MIRIS Unleashed

Paolo Dongilli Johann Gamper*

Faculty of Computer Science
Free University of Bozen/Bolzano
Dominikanerplatz 3, 39100 Bozen, Italy
{*paolo.dongilli, johann.gamper*}*@unibz.it*

Abstract. Documentation and easy accessibility of data on matters relating to national and ethnic minorities is becoming increasingly significant. Even over the last couple of years, legal standardisation in Europe, the enlargement of the European Union and the establishment of international instruments such as the Framework Convention for the Protection of National Minorities and the European Charter for Regional and Minority Languages of the Council of Europe have contributed to a greater awareness of minority issues. However, the rapid growth of the Web makes it increasingly difficult to locate, organize, and integrate the available information. We herein present the research project MIRIS which for the very first time offers a web portal to all relevant information on minorities of all member states of the Council of Europe. MIRIS contains both legal as well as non-legal documents and provides a coherent presentation with powerful query capabilities including an ontology-driven search.

1 Introduction

Due to a new sensibility towards minority questions, the documentation and easy availability of data on ethnic, national, and linguistic minorities is increasingly significant. While international treaties and documents of international organizations are mostly available on the Web, the gathering of extensive information on the legal standards in different countries, such as national legislation or case-law, is still a difficult task. The existing databases about ethnic/linguistic minorities are either limited to a few countries (e.g. MINELRES and CEDIME-SE, which are focusing on south-eastern European Countries only) or they only cover a few topics (e.g. Mercator, which focuses on linguistic rights only). Furthermore, those databases are mostly unstructured. Often, documents are simply listed without being organized by keywords or topics. From a content and technical point of view, the users search is not supported/backed up. Thus, a quick search or a comparison between different countries is often a difficult task.

Taking into account these problems, the ongoing research project MIRIS (Minority Rights Information System) aims to develop a comprehensive information system about ethnic minorities and autonomies in Europe. MIRIS offers for the very first time a Web portal to all relevant information on minorities of all countries which are member states of the Council of Europe. It will become a major consultation center for minority questions in Eastern and Western Europe. MIRIS is a project led by the European Academy of Bozen/Bolzano in cooperation with the Faculty of Computer Science of the Free University of Bozen/Bolzano (Italy).

*In cooperation with the European Academy of Bozen/Bolzano. Project funded by Propter Homines, Liechtenstein.

In the present article we will explore in detail the architecture of MIRIS and in particular its main components: the user interface, the ontology-driven search and the authoring tool as front ends of the whole MIRIS architecture.

2 The MIRIS Technical Architecture

The MIRIS system can be seen as a three-layered architecture (see figure 1) consisting of the following blocks:

- Applications and Services,
- Middleware,
- Remote and data services,

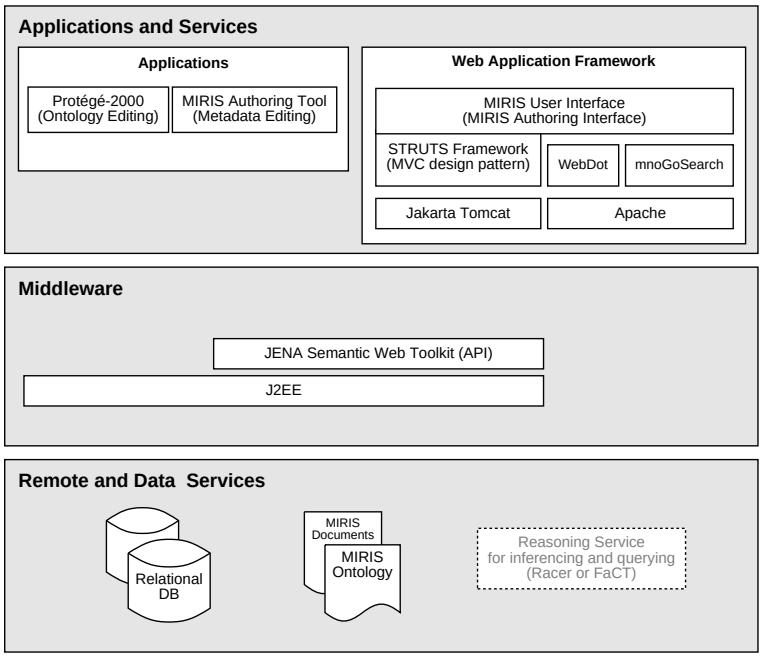

Figure 1: MIRIS system architecture.

