6th BILETA Conference
1991

Editor: Tim Green

MATURING AND AGEING IN INFORMATION RETRIEVAL. THE CASE OF CELEX

Constantinos Popotas [1]

Court of Justice of the EEC, L-2925, Luxembourg

Keywords: Information retrieval systems - Legal data banks - Integration of technologies

Abstract: CELEX, the public information retrieval system of the EEC, is an ideal environment for assessing inherent problems of information retrieval. Referring to a complex legal order and covering exhaustively the material thence emanating - in full text for most of its sectors - it represents an excellent support for judging research strategies against precision. On the other hand it can be used to illustrate problematic areas in information retrieval in order to propose improved approaches.

Introduction

There is nothing original in saying that information retrieval technology has not changed radically since the 5O's. Performances have of course been enhanced; still the existing systems did not claim a more substantial role than just providing information. Trapped in this preconception developers limited their approach to an effort to improve certain characteristics, but we should add here that existing criteria for judging performances of legal information retrieval systems seem to be quite vague and arbitrary, lacking reference to a universal measurement.

We also lack elements to appreciate when a system becomes mature and how handy the facilities it possesses are for the users. Volume seems to be an element much appreciated, which to our eyes means nothing when not seen in the context of the material actually existing This is probably the reason why systems start showing their age: managers emphasising maturity uniquely in the coverage area tend to forget to upgrade their services, if they do not find that it is uneconomical to do so.

Probably the only aspect by which we can judge the activity of a system - still in a quite subjective manner - seems to be interface friendliness, but there again we should not forget that it is in the interest of hosts to cater for a multitude of users with a variety of equipment ranging from TTY to WYSIWYG environments; still they could make an effort to replace their actual minimalistic approach by a multifaceted facility.

When does a legal information retrieval system become mature? How many years does this process take? These are questions that nobody can answer in advance.

The question of maturity and completeness seems therefore a very complex one. Very few systems can set the example in this domain. The reasons are relatively simple. First, because few are the systems that survived for longer than two decades. Fewer are the ones that can claim exhaustiveness of their data. Normally linked to broad legal orders capable of providing a representative but limited spectrum of documents in order to cover the needs of a variety of users, these systems were based on a preselection of their content for economic and practical reasons thus limiting eventually their pool of information.

In view of all these difficulties in obtaining a perfect system what kind of strategy can we develop? Instead of a conceptual model - panacea, we prefer an operational analysis based on the deficiencies of a system that has a very well delimited domain and use. There is a double reason behind that: first the personal experience with this system in the framework of a Community Institution; and the actual potential of it, being at the heart of a legal system that comes of age and manifests signs of information overflow.

CELEX - Brief historical introduction - Features and functions - Management aspects

CELEX (Communitatis Europae Lex) is nowadays a public information retrieval system. It was nevertheless conceived and initially implemented for internal use of the Community Institutions in an effort to master the ever increasing volume of the material produced. It obtained thus - from the very beginning - a function that exceeds by far the role of a common "informative" information retrieval system. It has been and is being used as a legal drafting tool, a testbed for case law consistency, and finally as a means for providing subproducts on paper form.

The first efforts to automate legal documentation within the Communities date back to 1967. It was still only in 1971 that CELEX, with a structure similar to its contemporary form, became operational and much later (early 80's) that it became accessible for public use through the EURONET networking environment.

Some 1300 external users are currently contracted with CELEX, accessing the system for around 15000 hours per year (statistics end 1990). Thousands of EEC officials access CELEX as mid- and end-users for their everyday work.

The full text organisation is the obvious choice for a legal order where the unambiguous wording is the *sine qua non* but fine linguistic variations maintain the necessary level of distinctions; extensive indexing needs to be applied in order to guarantee cross referencing and to maintain the possibility of alarm procedures given the complexity and particularities of the EEC legislative and judicial processes.

CELEX is today produced and managed by the Directorate General IX, division 1.4. with the participation of all the other Community Institutions and organs (Court of Justice, Council, BEI, Court of Auditors, Parliament).

The actual occupation in memory is of 5.5 giga for the lot of the linguistic versions. The system is operating on a Bull mainframe and uses the Mistral software as support. Some 115000 documents are contained in the French version which is considered the pilot language, while the user can obtain around 95000 documents through the other versions. This difference is explained by historical reasons: the French version - implemented as the very first - contains documents that are obsolete nowadays.

