# Managing Knowledge Extraction and Retrieval from Multimedia Contents: a Case Study in Judicial Domain

Elisabetta Fersini[1] , Mauro Cislaghi[2], Roberto Mazzilli[2], Fabrizio Callegaro[2], Stefano Somaschini[2], Roberto Muscillo[3], Domenico Pellegrini[3]

[1] Consorzio Milano Ricerche,
Via Cozzi, 53 - Edificio U5, 20125 Milano, Italy
fersini@disco.unimib.it

[2] Project Automation S.p.A.,
Viale Elvezia 42, 20052, Monza, Italy
{Fabrizio.Callegaro, Mauro.Cislaghi, Roberto.Mazzilli, Stefano.Somaschini}@p-a.it

[3] Italian Ministry of Justice,
Via Crescenzio 17/C, 00193, Roma, Italy
{Roberto.Muscillo, Domenico.Pellegrini}@giustizia.it

**Abstract.** In this paper we present the main challenges and opportunities in exploiting the knowledge embedded into multimedia judicial folders in criminal trials and their influence on the courtroom infrastructure. The paper describes the results of a one year analysis conducted in the Italian and Polish Courtrooms and how to face them in order to make this knowledge available to judicial operators, focusing on the criminal cases during the trials phase.

**Keywords:** digital libraries, audio processing, video processing, e-justice.

## 1 Introduction

The available information technologies have significantly expanded quantity and type of information obtained by End Users.

In addition to a virtually unlimited supply of unstructured text content, there is the opportunity to access an even broader set of static multimedia content, typically images, and dynamic digital sound recordings and audio-video. This raises the problem of indexing this content, to enable an efficient and effective information recovery by the judicial operators. In parallel to the traditional ways of indexing multimedia contents, providing manual addition of metadata to catalogue and recover the contents, in the last years techniques for automation knowledge extraction made relevant progresses. These technologies are able to extract automatically information from textual and multimedia folders and link this information to concepts directly used by the user, such as the recognition of scene changes in a video or time in a audio recording, identification of an object in an image or a video or a concept in a

text, identifying the emotional state of a person. This article analyzes the possibilities offered by the use of techniques for automatic knowledge extraction from multimedia contents, with specific application to the legal domain and criminal trials, illustrating on the technical point of view benefits and limitations of using these technologies. The specific application context, the criminal trial, has been developed in the JUMAS project, co-funded by the European Commission in the context of the ICT program 7th Framework program.

## 2 Content Management in the Judicial Domain

The ICT systems in use in the judicial domain [16] during trials can be divided into two major categories: the Case Management Systems (CMS), for the management of administrative and procedural information of the legal proceedings, and the electronic Court Systems (eCS) that support the execution of debate hearings and allow the management of the documentation produced during the trials. A CMS is a transactional system designed to manage into a relational database the events and the data provided by the Criminal Procedure and needed by the users involved in the management of the legal proceedings. The information inserted into a CMS follows the "life cycle" of the criminal case, such as case opening, investigation phase, pre-debate phase, debate and post trial enforcement and surveillance. A CMS may be considered as the storage of the static data related to the criminal trials.

Examples of the information processed in a CMS during a trial are: general information of the trial (registration number, type of trial, kind of criminal act, current status, list of the defendants with their lawyers, names of judges involved, etc.), legal acts to imprison the defendants or against the properties, requests by the defendant to review the acts, the separation or the combination of two or more proceedings, the outcome of the case: the sentence.

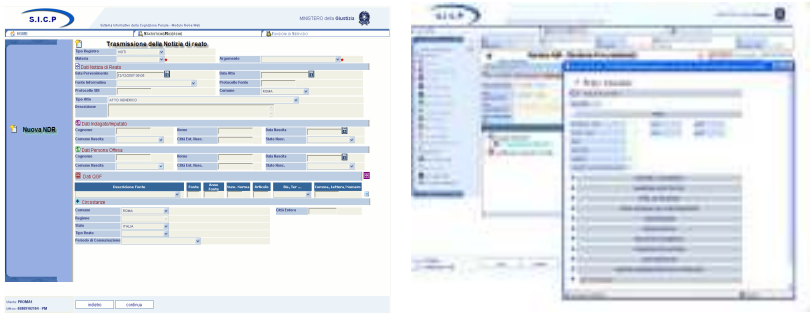


**Fig. 1.** Example of case management system from the Italian Ministry of Justice: the SICP system.

CMS can also include simple content, mostly textual, management functionalities, such as the storage of the hearing transcription or the assisted generation of trail documentation (e.g. the hearing report). Considering the complexity of the acquisition

and consultation process, the complexity of the infrastructure and jointly the importance of the documentation acquired during the legal proceedings, the most adopted solution is to use dedicated document information system, the eCS, integrated with the CMS, and dedicated to the management of the digital trial folder, including the access management to third parts (e.g. Lawyers).