In the following sections we will explain in detail, meaning and functionalities of the single layers building up MIRIS.

3 Applications and Services

This layer represents the interface to the two categories of users of the system: *active* and *passive users*. The team of jurists working for MIRIS are the active users who are responsible for the information-gathering process which began in 2001 and is focusing on some twenty-eight countries in Western and Eastern Europe. On the other hand the passive users are all the people interested in the implementation of minority rights: not only experts, public officials, NGO's (Non-governmental Organizations) and minority representatives but also researchers and students.

The applications will be presented first, and followed by a description of the underlying services.

Table 1: Overview of MIRIS documents (September 2003)

Document types	Internal	External	Total
Case Laws	223	25	*248*
Country Information	119	76	*195*
International Organization Documents	39	52	*91*
Info about International Organization Documents	23	25	*48*
Minority Information	105	78	*183*
National Laws	1376	612	*1988*
Reports	284	269	*553*
Treaties	84	48	*132*
Total	2253	1185	***3438***

3.1 MIRIS Authoring Tool

The active users are currently eight jurists (out of a total of sixteen experts who have been working for the project). Their main task is to create and store for all candidate documents a set of metadata which include administrative information as well as content information. These pieces of information are important for organizing, accessing, and presenting the documents. When properly implemented, metadata can crisply and unambiguously describe information resources, enhancing information retrieval and enabling accurate matches to be done, while being totally transparent and invisible to the user. Search specificity is increased (noise reduction — only good matches) and search sensitivity is boosted (i.e. silence or missed matches are decreased and signal-to-noise ratio increased — all good matches) [6].

Table 1 represents a snapshot of the documents we collected to date (September 2003). We distinguish between internal documents, which are physically stored in our database, and external documents, for which we only store a URL.

The MIRIS Authoring Tool (MAT) is the application used by the jurists to edit the metadata related to the documents they process. In particular with the MAT the active users can view, create, edit and delete the metadata which are stored in the database.

The first version of the MAT was a web application which has been dismissed and reprogrammed into a pure Java program. Figure 2 shows two screenshots of the metadata-editing phase of a Greek case law and a Bulgarian national law.

