A pan-European service disseminating national case law: the Caselex project

SEBASTIANO FARO, ROBERTA NANNUCCI *

1. PUBLIC SECTOR INFORMATION (PSI)

Public governmental bodies in exercising their tasks and duties collect, create, store and disseminate huge quantities of information of any kind (financial and business information, legal and administrative information, geographical, traffic, tourist information, etc.). As a consequence, the public sector can be considered the biggest resource of raw material for the creation of value-added information content and services. This is very important for citizens, business firms and professionals who can greatly benefit from this type of information released on the Internet.

The European Community focused its attention on this kind of information called “public sector information” (PSI). A directive on PSI re-use has been passed in 20031 together with strategies and policies oriented to guarantee PSI accessibility and availability. Furthermore, these policies were some of the objects of the eContent programme2 which aimed at supporting the production, use and distribution of European digital content and at promoting linguistic and cultural diversity on the global networks.

* S. Faro and R. Nannucci are senior researchers at ITTIG. R. Nannucci has been the administrative Coordinator of the Caselex project.


2. Case Law as an Essential Source of PSI

Case law is an important source of PSI as output of the judiciary, together with information of governmental bodies, as output of the executive and legislation, as output of the legislative: they jointly constitute the so-called “basic information of the democratic constitutional State” and it is generally accepted that these categories of PSI should be more and more accessible and available as wide as possible. Furthermore, case law represents the most dynamic source of law in view of the fact that increasingly case law becomes law: although theoretically court decisions only affect parties, courts (in particular lower courts) have the tendency to follow decisions of higher courts, especially those of the Supreme Courts, thus creating a new legal reality. This is particularly true if applied to EU legislation considering the relevant role played by both the EU courts and the national courts relating to its implementation.

However, contrary to legislation and information of representing public bodies - most countries offer full coverage thereof via the Internet - the dissemination of case law (beyond the conflicting parties involved) has most often not been considered to be a public task and not received sufficient attention, in particular on global networks: traditionally, publishers have taken that role within a national context and perspective, and in markets which can be described as fragmented and oligopolistic from a European perspective. Therefore case law was risking to be unaccessed and unexploited. Europe at large is not being served, which can be considered a missing link for the complete eGovernment fuelled services chain in the open and democratic Europe.

This situation is awfully negative if it is considered that, with the evolving European integration, legal professionals, mainly working in

---


4 In the last decades, the legal publishing market has become an oligopoly with three dominant international players, i.e. Reed Elsevier, Thomson and Wolters Kluwer. Today it represents a national focused industry with significant entry barriers, dominated by the three actors plus an additional small number of publishers selling either identical or differentiated products in national markets.
courts or public administrations, next to serving as sole practitioners or in small and medium-sized law firms, demand European knowledge on legal solutions to a greater extent as the legal issues questioned by their clients are increasingly turning to European business. The growing importance of European legislation fuels this need even further.

The national oriented supply of case law did not meet the need and expectations of the legal community. Moreover, elements such as the independence of the judiciary, its limited resources, the lack of clear dissemination and communication policies and conservative stance towards new methods of work and technologies had in many countries left the important online case law dissemination entirely up to voluntary exercise of single judges or court employees, or to specialized legal publishers.

Furthermore, European law should in principle be applied transparently and equally across the Member States. Such a reality is hardly to be achieved without access to any European eGovernment service allowing judges and other legal professionals to seek inspiration and best practice solutions to the very same issues. Today they struggle with identical legal challenges, without the opportunity to learn if, and how, other foreign courts already have tackled these issues. As a consequence Europe is losing out on knowledge capturing and sharing, in the form of case solutions, between courts, between courts and other legal practitioners, and between legal practitioners themselves who argue the cases in front of the courts.

Besides the obstacles driven from the market and the insufficient activity of the courts in disseminating their cases, there is a number of practical barriers preventing players from entering the market and produce cross border services. These barriers relate to the complexity and inconsistent state of online case law in addition to the absence of truly interoperable data across Europe. Furthermore, the different European languages create a significant problem next to the rather traditional culture of national court organisations, lacking experience with open competitive public-private partnerships.