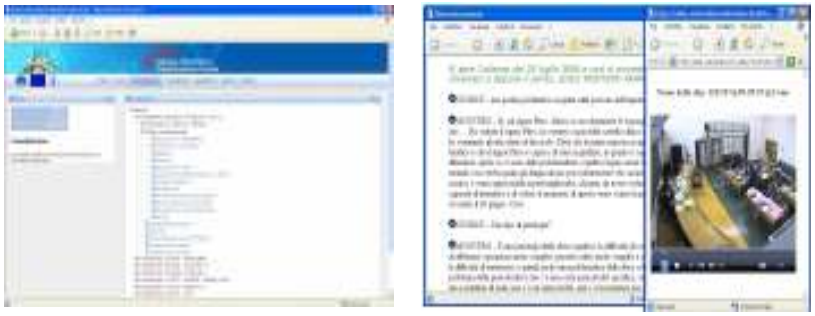

**Fig. 2.** Example of a court management system from the Italian Ministry of Justice: the SIDIP system.

Information managed by the CMS and the ECS are both an essential reference to ensure compliance with the provisions of the Code of Criminal Procedure and the main information source on which Public Prosecutor and attorneys build their arguments.

The complete set of data stored in the transactional database of CMS, but mainly the data managed by the eCS regarding the not structured contents (the documents), constitute the reference information source for the end users (prosecutors, lawyers, judges). During his/her legal activities, Prosecutors and lawyers use CMS and eCS to consult the contents and also to:

- obtain knowledge from textual contents, such as the extraction from the transcriptions of lists of names, places, dates;
- classify and search the textual contents based on a particular personal classification, such as collection all the specific instants of a defendant deposition to use them in their thesis;
- search and consult audio-video recordings of a hearing or part of them, synchronized with the corresponding textual transcription to review a particular deposition or other phases of the trial;
- produce audio-video-text highlights of a trial to support their legal activities.

The generation of the judicial decision is therefore based, both for judges, prosecutors and lawyers, on the detailed review of all trial documentation, including evidences gathered during the investigations and showed in the courtroom, becoming fundamentally to the objectivity of the conclusions and the economy of work. These activities require a significant manual work to allow the users to recovery, extract, reorganize and link the contents useful for their judicial activities. Audio-video

contents are accessible only sequentially, while the text documents acquired during the various judicial phases are organized in a chronological structure.

An influencing element in many judicial systems is the time latency in the availability of the information, directly consequence of the average duration of a criminal case. In Italy, the average [1] time from investigation beginning and the corresponding debate start is 1000 days; to this it has to be added debate duration, averagely 300 days for monocratic (with one judge) and 560 for collegial (with more judges) trials.

Every year about 550,000 manual transcriptions are produced, for a total of over 6.5 million pages, and a yearly cost of more than 20 million Euro. This do not include the costs for waiting the transcription, difficult to quantify.

A high percentage of cases are sent to Appeal and Cassation Courts ($2^{nd}$ and $3^{rd}$ level), where trials are carried out on the basis of all documents produced in the previous judgment level(s).

These data give evidence of the opportunity to provide advanced tools to categorize and retrieve efficiently the amount of rich audio-video-text, considering both the innate heterogeneity that the required time availability.

## 3. Automatic Management of Multimedia Contents in the Judicial Domain

The introduction of automatic knowledge extraction, considering especially the integration with the eCS, can give significant benefits for a full and efficient use of contents, especially unstructured audio-video-text, collected during the entire cycle of the judicial action.

The technologies currently available for automatic indexing of multimedia content are particularly focused on the extraction both syntactic and semantic information. The knowledge obtainable from the judicial multimedia contents can usefully be organized in a hierarchy of abstraction levels [2]. In case of video content, the basic physical properties of the video itself constitute the first level of information; the next level is the identification of the elements that compose the scene; the concepts related to the interpretation of the scene constitute an higher abstraction level.