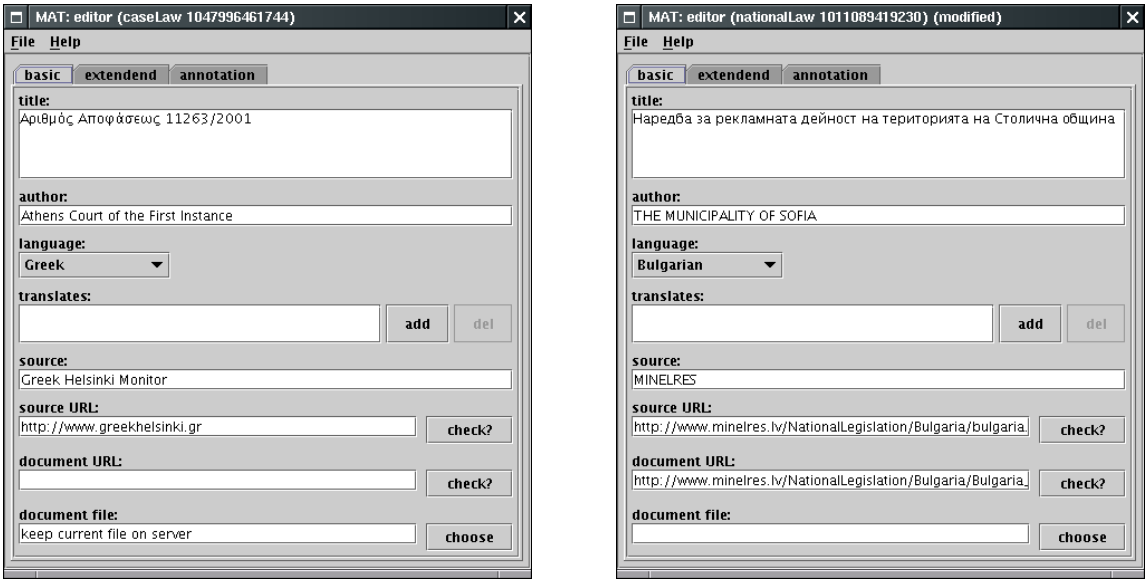

Figure 2: MIRIS Authoring Tool: metadata editing.

The items shown in the *basic* tab are only a part of the metadata fields accompanying each document; the hidden *extended* tab contains other fields such as *date of judgment*, *date of entry into force*, *guiding principle*, *document number*, *related minorities*, etc. These fields vary according to the document type chosen. The third tab (*annotation*) is used to annotate the documents with topics. The editors can navigate the ontology and they can assign one or more topics referring to each document as a whole or to the single sections of it.

3.2 Ontology Editor

We put a particular emphasis on the extension of our metadata with topics, which explicitly describe the content of the documents. We allow content annotation not only at the document level, but down to the level of law articles. The topics are organized in a domain ontology which represents common knowledge about ethnic minorities and provides the vocabulary required for describing the content of the documents. An excerpt of our relatively simple topics ontology is shown in figure 3. The ontology consists of a taxonomy built on the *isa*-relation and a *relatedTo*-relation which connects related topics.

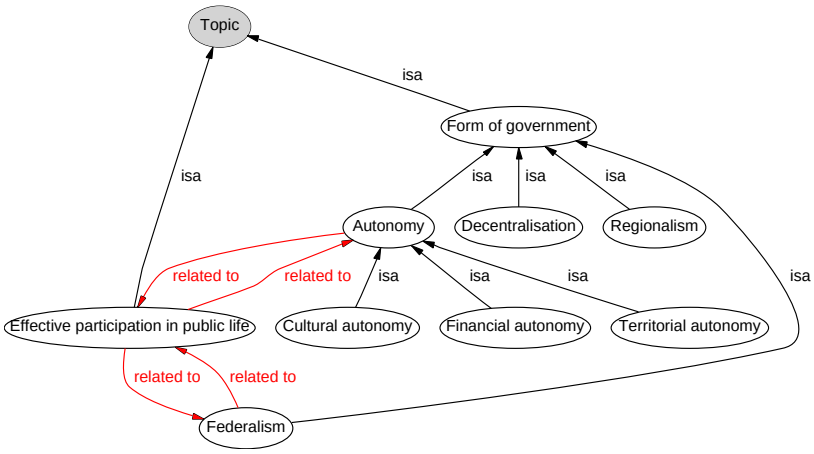

Figure 3: Extract of the topics ontology.

The ontology is currently being created by our lawyers using Protégé-2000, a powerful knowledge-base-editing environment developed at the Stanford University [7, 8]. It allows us by means of its flexible metaclass architecture to use configurable templates for the classes we include in our knowledge base.

We generate content annotation at different levels for internal and external documents. This allows fine-grained, content-based search in a uniform way independent of the location of the actual document. Additionally, for local documents MIRIS guides the user directly to relevant parts of the document. For external documents, only the entire document can be accessed and the user has to scroll to the indicated part.

3.3 MIRIS User Interface

An important aspect of the MIRIS system is to provide a coherent presentation and powerful query capabilities to information about minority rights. The following functionalities and interactions are offered:

- full-text search,
- metadata-based search, and
- ontology-driven search.