The Caselex project, financed by the European Commission under the e-Content Programme, during 2005 and 2006, contributed to bridge this gap.
3. THE CASELEX PROJECT

The Caselex Project, through a Consortium blending private and public sector parties, aimed at building a platform that could shift the public case law content from a disconnected group of national information sources to an integrated European knowledge service, offering access to national case law across European borders harmonising the different legal traditions and cultural backgrounds of European countries.

It overcame all practical barriers for establishing an online service able to support and link the vast repository of important case law of national and European courts with European relevance.

Caselex had four main objectives.

Firstly, it aimed to bring public case law resources from the state of national, public information sources to distributed European knowledge for legal professionals.

Secondly, it gave substantial added value to national case law. The simple collecting of case law from the various European countries would not have allowed end-users to take full advantage of its knowledge potential. Consequently, the Caselex service had a crucial role to play in adding context and meaning to the cases, also allowing navigation between cases. The legal community in Europe has no strong tradition in using foreign case law as source material in its professional life, despite recognising the value of cases related to a common denominator such as EU law. Consequently, Caselex had to put strong emphasis on the quality of the service and its content, and to be put in a legal context which makes sense to the user and allows the user to fully utilise the service offering.

5 The Project Consortium consisted of 10 partners belonging to several European countries. Some come from the scientific and research world such as the University of Tilburg (The Netherlands), CRID-FUNDP - University of Namur (Belgium), ITTIG - CNR (Italy). Some are publishers of legal databases such as First Law (Ireland), WordWave International (UK) and TIC - Ministry of Justice (Lithuania). Retriever/Schibsted (Norway) is engaged in sophisticated computer applications being a leading supplier of digital media intelligence services in Scandinavian countries and ZENC BV (The Netherlands) is a private firm expert in innovative counselling in the area of ICT and public policy. The other partners are case law content holders representing major European countries: Juridat - Cour de Cassation (Belgium), Direction des Journaux Officiels (France), Ministry of Justice, DGSIA (Italy).
The third objective was to offer an “inclusive” pilot system, not privileged or restricted by unfair pricing, contributing to access for all, allowing publishers and other intermediaries to exploit national case law.

Finally, Caselex aimed to build and launch a business model for Caselex as a viable European venture allowing courts as publishers to easily and cost-effectively disseminate court cases and judgments across Europe in a cost-effective manner; allowing publishers to exploit new content and market channels into new territories; allowing legal professionals to acquire and render better knowledge and advice.

As a result, Caselex has created a system offering an Internet based “one-stop-shop” service for important national case law linked to the common denominator of EU law, i.e. decisions of supreme and high courts in all the EU member States plus the EFTA States involved in the EEA, also holding cases from the European and EFTA courts, within selected areas of law connected to the implementation and application of EU law.

The service pays specific attention to the important judgments that can inspire cross-border legal solutions and ultimately be used as authoritative source for solutions in courts outside its respective country. In order to be of such value, Caselex focuses on case law that:
- is introducing a new principle or rule;
- is materially modifying an existing principle or rule;
- is settling, or materially tend to settle, questions upon which the law is doubtful;
- is for any other reason particularly instructive.

Caselex offers the case full text in the native language, with headnote, summary and additional metadata in English. Reading the headnote and summary, the user can make up a qualified opinion about the substance of the case, and if of interest, order a full text translation on demand into English, French, German, Italian and Spanish.

The figure below shows Caselex stakeholders, including target groups. The end-to-end chain of the Caselex service encompasses content producers and holders, intermediaries such as publishers in the public and private sectors, and the end-users. Caselex interacts with public content holders for raw data and with private and public content holders for raw data and value added data. Furthermore, the publishers can also support the distribution of the Caselex service, which will be accessible also directly to end-users.
4. BUILDING THE SYSTEM

To reach these goals specific activities and tasks were planned and carried out by the Consortium. Five main activities are here briefly described.