This modelling abstraction allows to classify the information content of a video into syntactic, generally coincident with the information related to the physical characteristics of images such as colour brightness or the corresponding event, and semantics, consisting of property related concepts such as the place where the scene takes place, the name of person / object present in the video, or the mood contained in the scene.

Recovery of multimedia content on the basis of semantic information is, compared with the syntactic level, closer to user needs. For example in a criminal trial the questions "shows the next intervention of the lawyer of the civil part" or "I want to hear again the testimony of Mr. Rossi" are more immediate and easier to interpret than "showing scenes that begin with the time t0 and t1" or "retrieves documents where there is the word Rossi". According to the previous approach is therefore possible to structure into levels the knowledge extracted from multimedia content,

with the specific objective of providing search modalities as close as possible to the users domain knowledge and efficient and automated methods of indexing and automatic extraction of the huge corpus of audio-video-text collected during the investigations and the trial hearings.

Analysis of the applicability of specific technologies of indexing and extraction of knowledge from audio, video and text contents in criminal justice domain follows.

### 3.1 Audio Recordings Processing

Automatic Speech Recognition (ASR) technologies actually available allow to extract many relevant information embedded in the audio recordings. It is possible to classify parts of the audio recording, for example by identifying the interval with and without speech; to identify the speaker and the start and end moment of the respective speeches; identify specific words (Keyword Spotting); generate the textual transcription from audio recordings (ASR), in similar way as it is done with the techniques for automatic character recognition (Optical Character Recognition, OCR); to identify the emotional state (neutral, nervous, scared, angry, sad, annoyed, etc.) of one or more speakers.

Considering the casual nature of the speech and of the context where it is allocated, it is impossible to derive a deterministic formula capable to create a link between the acoustic signal of the speech and the related sequence of associated words. Statistical-probabilistic formulations are therefore used, capable to model the problem of speech recognition on a non-deterministic point of view. In this modelling, the acoustic observations derived from the speech audio signal allow to estimate the probability that a specific sequence of words has been pronounced. In ASR state of the art, the more used modelling approach is a combination of two probabilistic models:

- an acoustic model, capable to represent the knowledge base of the system about phonetics, the pronounce variability, the time dynamics (co-utterance), etc.;
- a language model, capable to represent the knowledge base of the system on the concatenation and on the word sequences.

While the acoustic model has the main goal to estimate the probability that a given word is pronounced, on the basis of the acoustic observations, the language model has the objective to estimate in advance the probability of a words sequence.

In both modelling, Markovian models [5], [6], [7] represent one of the more used solutions. In particular, the acoustic model is composed by a concatenation of Hidden Markovian Models (HMMs) for each phoneme, while the language model assumes that a word sequence follows a markovian process of order n-1 (n-Grams, e.g. the probability to have a given word depends on the previous n-1).

Audio recordings of hearing in criminal trials and during interrogations, but also phone interceptions, are one of the mains sources where to extract automatically knowledge finalised to contents indexing, including those on video and text format, and to factual knowledge collection, useful in the decision process.

Automation generation of verbatim transcription starting from the digital audio recording of an interrogation or of a hearing opens scenarios of great interest. First, it has to be considered the possibility to supply, in a time frame almost proportional to the one of the duration of the audio recording, a transcription substantially equivalent to the one generated by transcribers. ASR algorithms provide also information useful for synchronising audio and text, so making possible to consult the text and hear the corresponding speech or to join the visualisation of an audio-video clip with the related subtitles ("Closed Captioned").

The possibility to perform direct search on the audio recording using keywords constitute a second relevant advantage. Thanks to the availability of the verbatim transcription synchronised word by word with the corresponding audio recording, the user has the possibility to search for terms or phrases in the ASR generated transcription, so activating the corresponding audio or audio-video clips playing. In case of application on vertical domain knowledge, such as the juridical one, the effectiveness of the queries can be achieved through query expansion through dictionaries (thesaurus) or ontology of the words used for the research on automatically generated texts.

Techniques for automatic audio processing thus allow to index the huge knowledge embedded into the audio and audio-video recordings with costs that are significantly reduced compared with the ones needed by other methods (manual transcription, stenotype o re-speaking technique).