Multilingualism

It is important to note here that CELEX reflects also the choice made historically for a multilingual legal order responding to the principles and aspirations of the participating national legal orders. It therefore maintains 7 linguistic versions (DE, DK, EN, FR, GR, IT, NL - ES and PT soon to be added), while the command language exists in three versions (FR,EN,GR). This multilingual aspect might have been an element of political importance at the very beginning of the Communities' existence, but it is nowadays a paramount operational parameter since all linguistic versions of the documents have equal binding force (with a relative exception for the texts emanating from the Court of Justice), every document being officially translated into all the other languages.

In technical terms this aspect puts an extra burden on CELEX management. The coexistence of different alphabets placed in the same ASCII code together with special characters that must be maintained in order to run standard software, means juggling all the time at the limit of standards - especially 8bit computing ones.

Communications suffer substantially from this situation, with important investment going on for data communication equipment programming - especially for extensive filtering and interactive setup.

ParlicularIties of the EEC legal order

One needs also to consider the particular characteristics of the EEC legal order. Law is implemented within the Communities' "territory" through a complex system, depending much on the collaboration between EEC and national authorities; it emanates from different institutions. Its interpretation is guaranteed uniformity through a unique Court of Justice but needs in some cases to be applied by national Courts.

The legislative material is in many cases volatile, in other cases so "long-living" that consecutive amendments pierce and patch the basic documents in such a way as to make it difficult to use with certainty. And it is a fast growing body; today it reaches the order of 30000 documents, of which some 12000 are still in force.

The case law, on the other hand, reaching nowadays 6000 documents, follows an inevitable expansionism of the scope of the EEC legal order, while preserving the respect for precedent; the equilibrium sought is expressed by a very careful weighing of the Court's approach.

Coverage

CELEX covers the entire Community production which is of binding force or is interweaved in the law making procedure. It organises the documents according to the following sectors (in brackets the number of documents available on 21.2.91):

Legislation
Sector 1: Basic treaties and treaties amending them, Accession treaties, Single European Act (2400)
Sector 2: Agreements and other acts adopted within the framework of external relations (1448)
Sector 3: secondary legislation (29722)
Sector 4: complementary legislation (369)

Preparatory documents (13589)
Commission Proposals, European Parliament Opinions, ESC Opinions, Court of Auditors Opinions, other preparatory acts

Case Law (5925)
- Judgments and orders

- Opinions of the Advocates-general
- Seizures,third party proceedings, opinions(avis) rulings

National provisions implementing directives (18420)

Parliamentary questions (36394)

Two more sectors are under development:
- national case law related to EEC matters
- doctrine of the EEC law

You may be aware of a gradual implementation of different sectors milestones to CELEX's life and maturity. Particular emphasis is given to the full text structure of the documentary units where this is feasible and justified by the persistence of documents in time, since it is commonly recognised that legal texts obtain a particular usage in the hands of lawyers.

Structure of documents

CELEX documents are characterised by a detailed backbone which accommodates both for full text and for indexed fields.

Full text fields are loaded directly from phototypesetting; the original text of the document is introduced without any intervention apart from the labelling of fields. Before any document is introduced into the system, groups of lawyers analyse its content and create a series of fields containing keywords or citations. Indexed fields concern dates of events or deadlines, involved persons, states or institutions in a codified form, subject matters and relations between documents. It is at this stage that intensive cross referencing between CELEX documents is created by use of a unique CELEX number. These citations have paramount importance for either tracing the legislative history of a piece of legislation, checking with precedent in case law, or examining the application of certain provisions.

The unique document number is the central field of CELEX. This number is compiled in such a way as to be significant according to a set of conventions that were adopted by the CELEX team; it can therefore be used as a research criterion since every user can formulate its form. For example **3**84**L**0450 denotes a Directive(**L**) number 450 in **Sector 3 (Legislation)** during 1984.

In the very same way references to other Community documents can be obtained. Where necessary finer localisation can be used, either by citing the article and/or paragraph, or - as was the case in the past - pages, or other subdivisions of a document. Unfortunately in CELEX the physical document is still considered to be the documentary unit. Fields and dictionaries are defined, but they do not represent entities "per se", consequendy the system cannot refer directly to them.

Profiles of users

The absence of dynamic referencing and displaying made many users consider that CELEX is a system complementary to paper material. The system can be used in order to locate first where to fmd the information; following that one needs to check with the Official Journal or the reports of the EEC in order to physically obtain the document.