4.1. Understanding the needs and potential of the market

Some surveys were executed for understanding the needs and potential of the market with the aim of seeking a deeper understanding of the international market place, particularly with respect to content holder, judiciary and industry scenarios, relating the whole of Europe.

*Content holder survey* to detail access to, and potential of re-using, case law coming from national supreme courts, high courts and other superior courts and national competition authorities. This survey aimed at understanding what is available and against which conditions, looking at all judicial systems across the entire EU plus the EFTA zones.

*Judiciary survey* to assess the knowledge production and use of the judges, and the potential of engineering backward integration into the content generation processes and rendering live, real-time court content.

*Industry survey* to assess the specific needs of the international scene and to test the concrete case of distributing Caselex content at European level.
4.2. Assessing the technological and legal barriers

Prior to convert the research findings into definition of the content, functional and technical specifications of Caselex, additional work was to be invested into assessing in detail the technological and legal barriers and challenges the project would encounter. So, a thorough understanding of key technical issues and how they potentially might become barriers to entering the market were addressed.

The data management process performed by Caselex service platform can be synthesized according to the interaction both with content holders and end-users. In particular the scheme of Caselex data management process can be represented by three main phases: data acquisition, processing and access.

Specific attention was given to legal and technical barriers related to main issues involved in each of these three phases, in order to obtain an exhaustive assessment (e.g. on issues such as interoperability for re-use, intellectual property rights, data protection, digital rights management, languages and customization)\textsuperscript{6}.

A detailed set of requirements and specifications for the Caselex service at macro and micro level came up.

4.3. Creating the content repository

Having gained a deep insight into the market for case law, including the barriers to entry, and consequently specified the Caselex solution, another activity was to cover the important work of providing the content foundation for the service and catering for the editorial process bringing added value to cases.

In order to feed properly the content into the repository, the elaboration of detailed editorial guidelines was provided for guiding the editorial staff in their task. Furthermore, great emphasis was placed on developing and employing suitable metadata, including a pan-European thesaurus, as “added value” is crucial to a well-functioning service.

\textsuperscript{6} The results are described in a more detailed way in M.A. Biasiotti, Re-using case law across Europe: legal barriers to be faced, in “Informatica e Diritto”, no. 2, 2006.
4.3.1. The metadata set specifically elaborated for Caselex

Since the simple full-text of a decision was not considered sufficient to provide a high quality service, it was considered essential for Caselex to enrich the full texts of selected cases with an articulated set of further information, generally called “metadata”. Getting the right metadata set in place required four steps.

- Finding the right document type format, which is the result of the merging between the full-text and properly edited metadata, stored in XML format.
- Analyzing existing metadata schemas such as the Dublin Core metadata initiative\(^7\), eGMS (e-Government Metadata Standard)\(^8\), Justice Sector Metadata Standard\(^9\), Italgiure Web metadata, EUR-Lex case law metadata\(^10\).
- Drafting the proper Caselex metadata set, being aware that it must be original and unique, even if taking account the best of the various examined sets.
- Implementing the metadata set elements, some of which are automatically generated by the system, while the other are to be edited manually by the editorial staff according to the editorial guidelines.

The Caselex metadata set elements were subdivided into four groups according to their functions:

- **Content**: metadata relevant to the case content, extracted or inferred by analysing the case.
- **Cross-referencing**: metadata relevant to cited or citing acts connected with the case.
- **Context**: added information providing tools to improve the comprehension of the case.
- **Administration**: information concerning the Caselex documents management.

For each defined element data provided were: *name* (the unique token assigned to the element), *label* (the human-readable label assigned to the element), *definition* (a statement representing the concept and essential nature of the element) and *content* (value assigned to the element). Where

\(^7\) http://dublincore.org/.
useful, the following additional information was also provided: notes (additional information useful in the application of the element); creation (information relating to the data creation methodology and/or to specific skills required for the metadata editor); example (used both for suggesting lists of standards or controlled values usable for editing the element content and for identifying the information useful for editing the element in the raw documents).