**Fig. 3.** Example of consultation of a transcription synchronised with the corresponding audio recording

In criminal justice, and more in general in other domains such as journalism or security, the applicability of the above described audio processing methods is affected mainly by the quality of the available audio recordings. Many factors may negatively influence the performance of ASR systems: presence of multiple speakers and cross talking in the microphones, reverberation introducing audio signal distortion, the

presence of background noise or whispering, the heterogeneity of lexica and language, including the usage of words not belonging to the common language (person names, acronyms and abbreviations, technical terms and dialectal expressions), the spontaneous speech characterised by uttering, hesitations, false starts, shouts, data compression and signal loss during the audio acquisition.

Quality of interrogations and hearing recordings, can be significantly improved in relationship with the subsequent automatic processing by ASR systems. It is possible to dedicate separate recording tracks to the different speakers, to manage the activation/deactivation of microphones, to use directional or even personal microphones, to enhance the courtroom acoustics with architectural interventions and by digitally processing the collected audio signals, to use sampling and coding systems with reduced loss of information (lossless codecs), train and tune ASR language and acoustic models with in-domain lexica and specific inflexions, thanks to the wide library of available transcriptions and recordings in criminal justice. The main challenge is to find the right balance between costs, recorded audio file dimension and ASR performances.

### 3.2 Video Recording Processing

Automatic video processing offers, in analogy with audio processing, important advantages in terms of reducing the complexity of interpretation, search and retrieval of multimedia contents. Application fields of today available techniques for video processing range from automatic identification (detection) of scene elements (such as objects or persons) ed their movements (tracking) to automatic extraction of features (colour, brightness) referred to the scene or to portion of it to be used to extract properties of the videoclip (such as detection of abandoned objects in case of security applications), to the identification of specific behaviours (referred to persons or groups) or situations (for example fights), to automation generation of summaries (storyboards) of long video sequences.

The first studies to extract high level semantics from multimedia data were focused on manual text annotations. These methods were extremely expensive in terms of human effort. More recent approaches to the problem introduce the automatic semantic annotations [8], [9], [10].

Different sequential steps are required in processing video recordings to create a link between low level information coming from cameras and high level concepts useful for the end users.

Video contents are initially split into “atomic units”, called frames, eventually filtered from the noise generated by the acquisition equipment. The next step consists in an activity known as video time segmentation (SBD *Shot Boundary Detection*), that identifies the transitions between continuous sequences of sequential frames acquired from a single camera (shot). SBD activity finalised to shot boundary identification is performed considering the length of the time interval, the identification and classification of transitions between frames, the identification of scene changes, based on colour histograms, on brightness changes, on contrast and different intensity of pixels.

At the end of the video segmentation, a further analysis is linked to the semantic annotation of concepts.

According to knowledge acquisition and representation processes, *semantic annotation* may be distinguished into *implicit*, implemented through machine learning techniques, and *explicit*, implemented through model-based approaches. In case of implicit annotation, the associations between "low level features" and "semantic concepts" are learnt automatically through Neural Networks, Hidden Markov Models [11], Bayesian Networks, Support Vector Machines [12] and Genetic Algorithm. In case on explicit annotation, a priori knowledge (in terms of facts, models, rules) capable to provide a semantic model coherent with the domain is used.

The low cost and relatively simple management are increasing the number of cameras also in the courtrooms. Availability of witnesses and defendants hearing recordings enriches with important relational elements the otherwise anonymous textual transcriptions. Movements of face, eyes, body or hands that accompanies the evidences of a witness or a defendant are information elements of great importance, not available in transcriptions or audio recordings.

**Fig. 4.** Example of algorithm for identification of the faces of participants to a hearing.

Video recording may also be seen as one of the starting points for searching and navigating into the knowledge collected during the hearing. The usage of algorithm for the automatic generation of the summary (storyboard) of a hearing makes automatically available the chronological index of the speeches (for example it is possible to request the consultation of the forth speech of the judge). The possibility to identify faces allows to lay automatically on the video the label with the names of each participant to the hearing for an immediate contextualisation of the events. From video recordings is also possible to go directly to the verbatim transcription for example in the form of video subtitles (Closed Caption Text).

Automatic extraction of information from video recordings, jointly with the audio recordings, transforms without additional effort of the end users the video of a criminal and civil trial hearing into a navigation tool to gain access not only to the verbatim transcription, but also to other types of available contents (hearing report, documents submitted during the hearing, etc.).