Full-text Search About 40% of the documents available in MIRIS are stored in a keyword index available for full text searching. For each document found, the title, a short text sample, as well as type and size of the document are indicated (see figure 4). For implementing the full-text search we use a slightly modified version of *mnoGoSearch*[1], a full-featured SQL-based web search engine; for UNIX it is free software covered by the GNU General Public License (GPL). *mnoGoSearch* consists of two parts. The first part is the indexing mechanism (indexer). The indexer walks over html hypertext references and stores found words and new references into the database. The second part is a web CGI front-end which provides search using data collected by the indexer.

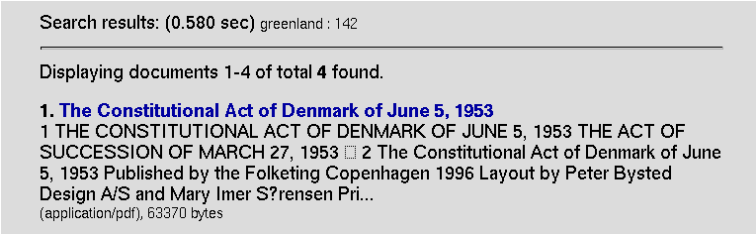

Figure 4: Screenshot of the MIRIS full-text search.

Metadata-based Search Figure 5 shows instead a screenshot with the components of the MIRIS interface for the *metadata-based search mode*.

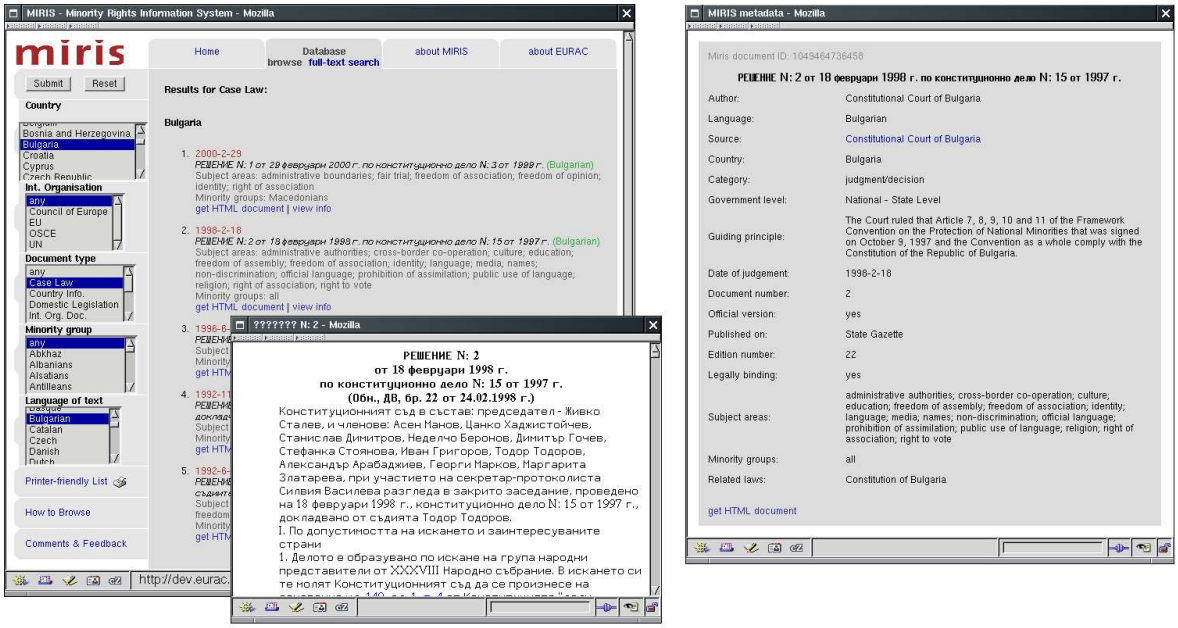

Figure 5: Screenshot of the MIRIS metadata-based search interface.

On the left-hand side of the main window (left), the user can specify up to five query parameters by choosing values from several select boxes. In the example shown, we are interested in all case laws of Bulgaria in Bulgarian language. The right-hand side of the same window reports the result of the query ordered by country (only one in this case) and by date. The result list contains for each hit some important metadata including the date of the document, the topics which are regulated in the document (related subjects), and the minorities

[1] http://search.mnogo.ru/

for which this document applies. The user can choose then to visualize the document in the proposed format (window in the middle) along with a full version of the pertaining metadata (window on the right).

