

2009

# A Model For e-Government Digital Document

Flora Amato

*Department of Informatica e Sistemistica, Universita' degli Studi di Napoli "Federico II", flora.amato@unina.it*

Antonino Mazzeo

*Department of Informatica e Sistemistica, Universita' degli Studi di Napoli "Federico II", mazzeo@unina.it*

Vincenzo Moscato

*Department of Informatica e Sistemistica, Universita' degli Studi di Napoli "Federico II", vmoscato@unina.it*

Antonio Picariello

*Department of Informatica e Sistemistica, Universita' degli Studi di Napoli "Federico II", picus@unina.it*

Follow this and additional works at: <http://aisel.aisnet.org/mcis2009>

## Recommended Citation

Amato, Flora; Mazzeo, Antonino; Moscato, Vincenzo; and Picariello, Antonio, "A Model For e-Government Digital Document" (2009). *MCIS 2009 Proceedings*. 49.  
<http://aisel.aisnet.org/mcis2009/49>

This material is brought to you by the Mediterranean Conference on Information Systems (MCIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in MCIS 2009 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact [elibrary@aisnet.org](mailto:elibrary@aisnet.org).

## A MODEL FOR E-GOVERNMENT DIGITAL DOCUMENT

Amato, Flora, Department of Informatica e Sistemistica, Universita' degli Studi di Napoli

"Federico II", via Claudio 21, 80125, Naples, Italy, flora.amato,@unina.it

Mazzeo, Antonino, Department of Informatica e Sistemistica, Universita' degli Studi di Napoli

"Federico II", via Claudio 21, 80125, Naples, Italy, mazzeo@unina.it

Moscato, Vincenzo, Department of Informatica e Sistemistica, Universita' degli Studi di Napoli

"Federico II", via Claudio 21, 80125, Naples, Italy, vmoscato@unina.it

Picariello, Antonio, Department of Informatica e Sistemistica, Universita' degli Studi di Napoli

"Federico II", via Claudio 21, 80125, Naples, Italy, picus@unina.it

### Abstract

The presence of a great amount of information is typical of bureaucratic processes, like the ones pertaining to public and private administrations. Such information is often recorded on paper or in different digital formats and its management is very expensive, both in terms of space used for storing documents and in terms of time spent in searching for the documents of interest. Furthermore, the manual management of these documents is absolutely not error-free. To efficiently access the information contained in very large document repositories, such as public administration archives, techniques for syntactic and semantic document management are required, so to ensure a large and intense process of document dematerialization, and eliminate, or at least reduce, the quantity of paper documents. In this work we present a novel RDF model of digital documents for improving the dematerialization effectiveness, that constitutes the starting point of an information system able to manage documental streams in the most efficient way. Such model takes into account the important need that is required in several E-Government applications which, depending on authorities or final users or time, provides different representations of the same multimedia contents.

**Keywords:** *Semantic Document Management, Information Retrieval, Knowledge Engineering*

### 1 INTRODUCTION

E-Government based applications need suitable data models for the representation of the composition of their different media elements or multimedia document models. Nowadays, in fact, almost all the novel bureaucratic processes, are characterized by both formatted text and a huge quantity of multimedia data (e. g. audio, still images, sometimes videos) documents, which need to be properly managed, stored and distributed. Multimedia document models are employed to model the semantic relationships between the media elements participating in a multimedia document.

Our research project in the Italian notary domain, [Amato et al., 2008] is an example of an advanced multimedia application that emphasizes this need of a model for multimedia material that are part of legal checked documents. In this work we propose a novel model of multimedia document that is particularly suitable for the management of the whole flow of digital documents and that in particular, allow: i) automatic information extraction from digital documents; ii) retrieval; iii) semantic interpretation of the relevant information presented in the document, iv) storing and v) long term preservation. The model described in this paper integrates three fundamental aspects that are strictly related to the evolution of multimedia documents:

- the structure of documents used in e-Gov applications,
- the organization and management of multimedia data,
- the presentation of the same information using different formats.

