Integrated legal information retrieval – new developments and educational challenges

Kees (C.) van Noortwijk, Rotterdam

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 Traditional information sources such as books and journals have to a large extend been replaced by their digital counterparts.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.[1] 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 organized differently and make use of various specific 'disclosure mechanisms', 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 'conceptual information retrieval'. Examples of the latter include the automatic classification of documents and the searching for documents 'similar to' one that was already retrieved.[1] Furthermore, the filing of search results and the inclusion of signaling 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.

Making use of such more advanced options requires specific skills. Many law schools already offer 'information skills' courses to their students.[1] 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. In itself that is understandable, especially because such more advanced functions often require a certain level of familiarity with the wide range of different sources available within modern integrated disclosure systems, which undergraduate students might not yet possess. But it is not just that, even experienced lawyers sometimes have trouble using advanced search tools. They know exactly what they are looking for, but lack knowledge about certain technical aspects of searching, and therefore get suboptimal results.

Given all this, it is essential both to improve education with respect to the use of advanced disclosure systems for digital legal content, and to continue efforts to make these systems – not only the basic functions, but also the most powerful options – easier and more straightforward to use. Examples of both will be given in this contribution.

Digital legal sources

Although lawyers have often been said to work in very traditional ways, they have been using digital sources for over three decades already. Online databases with full text retrieval systems were already used by legal professionals and legal researchers 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.[1] 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 practicing lawyers.

Given those facts, one would expect that using digital legal information would be a piece of cake for every practicing lawyer 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 that more advanced functions require additional study and practice, the time needed for which is often not invested. The question is then if that is really a problem. Shouldn't modern computer software 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 don't really need any special information skills. Anyone could do this!

Given the fact that many retrieval system, including specialized ones, 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? In rare cases it might be, but usually it is not. Specifically professionals need *complete* information, in order to be able to assess the subject properly. But when can we consider our information to be complete? How many 'hits' are necessary for that? That's hard to tell, as a user normally has no idea about the actual number of documents on a particular subject that is 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. For the 'precision' factor – the ratio between the number of relevant 'hits' and the total number of hits from a certain query – that is usually easier.[2] And users therefore often have the idea their search actions are successful when the majority of the presented hits proves to be relevant. But 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.

I would even want to take this argument further, by stating that for a lawyer the recall factor is of much higher importance than the precision factor. If a lawyer misses even one single document, let's say a relevant case report, that could make him loose the case for his client. Therefore, the focus for lawyers should definitely be on optimizing recall, even at the cost of precision.

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 available 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

[1] Leith & Hoey 1998, p. 73.

[2] See for a further explanation on this Meadow, Boyce & Kraft 2000, p. 321-328.

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. 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. All in all, 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. 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 actual searching is performed by the original database search engines 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 analyzed 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.

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 are effectively joined together to form one single collection, the system can become the focus point for gathering and storing information for a whole organization. For instance when it is equipped with a possibility to group retrieved documents (or, even better, links to these documents) in custom dossiers (or files) and to add extra information to such dossiers. Within

organizations, the dossiers could be shared with colleagues, making this a very effective way of managing knowledge and know how.

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) that are 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. But 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 organization (only available to themselves, not to other organizations). 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 optimize 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 specialized organizations that offered CI technology for the legal market commercially. After a few years, two of these organizations remained: Legal Intelligence[3] and Rechtsorde.[4] 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.

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.

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[5], 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. But 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's why it pays, specifically in legal information systems, to optimize 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

[3] http://www.legalintelligence.com
[4] http://www.rechtsorde.nl
[5] Or 'free format' data, as Leith and Hoey (1998, p. 32) called it, to distinguish it from record-based collections of data.

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.

Linked content
The second example concerns that, even if certain relevant documents are not part of the set that is eventually retrieved by a query, such documents can still be obtained from the system, improving recall. For this, the links between documents that are established (by the CI system or by, for instance, a publisher) are vital. 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 de CI system.

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, again enhancing recall.

Subsets that can be distinguished easily, and essentially for every document, are based on such characteristics as the source they were 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. 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 organization) is usually a good way to achieve that.

But 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'.[6] 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

[6] See for instance Mitchel 1997 and Van Noortwijk, Visser & De Mulder 2006.

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 in CI systems) more effectively.

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 paragraphs, 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 organizations, including the major law firms. At the same time, examples were given of functionalities these systems contain, 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.[7]

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, and these days also get (usually quite brief) introductions to the major digital sources. But this often stops at a rather basic level. 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.

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 without sufficient knowledge about the (often very 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.

Especially the notion that high recall rates are important for lawyers, often surprises students. 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 discussed 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 recognizing 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.

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. These functionalities are of importance to legal professionals because they can improve the recall rate of a search operation: a larger proportion of the relevant documents present in the huge, combined

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

data collections can be retrieved, which can be vital for any lawyer involved in, for instance, a legal dispute or in litigation.

Because of this, advanced legal information systems should also be demonstrated 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

Hazelton 2012
Penny A. Hazelton, ‘Law Students and the New Law Library’, in: Rubin 2012, p. 158-182.

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

Margolis & Murray 2012
Margolis, Ellie, and Kristen E. Murray, "Say goodbye to the books: Information literacy as the new legal research paradigm." U. Dayton L. Rev. 38 (2012): 117.

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

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

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

O’Grady 2015
O’Grady, Jean P., *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

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

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

Thompson 2001
Thompson, Paul, "Automatic categorization of case law", *Proceedings of the 8th international conference on Artificial intelligence and law*, ACM 2001.