Ontology-driven Search The last component of the MIRIS user interface is the newly developed prototypical web application for the ontology-driven search. The user can navigate the nodes of the topics ontology following the existing relations: *broader term*, *narrower term* and *related term*. Whenever a node is clicked, the term (topic) is shown along with its description, possible synonyms, comments and the relations to other terms. A navigation graph is generated and the user can see the selected term, its immediate super-topics and sub-topics. The system displays also the number of documents which are tagged with the chosen topic and the user can decide to finally see the list of texts found (see figure 6).

There are some extensions we want to implement in the user interface of this component. First it would be useful to additionally show the number of documents pertaining not only to the selected topic but also to the broader, narrower and related terms. Second, the result set, for performance reasons, should only be an overview of the documents, organized in a table e.g. by document type and country (to name just two search criteria). Then it's up to the user to refine the search selecting a specific country and/or one of the document types getting the complete set of information about the documents found. Third, we could let the users choose more than one topic. This requires though a more sophisticated query input mechanism, which will be part of our future work (see section 6).

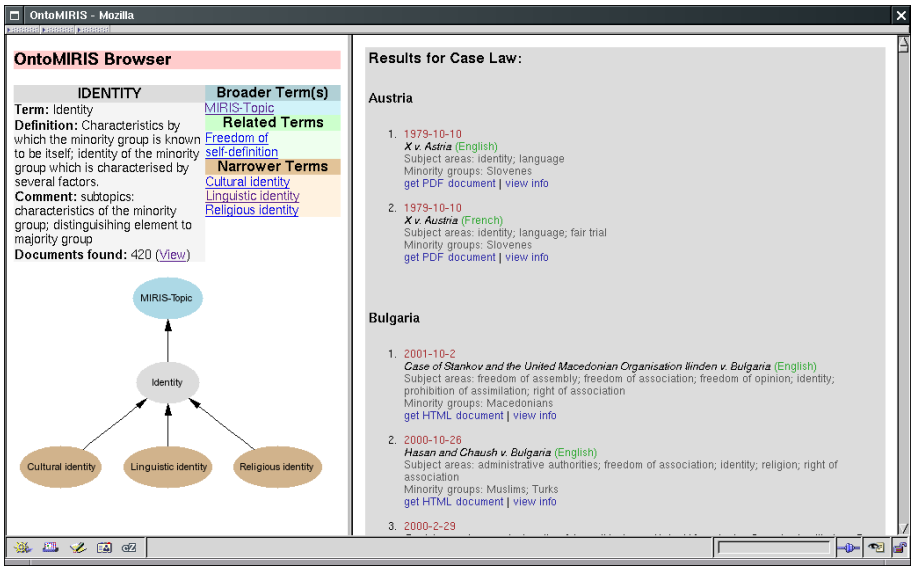

Figure 6: MIRIS ontology-driven search

3.4 Underlying Tools and Services

The MIRIS User Interface is a web application built using the Jakarta Struts Framework[2]. Struts combines two of the most popular server-side technologies – JSPs and servlets – into a server-side implementation of the Model-View-Controller (MVC) design pattern. Struts is currently supported by all the major application servers including BEA, SUN, HP, and (of course) Apache Jakarta-Tomcat, the JSP/servlet container which runs on the MIRIS server.

[2]http://jakarta.apache.org/struts/

The web server Apache is also installed to execute among others the CGI-scripts WebDot[3], and mnoGoSearch (see paragraph *Full-text Search* in section 3.3).

In our system we use WebDot to create the navigation graph for the ontology-driven search (see section 3.3). WebDot is a graph layout engine (based on the Graphviz[4] package from AT&T Research) that converts a graph description from a `.dot` file into an image. A graph can contain URLs so that the nodes and edges are clickable. To respond to user mouse events, a map must be provided. This mapping information is generated dynamically from the same graph source file as the image itself (see figure 7).