However, it is important to notice that we do not claim here to provide a comprehensive model

of multimedia documents; we are focusing here on the aspects of such documents that have a great impact in e- Gov processes. To best understand the core idea of our paper, let us consider a criminal investigation in the Italian domain, that can be easily extended to the case of different countries. All the documents related to the criminal trial are collected into a folder containing verbal transcription of the interrogations, videos, crime scene images, picture of the victims and so on. All the different part of the documents forms the legal documents, using a set of rules established by the Law. The different part of the documents, (or segments) may be accessed only by authorized users and are presented in different ways to different users (the police, the judge, the lawyers and the criminals) in different contexts.

The main idea here is that a multimedia document consists of objects such as text, images, drawings, structured data, operational codes, programs and movies, that, according to their relative position on the support, determine the shape and, consequently the structure of the document itself through the relationships between them. During the various and different e-Government processing phases, that are really different from an application domain to another, a document is processed and eventually stored on various kinds of media, properly defined in order to archive and preserve papers, photographic films and microfilms, VHS cassettes, Magnetic Tapes, DVD disks, and more.

In this paper we will provide a novel RDF based model for digital documents and we describe a system for multimedia document management, in particular for those regarding archiving and long term preservation, in order to improve the dematerialization effectiveness, that constitutes the starting point of an information system able to manage documental streams in the most efficient way. The proposed document model is characterized by the separation between presentation and content, allowing to solve, among others things, open problems related to the technology evolution as the juridical validity of a document, and the support of different multimedia type. The paper is organized as follows.

The next section presents the proposal for digital document model; section 3 outlines the general architecture of an e-Gov information system suitable designed on the basis of the digital document model; eventually some preliminary results about the prototypic implementation of the system, are reported in the section 4.

## **1.1 Related Work**

Fast access to multimedia information requires the ability to search and organize the information. In such an area the main objective of the researchers is to index in an automatic way multimedia data on the base of their content in order to facilitate and make more effective and efficient the query processing.

In the following, supported by the related state-of-the-art, we describe the major challenges in developing reliable image and text database systems. In the Image Database Systems field, in the last decade, most of researches are focused on Content Based Image Retrieval (CBIR). The CBIR is characterized by the ability of a system in retrieving relevant information on the base of image visual content and semantics expressed by means of simple search-attributes or keywords. Traditionally, CBIR addresses the problem of finding images relevant to the users' information needs from image databases, based principally on low-level image global descriptors (color, texture and shape features) for which automatic extraction methods are available, see [Smeulders et al., 2000],[Lew et al., 2006],[Datta, Joshi, 2008] for details.

More recently, it has been realized that such global descriptors are not suitable to describe the actual objects within the images and their associated semantics. For these reasons, two main approaches have been proposed to cope with this deficiency: firstly approaches have been developed whereby the image is segmented into multiple regions, and separate descriptors are built for each region; secondly, the use of salient points has been suggested. Following the first approach, different systems like, SIMPLicity [Wang et al., 2001] and Blobworld [Carson et al., 2002] have been developed. The second approach avoids the problem of segmentation altogether by choosing to describe the image and its contents in a

different way. By using salient points or regions within an image, in fact, it is possible to derive a compact image description based around the local attributes of such points [Hare, Lewis, 2005].

Our proposal [Boccignone, et al., 2008] follows the second approach avoiding the problem of early segmentation and exploits color, texture and shape features in the principled framework of Animate Vision, according to which is the way that features are dynamically organized in the Where-What space that endows them with information about the context in terms of categories. The discovered semantic knowledge in terms of categories and relations among them is part of a particular folksonomy produced by humans through the Flickr image management system [Capasso et al., 2008]. It is worth recalling that the use of context/semantics for improving retrieval process is also taken into account by Wang et al. [Wang et al., 2001], in the form of categories, by Del Bimbo et al. [Corridoni, et al., 1999], [Colombo, Bimbo, 2002], in terms of color-induced sensations in paintings, and clearly addressed by Santini et al. [2000], through a mechanism of similarity tuning via relevance feedback. Finally, more recent systems, such as Cortina and ALIPR [Manjunath et al., 2007], [Li, Wang, 2006] have as goal the automatic classification of images on the base of low-level features and high-level human annotations.