Among metadata elaborated for Caselex the most suitable for accessing cases are the headnote, the summary, the EU core provision and the national core provision.

The headnote is the answer to the question the judge has been requested to deal with. It refers to the essence of the case which is relevant for Caselex and deals with the European law component in relation to the national issue that has been addressed by the court.

The summary contains a logical reasoned precis on the basis whereof, how and what the court has decided. It also contains the result of the case and (where stated) any relief given. After reading the summary, the user should have a full understanding of the legal substance of the case, however without knowing all factual details. A typical Caselex summary is between 250-400 words. It does not represent a legal analysis of the case but rather reflects key facts, holding, decision and results.

The EU core provision is the EU rule that the court had to interpret, based on the arguments brought forward by the parties.

The national core provision is the provision through which the EU core provision has been implemented within the national legislation of the member State (whenever the EU Core provision is part of a directive).

4.3.2. The Caselex multilingual thesaurus

Starting from the fact that the area of case law of interest for Caselex is currently lacking of a common thesaurus and that only proprietary versions do exist, not available to the legal community and the data interchange, Caselex built a multilevel, hierarchical thesaurus for multilingual application by the service itself, the courts, and the publishers. It covered eight languages, being English, Danish, Dutch, French, Italian, Lithuanian, Spanish and German.
Not prepared to invent the wheel twice, at first current solutions have been considered (Eurovoc, Lois produced by Ittig; Glin.gov; Eclas, the system of the EC library; Pica, the Tilburg University system in Dutch, French and English; Jurivoc, in Italian, German and French; Gemet, in all EU official languages; Juridat, le thesaurus juridique belge, in French and Dutch).

In order to build the first layer of the thesaurus two activities were carried out; firstly, identifying significant descriptors for each area of law relevant to Caselex and then giving the hierarchical structure and organization to these descriptors.

For each area of law an essential number of descriptors representing legal major concepts involved was identified. After, a comparative view of existing descriptors in other thesauri was conducted. Once the list of each area of law (“sector” in the thesaurus organizational table) had been defined, the hierarchical structure was applied, classifying the most general descriptors as “top terms” (TT) of the hierarchy and the most specific ones linked to each top term as “narrower terms” (NT); these could be organized into two different levels (NT1 and NT2). An example is provided in the following table 1.

<table>
<thead>
<tr>
<th>From the list of the descriptors ...</th>
<th>... to the thesaurus</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Area of law</strong>: Intellectual Property Right</td>
<td>= <strong>Sector</strong>: Intellectual Property</td>
</tr>
<tr>
<td>General concept relating to this area: Industrial property</td>
<td>= <strong>TT</strong>: Industrial property</td>
</tr>
<tr>
<td>Specific concept related to Industrial property: Trademark</td>
<td>= <strong>NT1</strong>: Trademark</td>
</tr>
<tr>
<td>More specific concept related to Industrial property and Trademark: Domain name</td>
<td>= <strong>NT2</strong>: Domain name</td>
</tr>
</tbody>
</table>

Table 1 - Example thesaurus build up

Afterwards the second layer of the thesaurus, the multilinguality, was considered. The starting list of descriptors came from the English language, so it was necessary to adapt it to the Caselex multilingual needs, i.e. rendering each descriptor into the other languages relevant to Caselex without losing the legal significance of the concept. Reusing multilingual official sources available at EU level (such as Eurovoc, Iate or other EC specific glossaries) allowed the performing of this task.
Looking at the future, Caselex is fully aware of chances offered by the semantic Web approach. Addressing the next frontier of information retrieval, the Caselex thesaurus could be enhanced in two directions: (i) identifying additional descriptor sources and (ii) enriching the knowledge structure management through conceptual relations semantically specialized and formally defined.

