The International Information & Library Review* 37.4 (2005): 383-393.

Stonebraker & Hellerstein (2001)
Stonebraker, M. and Joseph M. Hellerstein (2001), 'Content integration for e-business', Timos Sellis and Sharad Mehrotra (Eds.), *Proceedings of the 2001 ACM SIGMOD international conference on Management of data (SIGMOD '01)* (New York, NY, USA: ACM 2001), p. 552-560.

Thompson (2001)
Thompson, Paul (2001). 'Automatic categorization of case law', in: *Proceedings of the 8th international conference on Artificial intelligence and law* (ACM 2001).

---

[1] Erasmus School of Law, Rotterdam, The Netherlands.

[2] Beale, H. (Ed.), *Chitty on Contracts*, Sweet & Maxwell 2015.

[3] Allen, M., *Textbook on Criminal Law*, Oxford University Press 2013.

[4] http://www.legalintelligence.com

[5] http://www.rechtsorde.nl

[6] Or 'free format' data, as Leith and Hoey (1998, p. 32) called it, to distinguish it from record-based collections of data.

[7] A popular algorithm to achieve that is TF/IDF, in which the total frequency of a term is divided by the inverse of the number of documents the term appears in. This will diminish the 'weight' of common terms appearing in almost every document. Several refinements to this algorithm have been proposed, of which the (Okapi) BM25 variant (Spärk Jones, Walker & Robertson 2000) is probably most commonly used.

[8] See Long & Chang (2014) for a comprehensive overview of ranking methods applicable to 'vertical search engines' (information retrieval systems for a limited domain).

[9] See also Hazelton (2012), in which an overview is given of the way digital information sources can be integrated in more traditional library environments.