This impression was reinforced by the fact that - due to technical limitation - up to very recently, CELEX material was displayed uniquely in capital letters - something that in other terms made downloading meaningless.

This feature on its own was enough to disorientate users. Nobody expects anything more than the simple presentation of the information with eventually a printout - not of extreme elegance we must admit.

But, we can still isolate seven potential groups:

1. authors of documents. Practitioners fall within this category, but Community staff could well justify on their own aspirations possible extensions of CELEX features. What they expect from the system is to selectively operate on the text and the information in order to localise texts to be extracted or to group references to documents that they must cite.
2. linguists. They tend to use the system in a particular way: either as translators working for the Institutions or external workers attempting to extract precise expressions, they need to match expressions in a source language with a target one. CELEX in its current form proves difficult to manipulate as far as this use is concerned, still not without interest.
3. academics seeking to formulate a theoretical approach to EEC law. Given the lack of dynamic functions or statistical features within CELEX, the necessary abstraction and synthesis can be reached only by meticulous and imaginative use of the system. Even so, the volume of the printed EEC material justifies the investment in obtaining a thorough knowledge of search techniques.
4. documentation services that compile indexes and tables of authorities.
5. occasional users interested in having the latest selective information in specific domains.
6. students or lawyers wishing to learn Community law.
7. finally - not just a potential group - "managers" of EEC law, responsible for the follow up of the legislation. They furnish substantial feedback to the system.

Environment and constraints of use

Most contemporary information retrieval systems are owned by commercial enterprises. It is true that even in these cases quality of services remains the central aspect. Celex though can be clearly distinguished, since its main objective is not to earn money but to provide a working platform for Community staff and to diffuse, in the broadest possible manner, the Community law. Therefore it keeps itself away from the major dilemmas of its concurrent systems.

It is clear in the options taken by its creators that the structure of the system reflects the internal needs of the Community Institutions more than the intention to cater for the public. The mixed fulltext/indexing technique might appear contradictory to the intention to provide a widespread diffusion system, since the system becomes extremely bulky. On the other hand what this approach means is that external users have access to the same views as internal ones.

The command language is decent enough to provide for most of the facilities that are considered necessary to IR; the complaints against it refer normally to its mnemonic, the codified form which is difficult for the user to remember unless he is involved in daily use of the system.

Having to deal with the presumption that users would access the system through a variety of equipment, the system did not opt for technical sophistication; instead the directions were towards uptodateness and data independence while facing the main characteristic of Community activities: the will to operate in a multilingual environment. Undoubtedly this last feature added an extra burden to the already cumbersome system that was created in the beginning of the 70's. Raving to serve a population of 12 countries and 9 languages CELEX had to tackle the problem of technical constraints Anglo-Saxons are normally not aware of. As a direct consequence, a minimalistic profile catering for everybody's needs prevailed, a solution which is not satisfactory, since it creates an inflexible level of user-computer interface.

As was previously said, no dynamic dimension was added to references and citations, something that could allow for hypertext type navigation. Nevertheless experienced users are in a position to reproduce the historical dimension of a document - not without substantial effort - if they want to go beyond simple listings. One of the subproducts of CELEX - the "Repertoire des actes communautaires en vigueur" exploits exactly the detailed analysis of documents.

Uptodateness suffers recently from an effort to extend the struciure of the data base.While indexed fields are created within a reasonable time limit, full text fields - especially in the case law sector - represent an important backlog

After an initial period of 7bit communications and storage which placed lawyers in the uncomfortable position of working with text composed uniquely of capital letters, the choice was recently made for 8bit technology. Greek was the factor necessitating such a shift since it is placed in the upper part of the ASCII code, together with the accented characters of other languages. This brought together a whole wave of modernising efforts, once more as a recognition of the fact that CELEX was becoming old-fashioned.

Future developments

Aware of the shortcomings of the existing solutions, the CELEX group has opted for a modernisation of the data bases, starting with the management problems. The idea behind the whole operation is that file transfer programs can be used to pass the production of the different Institutions onto one computer in the Commission where the management and the preparation of the data will take place. When the documents are grouped and ready, uploading to the mainframe used for hosting the data bases will be initiated. A DBMS is used as the centralising scheme.

Another domain which is under development is the creation of intelligent interfaces. It is recognised though that such a facility should be available to non-experienced users while command languages should continue to constitute the core of the research procedure.