4.3.3. Feeding the content into the repository

Finally, a repository of cases was built, covering over a thousand of cases from 18 Member States and from 1 EFTA country at the end of the project (December 2006). Cases refer to a period starting from 2000 up to 2006. Furthermore, some other databases were also built, containing cases from the European courts and the EFTA courts, on the base of a selection connected with the Caselex core areas of law, and provided with the metadata set created for the repository. The cases gathered in the system cover four areas of law: intellectual property law, labor law, competition law and company law. In some countries other four areas of law are covered, being: environmental law, consumer protection law, ICT law and banking law.

In order to link decisions from different European countries dealing with the same EC rule a content map was also implemented, allowing a user in country A to check how a judge in country B interpreted a certain article in a specific EC directive. The Caselex content map is the most important content mechanism in the system: it links national law applied in cases across national boundaries through community law. Using the content map, a Caselex user is able to see whether there are any foreign cases on a statutory article in her/his national law. Consequently the user does not need to have legal expertise of foreign countries in order to find cases of importance to their own work.

4.4. Implementing the service

After having analyzed the available off-the-shelf technologies the prototype was developed along a modular approach. Caselex chose open source solutions as the base operating system for its applications, based on the following facts: free of charge, lowest total cost of ownership;
documented, open source, to be modified for special needs; proven stability; well organized community for bug corrections, tweaks and twists; best “performance/hardware cost” ratio; linear scalability. The exception regarding free software in the Caselex application is the choice of a “free text search engine”. Although some products are emerging – like Lucene – there are only a limited number of commercial platforms offering sufficiently advanced functionalities. Caselex utilized Fast Search and Transfer\textsuperscript{11} based on proven performance and scalability.

4.5. Disseminating and exploiting the service

In order to prepare the foundation for a sustainable pan-European venture a viable business plan was established supported by some marketing activities strengthening brand building and take-up of the service; they are:

- Business plan with roll-out;
- Corporate identity and presentation material: corporate identity; brochure; Web presentation;
- Partnering: creating partner programs; networking with potential customers and venture capital;
- Promotional activities: speeches and presentations at public events and expos; public relation through mass media; direct mail to publishers.

The business plan with roll-out demonstrates the business case of Caselex; its target audience is legal professionals, with the highest commercial value coming from barristers and solicitors, in-house legal counsel, consultants and paralegals, and policy makers. Judges are considered to be extremely important reference clients, but without a revenue potential like the other groups. Despite this fact the group is expected to be one Caselex’ largest user groups in terms of numbers. Scholars and students are also considered potential users of the service, adopting an appropriate price strategy.

5. The Caselex service

The present version of the Caselex system\textsuperscript{12} is described from two different points of view, system design and user experience.

\textsuperscript{11} http://www.fastsearch.com.
\textsuperscript{12} www.caselex.eu.
5.1. System design

The Caselex technical platform has been developed along a modular approach to easily allow for scalability, extensibility and interoperability. Figure 2 shows the key production and end-user modules.

The **harvester module** is responsible for gathering content from the various case law content holders in private and public sectors, and, if needed, converts the content to Caselex native format (Caselex' own XML format – CaseleXML). The harvester is also responsible for quality assurance of the content and the adding of some metadata. Finally, it injects the received content, if well formed, into the Caselex database.

The **storage and retrieval layer** includes a set of components for storing, modifying and retrieving data necessary to run the Caselex service: relational data base management system, text retrieval engine and access control and authentication. To support data storage and information retrieval, interoperable data formats are essential and consist of the representation of case law (CaseleXML) and the handling of cross-references (Caselex URN standard).

As an international standard for representing case law in machine readable form does not exist, the representation of the full text of the case and the associated metadata has been developed in the Caselex system in order to:
- serve as a preferred format when receiving data from content holders;
- delivering search results;
- exporting/syndicating data to partners;
- transferring data between various Caselex modules.