The effectiveness of processing technologies and their consequent applicability is affected by the quality of the video recordings generated into the courtroom. Problems they may limit the performances of video recordings processing are mostly connected to insufficient courtroom lighting, to the usage of analogic cameras, to camera positioning not compatible with the automatic video processing (for example shots

containing parts of the courtroom useless for search and navigation purposes), to analogic recording systems (VHS) that compromise recordings quality.

As for audio recordings, it is possible to effectively intervene to enhance the quality of the video shots in connection with their automatic post-processing at acceptable costs. A practically costless operation is to arrange shots having the hearing actors in foreground, in particular considering the great usefulness of connecting the hearing of the depositions with the observation of the movements of the body and the hands. A correct lightning and a correct static camera with PAL resolution or at least 640x480 pixels and a minimum sampling frequency of 800 kb/sec will provide a not compressed signal that will result as adequate for being processed by the algorithm of automatic extraction of knowledge.

### 3.3 Managing textual information

The software available for automatic text indexing have been the enabling technology for the global diffusion of the Internet. Search engines make constantly accessible to the users, through simple keyword-based queries, the creation of new knowledge. The application of Information Retrieval technologies is not limited to indexing, search and retrieval of text, but it extends to the wider knowledge management. These technologies are fundamental, for example, for automatic document classification in thematic portals, for extracting valuable structured knowledge from unstructured data (template filling) and for automatic entity and relationships extraction in specific domain (e.g. to highlight obvious or embedded links between facts).

An Information Retrieval (IR) system can be viewed as composed by three main core elements: representation of texts, representation of search queries (user's query) and retrieval function stated according to a specific notion of relevance. The approaches for realizing a retrieval model are basically of three types:

(1) keyword-based, aimed at modelling the relevance of a document (with regard to a query) according to Boolean conditions;
(2) statistical-probabilistic, which makes use of clustering techniques, frequency analysis and conditional probabilities;
(3) semantic-based, aimed at using linguistic and conceptual representations, such as dictionaries, ontology and semantic networks.

Alternatives approaches are represented by hybrid-techniques, aimed at combining "semantic-based" [13] and "keyword-based" [14] retrieval methods joint with automatic relevance feedback and reinforcement learning models for automatically infer future user behaviour.

The IR systems play a crucial role within legal domains, especially concerning the huge amount of data available in a textual form derived from the digital version of paper proceedings and judicial actions.

Important applications of Information retrieval technologies are related not only to traditional search and retrieval of textual documents, but also to the automatic extraction of structured information from a non-structured form and to the automatic

text classification. These kinds of functionality represent an important opportunity for improving the efficiency of accessing and sharing penal juridical knowledge.

The automatic extraction of information from a non-structured source - records and documents of criminal proceedings as evidences, reports of inspections, and even audio/video clips - to structured records allow judicial actors to extremely simplify the filling process of juridical data bases.

For fighting the organized crime, shared hierarchies of data bases (Local, National and Trans-national) have been exploited for supporting investigation activities and in particular for simplifying the identification of connections between different criminal facts and for ensuring the completeness and the availability of information in real time. For this purpose each magistrate have to insert facts of interest about the investigative phase – such as inquired subjects, the relationship between theme, the mentioned criminal facts, etc. - coming from its assigned proceedings.

The filling of these data bases is obtained through a complex and laborious procedure of analysis that, starting from the legal proceedings through manual document categorization and concept identification, inside maps non-structured contents to structured (predefined) database entries. The procedure of extraction of information is carried out according to a specific and formalized methodology that consists of several phases: (1) reading and understanding of contents within textual documents, (2) concept identification, (3) database filling, (4) resolution of eventual ambiguous entries, (5) consistency verification of new information inserted with those already present in the original data base. Once inserted the informative contents in a structure form, the consultation functionalities are available to magistrates through textual queries and evocative representations.

A further area of interest related to the judicial domain is represented by the automatic classification of textual documents. This possibility is extremely useful in case of investigative proceedings or judicial hearing characterized by a huge amount of documents. Also in this case the proceedings are currently manually managed, implying then a high-cost of human effort by magistrates and court clerks.