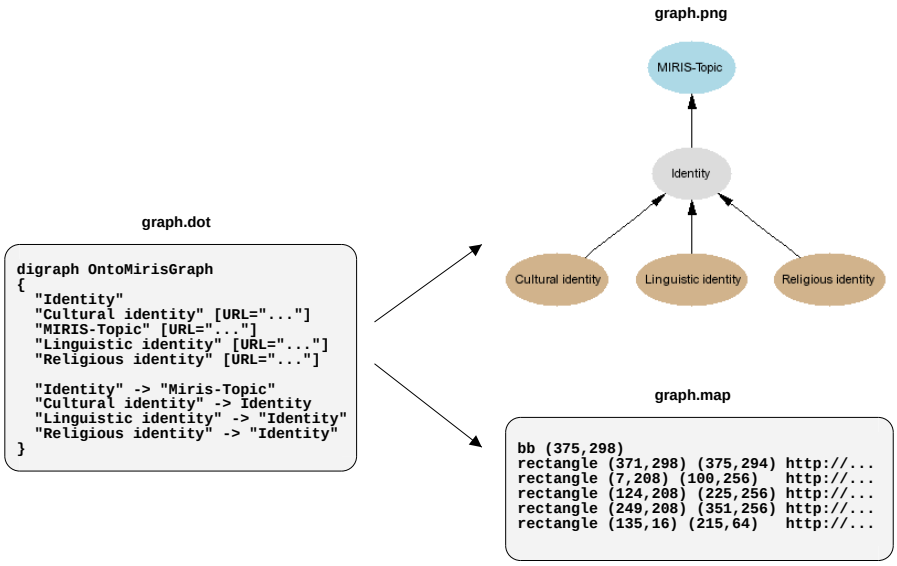

Figure 7: Conversion of a `.dot` file into an image and a map.

In MIRIS a Java class creates first the `.dot` file on the fly, depending on the current topic that the user is exploring. As soon as the `.dot` file is saved on the system, WebDot can retrieve it to generate the image and the map for the requesting browser.

4 Middleware

All the applications we developed (the Miris Authoring Tool, the Miris User Interface and some other unmentioned routines for our jurists) are platform-independent programs built upon J2EE. The same applies for the application server we use (Apache Jakarta Tomcat) and the ontology editor Protégé-2000 which are pure Java applications.

For the ontology-driven search we also use the JENA semantic Web Toolkit[5], a Java API for manipulating RDF graphs. We use it for accessing our ontology of topics which is exported from Protégé-2000 in RDF format.

Finally we have WebDot which is not based on Java: it is written in Tcl language and requires the Tcldot package from Graphviz.

5 Remote and Data Services

The central repository of the MIRIS metadata set is stored in an Oracle 9i database. Whenever a piece of metadata is requested, it is retrieved from the database as XML file and rendered

[3]http://www.graphviz.org/webdot/

[4]http://www.research.att.com/sw/tools/graphviz/

[5]http://www.hpl.hp.com/semweb/

to HTML via an XSL stylesheet for being displayed to the requesting user.

As far as the documents are concerned, rather than storing them in the database, we decided to save them directly on the MIRIS server for a faster access. The MIRIS ontology is also stored on the server. It is possible for our jurists to change it uploading a new one in case some corrections need to be made. An ontology change is automatically detected and the new uploaded ontology is reread and deployed by the MIRIS Authoring Tool and by the ontology-driven search application.

Nearly all of the relevant information is stored in human-readable documents, such as the state reports to which the member states are committed, minority reports, international law contracts, national laws, etc. The most important documents though, are encoded for structural information using XML, which not only allows a coherent visualization of the documents but also a more specific and fine-grained search. We decided to use the DocBook XML document type definition (DTD) because of its general nature. DocBook has been adopted by a large and growing community of authors writing documents of all kinds, it is easy to understand and there is a rapidly expanding support for it in a number of open source environments. Since structural encoding is a very time-consuming task, as stated above only the most important documents and documents which rarely change can be stored locally.