As far as the presentation aspects of the multilingualism are concerned, "comprehensive" presentation of characters, by use of two logical sets catering for all European languages, is now being implemented. Texts will be converted to normal typographical form, thus becoming susceptible to direct exploitation.

Attempt of a synthesis - proposals

Undoubtedly Celex proved its utility over a long and turbulent period. But we should not lose sight of some of its major disadvantages and problems in order to propose further improvements. In fact what we intend to do is to generalise our knowledge of its problems in order to extract ecumenical solutions.

There is a fundamental question to ask at this stage: is CELEX sufficiendy mature? The answer, connected to the life of EEC communities, should be positive. Very few systems can claim extensive coverage of the documents produced by a legal order. Moreover, rare are the cases of systems that maintain the structure and content of their data in such a uniform and consistent manner.

But, I cannot help criticising CELEX for the second aspect presented in the title of this paper: focusing on exhaustiveness of content is not sufficient in our era of invasion of lawyers' offices by microcomputers. CELEX needs to keep up the modernising effort, it has to be "renovated" by an opening towards new ideas. It's illogical to fear a change of its aspect or a negative impact on its reputation since this particular system combines a rigorous structure with sufficient data independence which guarantees a smooth evolution.

Undoubtedly the established point of view that information retrieval systems serve mainly the diffusion of information limits everybody's angle of view and restricts future developments. Nevertheless, problems inherent to the domain in which the information retrieval technology is applied have a substantial impact on the characteristics of any system.

The multilingual structure of CELEX allows for matching searches in the different linguistic versions that put under stress measurements of precision, as well as questioning the efficiency of translation techniques and the possibilities of constructing thesauri to cope with translation deficiencies. Take for example a very standard expression in the case law:

"Il est de jurisprudence constante"

(French being the pilot language for the Court of Justice, it can be used as the basis for our measurements.)

The English equivalent is:

"(the Court) has consistently held"

A research based on the two expressions gives different results. Some 600 documents exist for French, around 250 for English. In order to avoid inferences due to a bad updating record, since the EN version is not complete, we need to take a common range of documents. Even in such a case the research outcome differs, mainly due to translation discrepancies. CELEX, not having exploited the possibilities of thesaurus construction, is not in a position to automatically create descriptor equivalences which would allow for parity of results independent of the language version; on the other hand, Community law is of such a complexity that the effort to create thesauri demands a quasi-interpretation of Community law concepts, entailing an unacceptable risk element.

At this point we should focus on another problem of our particular multilingual environment. The coexistence of different versions with equal binding force necessitates comparative examination by the lawyers in order to isolate the one which supports a particular point of view. Nevertheless Celex demonstrates a major deficiency at this level, not allowing for comparative examination of texts in the form of "horizontal hypertext navigation". What we would like to see defined is the possibility of matching the same entity onto different aspects. The contribution of such an application would have been substantial, given the paramount importance of an internal comparative method for the development of EEC law.

Further extensions using techniques of "frequently repeated extracts of text" applied to documents presenting similarities in structure and/or content are of considerable value in some domains. Leaving aside translators who may obtain a direct gain, legislative technique and judgment drafting have much to benefit from such a system - and not only in style. Inevitably, such an approach would involve a justifiable shift from the current definition of the word as the basic structural unit towards the less ambiguous level of phrase/expression.

A layer of statistical functions should be added to every "decent" legal information retrieval system since there exists a considerable amount of information that can be derived from existing data.

Undoubtedly hypertext extensions would have allowed for more efficient organisation. To a certain extent CELEX has been anticipating a hypertext application, since the references to other documents, by their structure and utility, resemble a similar solution within the limitations of traditional IR technology. The problem of course is that these references are not incorporated into the actual text; however, it's difficult to imagine hypertext nodes appearing in the middle of official documents that are intended for publication. Technically speaking, however, there exist ways of making hypertext links - meaningful only in a computerised version of the text - disappear when the document is transformed into typography format. And it remains a mystery why in this era of intense computerisation the paper form remains so important.

A valid question is who will be responsible for establishing the hypertext nodes. If we stick with the idea that the original text should not be submitted to meta-interpretations then definitely the person to "hypertextise" a document is the one who compiles it. In fact one could imagine an environment where legal secretaries compile their work directly on terminals, by querying CELEX and when necessary extracting and copying what they select, thus maintaining all the pre-established internal links.