There are no limitations about the applicability of IR technologies to the judicial domain, also taking into account the current conditions and limitations of algorithm effectiveness. In case of adoption of semantic models, which exploit conceptual modelling for retrieving information, there will be necessary to provide up-to-date linguistic and concept representations. Automatic classification and extraction methods are rarely characterized by an accuracy of 100%.

## 4. Conclusions

The applicability of indexing and automatic knowledge extraction technique from multimedia contents offers concrete possibilities for improving the power of current systems of Case and eCourt Management [4]. It is possible to figure out a scenario in information management automation in which simple activities of storing images, audio, video and text could be combined with post-processing activities for automatic knowledge extraction (completely understandable by the users). The obtained information could be directly used by end users, such as the transcription of audio

recordings or list of concepts of particular interest from an unstructured text, or they can be used to improve the content retrieval activities (such as indexing of automatic audio or audio-video recording, automatic association of frames or images with other related contents) without manual annotations or metadata insertion. An effective application of techniques for automatic knowledge extraction from media contents seems to have a greater chance of success if applied to vertical domains of knowledge and in contexts where it is possible to govern the processes of acquiring audio or audio-video contents. This is the case of criminal legal domain where, taking into account the importance of benefits that the application of these techniques could provide, several initiatives related to 7th FP of the European Community have been proposed to assess and quantify the real effectiveness of this kind of automation.

**Acknowledgments**

This work was partly funded by the European Union under the ICT program of FP7 project JUMAS, Contract No. 214306.

## References

1. Eurispes: Indagine sul Processo Penale. Roma, (2007).

2. A. Mittal: An overview of Multimedia Content-Based Retrieval Strategies. In Informatica 30. pagg. 347-356. (2006).

3. N. Dimitrova: Multimedia Content Analysis and Indexing for Filtering and Retrieval Application. Informing Science Special Issue on Multimedia Informing technologies – Part.1 Volume 2. No 4. (1999).

4. JUMAS Project Judicial Management by Digital Library Semantics. ICT Programme 7th FP EC Funded Research Project, http://www.jumasproject.eu/ (2009).

5. L.R. Rabiner: A tutorial on hidden Markov models and selected applications in speech recognition. In Readings in Speech Recognition, A. Waibel and K. Lee, Eds. Morgan Kaufmann Publishers, San Francisco, CA, 267-296 (1990).

6. M. Federico: A system for the retrieval of Italian broadcast news, Speech Communication, Volume 32, Issues 1-2, Accessing Information in Spoken Audio, September 2000, pp 37-47, ISSN 0167-6393 (2000).

7. M. Federico, N. Bertoldi: Broadcast news LM adaptation over time, Computer Speech & Language, Volume 18, Issue 4, October 2004, pp 417-435 (2004).

8. G. De Silva, T. Yamasaki, K. Aizawa: Evaluation of video summarization for a large number of cameras in ubuquitus home, in proceedings of the 13th annual ACM international conference on multimedia, pp 820-828, (2005).

9. A. Jaimes, T. Echigo, M. Teraguchi, F. Satoh: Learning personalized video highlights from detailed MPEG-7 metadata, in proc. of the IEEE international conference on image processing, pp 133-136, (2002).

10. Y. Takahashi, N. Nitta, N. Babaguchi: Video summarization for large sport video archives, in proc. of the IEEE international conference on multimedia and expo, pp 1170-1173, (2005).

11. Assfalg, J., Berlini, M., Del Bimbo, A., Nunziat, W., Pala, P.: Soccer highlights detection and recognition using HMMs, IEEE International Conference on Multimedia & Expo (ICME), 825-828, (2005).

12. Zhang, L., Lin, F.Z., Zhang, B.: Support vector machine learning for image retrieval, International conference on image processing, (2001).

13. Metzler, D. and Bruce Croft: W.: Linear feature-based models for information retrieval. Journal of *Information Retrieval.* 10, pp. 257-274 (2007).

14. Salton, G. and Buckley, C.: *Term Weighting Approaches in Automatic Text Retrieval*. Technical Report. UMI Order Number: TR87-881., Cornell University (1987).

15. *EPOC I – II – III European Pool Against Organized Crime*. Eurojust, European Commission (2008).

16 M. Velicogna: Use of information and communication technologies (ICT) in European judicial systems, Council of Europe, European Commission for the Efficiency of Justice (CEPEJ), CEPEJ Studies No. 7 http://www.coe.int/t/dghl/cooperation/cepej/series/default_en.asp (2008).