6 Conclusions and Future Work

In this paper we have presented MIRIS, an information system about minority rights and autonomies in Europe. Using the tools provided by the MIRIS system, the jurists working for the project have cataloged to-date 3438 documents from Eastern and Western Europe, providing metadata descriptions for each resource (document) found. The content of the documents has also been described using topics picked from an ontology built by our domain experts. All the information collected are made available to the public through the MIRIS user interface including a metadata-based search and an ontology-driven search.

For the future, our main efforts will converge towards three main points.

First of all we want to map (part of) our metadata fields onto the Dublin Core metadata element set (DCMS) which consists of the following 15 elements: *title*, *creator*, *subject* (our topics), *description*, *publisher*, *contributor*, *date*, *type*, *format*, *identifier*, *source*, *language*, *relation*, *coverage* and *rights*. The reason is that the DCMS is now an international standard (ISO 15836:2003(E)) for cross-domain information resource description and we want to conform to it. For further details about the Dublin Core Metadata Initiative see [3].

Second, we want to extend our topics ontology to a broader ontology including most of the concepts and relations which are related to our project. Our final aim is to build a good-quality ontology where our domain experts make a full use of the chosen language (DAML+OIL [2], OWL [9] or ODL$_{I^3}$ [1]), in a way that the ontology can capture the author's rich knowledge of the domain.

Third and last point, we will interface the ontology editing phase and the MIRIS Ontology-driven Search application with one of the free Description Logics reasoners available (FaCT [5] or RACER [4]). Description logics (DLs) are knowledge representation languages tailored for expressing knowledge about concepts and concept hierarchies. DLs models a domain in terms of *individuals* (modelled objects), *concepts* (descriptions of groups of objects sharing common characteristics) and *roles* (the relationships between concepts and individuals). The abovementioned languages (DAML+OIL, OWL, and ODL$_{I^3}$) can be viewed as dialects of DLs. The presence of a DLs reasoner in the background of MIRIS will be a great help for our jurists who will be able to automatically classify the ontology concepts and check them for satisfiability. On the other hand our passive users will be presented a richer interface

where they will have the opportunity to incrementally build a query choosing concepts and properties (relations or attributes) via an ontology-driven diagrammatic support.

References

[1] S. Bergamaschi, S. Castano, and M. Vincini. Semantic integration of semistructured and structured data sources. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 28(1):54–??, 1999. Special section on semantic interoperability in global information systems.

[2] The DAML+OIL web site. http://www.w3.org/TR/daml+oil-reference, 2003.

[3] Dublin Core Metadata Initiative. http://dublincore.org, 2003.

[4] Volker Haarslev and Ralf Möller. Description of the RACER System and its Applications. In *Proceedings of the International Workshop on Description Logics (DL-2001)*, Stanford, USA, August 2001.

[5] Ian Horrocks. FaCT and iFaCT. In P. Lambrix, A. Borgida, M. Lenzerini, R. Möller, and P. Patel-Schneider, editors, *In Lambrix et al.*, pages 133–135, Stanford, USA, August 1999.

[6] Boulos M., Roudsari A., and E. Carson. Towards a semantic medical Web: HealthCyberMap's tool for building an RDF metadata base of health information resources based on the Qualified Dublin Core Metadata Set. *Medical Science Monitor*, 8(7):124–136, 2002.

[7] N. F. Noy, R. W. Fergerson, and M. A. Musen. The knowledge model of Protégé-2000: Combining interoperability and flexibility. In *Proceedings of the 12th International Conference on Knowledge Engineering and Knowledge Management (EKAW'2000)*, pages 12–32, October 2000.

[8] N. F. Noy, M. Sintek, S. Decker, M. Crubezy, R. W. Fergerson, and M. A. Musen. Creating semantic web contents with Protégé-2000. *IEEE Intelligent Systems Journal*, 16(2):60–71, 2001.

[9] Web Ontology Language (OWL). http://www.w3.org/2001/sw/WebOnt/, 2003.