It goes without saying that this task requires a specific editing environment. The question still is whether this should be incorporated within the IR environment, since users generally prefer to work at a local level. It seems therefore necessary to maintain a certain degree of independence at this level. It is also at this level that intelligent interfaces are invited to play a role. Maintaining the possibility of changing the working parameters means fully exploiting both the IR and local environments. Editing should not be limited to the files that are downloaded but should also include queries that fail to pass the consistency check of the IR system. The Court of Justice has already started implementing a working environment that offers solutions to this problem.

Such an intelligent interface can also be used for dynamic querying and linking. A navigation effect can be obtained by selecting terms that refer to document entities while dynamic redefinition of queries can be invoked by selecting displayed terms.

A reservation to be expressed at this point is that technically intensive applications tend to restrict creative writing, as authors will be influenced by mannerisms. Inconsiderate reproduction of existing case law might lead to fossilising expression and reasoning, since the user of such an application will be tempted to adapt the interpretation of the facts to an existing solution rather than developing an original one; such "standardisations" are usually justified by an effort to avoid redundancy.

In fact, the European judge was, in some cases, so charmed by the electronic medium that he was tempted to adapt his working method to the presumed requirements and nature of computers. It is along this line that Justice Pescatore, one of the most influential presidents of the Court of Justice, expressed the opinion that legal concepts should be uttered explicitly, univocally and in an homogenous manner. It is true that a certain degree of standardisation is necessary. But we should keep in mind that there are other methods that can permit an automated assimilation of concepts.

It is therefore the opposite point of view that must prevail. The text should not be compiled according to rules other than those which concern its physical formatting. Explicit references to the legal base is wished and feasible - in the community legal order we have the luxury of citing them unambiguously allowing thus for a standard way of uniformising research.

Our feeling is that CELEX should be redesigned in the role of a tool allowing the synthesis of computer-assisted research into a text, by mirroring the stages of a - not necessarily sequential - search.

Much the same applies for purely informative research where specific devices or routines would allow for extending traditional retrieval by a dynamic navigation permitting a trial and error approach and a constant redefinition of research criteria; if nothing else such an organisation would create an excellent - not precoordinated - training and learning tool.

In conclusion, it is obvious that, in order to permit a better assimilation of legal materials by lawyers, we should proceed to integration of different technologies. The basic form of such an integration should contain elements from DBMS, IR, Hypertext and text editing. With contemporary platforms e.g. Unix, integration does not necessarily mean concentration on a single level; solutions may then cover several services.

At the management level we need tools that will allow for global data transformation in order to upgrade their use, as well as in order to maintain data independence in case of change of environment. Editing needs to be an integral part of such an approach if we want to shift the use of IR into a real tool. It should cover both queries and extracting or modifying text. Some contemporary communication software allows for at least the mirroring of the session onto magnetic media. What we also need is an editing function allowing selective handling of extracted text, connected to automatic referencing. It would be preferable that editing be located at the local level simultaneously with the communication software in order to allow for a universal approach to different hosts. Editing of queries should be incorporated. Part of it could be dynamic querying, which would selectively invoke terms either as research terms or as hypertext nodes.

Hypertext, on the other hand, should allow for "historical" navigation but also for an horizontal linking - the possibility of maintaining the same localisation while moving through linguistic versions. The idea of dynamic linking seems indispensable for a system that is constantly evolving and needs to transcend manual intervention.

Finally, utilities that facilitate a synthetic approach (statistical functions, linguistic analysis) can make researchers conscious of the inner characteristics of a particular database and permit a finer definition of search strategy.

Conclusion

The initial perception of information retrieval systems as purely informative means prejudiced IR applications, its use and development, because it created a psychological tendency to consider them external to practical work.

A typical symptom of this approach is the survival of mid- users at an era when computer technology has become extremely widespread.

Enriching the functionalities of IR, or to be more precise, integrating neighbounng technologies, could allow for the creation of productive authoring systems permitting the "gleaning" of the necessary research elements, from the initial brainstorming up to the final synthetical compilation of legal reports.

Of course redefining the role of information retrieval and placing it within an environment where legal research and legal writing are contained in an integral whole is desirable but not easy. The importance of such an undertaking is, however, justified by practical benefits, not only at the level of Community Institutions. Practitioners have much to gain from such a method of work unifying in one step both the preparation and the synthesis. Academia finally must assume the responsibility of proposing technical solutions; and of course it may benefit from focalised training of young scholars.

---

[1] The opinions expressed in the present paper reflect uniquely the author's point of view; they do not voice policies of the EEC jnstitutions.