In order to facilitate the maintenance of all links, a Caselex URN (Uniform Resource Name) scheme has been developed\(^\text{13}\). As to the editor and end-user module, a separate web interface has been produced for editors, including country correspondents, legal experts and translators involved in the value added creation. The editorial user interface contains different functionalities enabling different types of value added (adding/deleting/editing metadata records, translations of headnotes and summaries), content management, handling of translation orders, subscriber management, etc. The end user interface and the editorial interface are built on the same principles: the main difference is of course that in the editor’s case it is possible to store as well as retrieve information. In addition, the editorial user interface is more advanced in terms of functionality since the users of this interface can be considered professional. Also the graphical basis is the same for both the editorial and the user interface.

Finally, Caselex has built a log and statistics module allowing management statistics, usage and service analysis, and any information needed for clients management.

5.2. User experience

The Caselex users find case law relevant to their work or study, by using foreign cases to identify solutions to national legal issues or finding solutions to issues outside their country. Browsing can be done in different ways and can also be combined with free text and advanced searching. Results can be stored in personal archives, with personal notes to be added to the cases and customised alerts on new cases of specific interest, to be received by e-mail. Retrieved cases include multilingual headnotes and

---

\(^{13}\) An URN is a kind of unique identifier for a resource, independent from its location. When storing a reference from a document, the URN of the referenced document will be stored together with the document rather than the location of the referenced document. When a document changes its location, the new URL has only to be entered once in the Caselex database (the URN register).
summaries added by the Caselex editorial staff, next to multilingual classification and additional selected metadata. The full text of the case which is in the language of the source court can be translated on demand manually following an online order. Each user can set up “My Caselex”, a personal homepage where he/she easily can see the latest news of interest and get easy access to own knowledge stored within Caselex.

There are a number of ways to get to the relevant cases. When starting from scratch, two main accesses are possible: browsing and searching.

Browsing can be performed through:

**EU legislation.** All cases within the Caselex repository are connected to one or more core EU provisions. Other referred EU provisions are also added. So selecting one or more EU provisions will list the cases which have this/these provisions added as metadata.

**National legislation.** The core national provision is the antipode to core EU provisions and is the provision which implements or otherwise equals the EU core provisions. These connections are entered into the content map. In most cases the national judge starts with this national provision and then tracks it back to the EU provision. As well as for EU provisions, also other referred national provisions are added. Selecting one or more national provisions will list the cases which have this/these provisions added as metadata.

**Legal area.** All cases are related to one or more legal areas, for instance “competition law” or “labour law”. The end user can select one or more legal areas.

**Country.** The end user can select case law connected to one or more countries from which the cases originate.

**Thesaurus/subjects.** All cases have one or more subjects from the Caselex thesaurus added as metadata. The end user can browse through these subjects and select one or more of them and the result list will display the cases with these subjects added to them.

Searching may be subdivided into multilingual search and monolingual search.

**Multilingual search.** All cases have summaries in English. In addition, cases are classified via subjects from the Caselex thesaurus. The thesaurus is multilingual; this means that cases in the languages covered by the thesaurus may be retrieved in the end user native language. As mentioned,
cases may also be searched via EU provisions which are not dependent of any language.

**Monolingual search.** When the user is making a search for criteria not found in the metadata or in the summary or does not know any case details or links to EU provisions, then the user may retrieve documents searching in the full text content of the case, available only in the original language, unless a translation has been ordered.

6. CONCLUSIONS

Caselex has managed to organize a content production process which now is ready to be scaled for 31 European jurisdictions (EU and EFTA countries), establishing a network, an infrastructure and a content repository allowing the European legal community to access and search national case law of European interest in a fully integrated way.

In view of its objectives and long-run goals, Caselex will be further financed (2007-08) by the EC under the e-TEN program to consolidate its availability to satisfy the needs and requirements of interested stakeholders all over Europe. Furthermore, since the service is citizens-oriented the attention will continuously be given to the updating and enlarging of semantic tools able to enhance a real and widespread accessibility of legal information, as suggested by the PSI directive.