The Text Database Systems, instead, require the use of different techniques from interdisciplinary fields regarding legal ontologies from both theoretical - in order to define legal lexical dictionaries - and application - for organization, storage, retrieval purpose points of view. In order to represent legal knowledge, several works have been proposed, such as: Breuker's Functional Ontology of Law [19], Frame-based Ontology of Visser [Visser, Bench-Capon, 1996], McCarty's Language of Legal Discourse [McCarty, 1989] and Stamper's Norma [Stamper, 1991]. As a consequence of such theories, several ontologies are now available, such as Ontology-based Legal Information Environment (ON-LINE), Dutch Unemployment Benefits Act (DUBA) and Cooper-ative Legal Information Management and Explanation (CLIME). Several approaches that are based on the wordNet project have been also done: in particular, in Italy, JurWordNet [Tiscornia, 2003] is the first Italian legal ontology.

In order to perform identification of concepts and document classification for automatic document description, several works have used pattern recognition techniques, as SCISOR [Jacobs, Raul, 1990] and FASTUS [Hobbs et al., 1992]. In the system BREVIDOC, documents are automatically structured and the important sentences are extracted, these sentences are classified according to their relative importance [Miike et al., 1994].

From the NLP point of view, legal research concentrate on the development of thesauri, machine learning for features recognition, the disambiguation of polysems, automatic clustering and neural networks. The most important systems are FLEXICON, KONTERM, ILAM, RUBRIC, SPIRE, the HYPO extension and SALOMON [McCarty, 1989].

## **2 THE E-GOVERNMENT DIGITAL DOCUMENT PROPOSED MODEL**

The core aspect related to a novel and efficient dematerialization process is the idea standing beyond the common concept of document. In Italy, an e-Gov digital document model regulated by recent laws about Public Administration organization. The starting point of the model is the Document definition of the dpr 445/2000, art. 1, comm. 1, lett. a, that state that the representation of the information contained in a document can be unbind from the paper support, and that a document can contain multimedia elements. The proposed model for the bureaucratic document is showed, as RDF graph, in fig. 1.

In order to optimally manage and preserve the real useful information contained in a certain document, despite of the required different presentation formats, it is necessary to provide a novel model for a multimedia document, pointing out how to identify and characterize what is the minimal content of the document itself, given a certain normative context, and relate this minimal content to a presentation level, depending on different users at different times.

The proposed document model is composed by three layers, defined in order to manage and preserve the real useful information contained in the multimedia documents, despite the required different presentation formats. The content will be processed in order to make possible semantic procedure on it, and will be showed in different way, subjected to the Italian normative context, depending on different users at different times .

The layer in which the relations on he documents are grouped are described in the following.

1. Data Management Layer: describes the semantic minimal content (or kernel) of a document, usually codified by different media types. This layer manages the different data types, furnishing all the necessary functionalities and facilities operating over a certain single media; for example, information extraction and indexing over texts, images, videos, audios and son on.
2. Composition layer: provides a proper integration of the heterogeneous data sources, having the aims of regulating the coexistence of the different objects within the context of a single document.
3. Presentation layer: this layer regulates the way in which the information has to appear to a single user within a certain context in different times.

In appendix A we report the full RDF serialized description of the model depicted in figure 1, in which the set of documents related of a single thing is enveloped in a folder. Every document is memorized in a proper format, chosen on the basis of the authority needs or the available technology (for example, it can be memorized in pdf, doc or odt), and is correlated by property, as the name of the author, the date of creation and change. The access right, indicating who and with which privileges the document may be accessed, are associated to the document itself. The Presentation layer codifies this kind of proprieties, associated to the modality on which the document is presented to the final users.

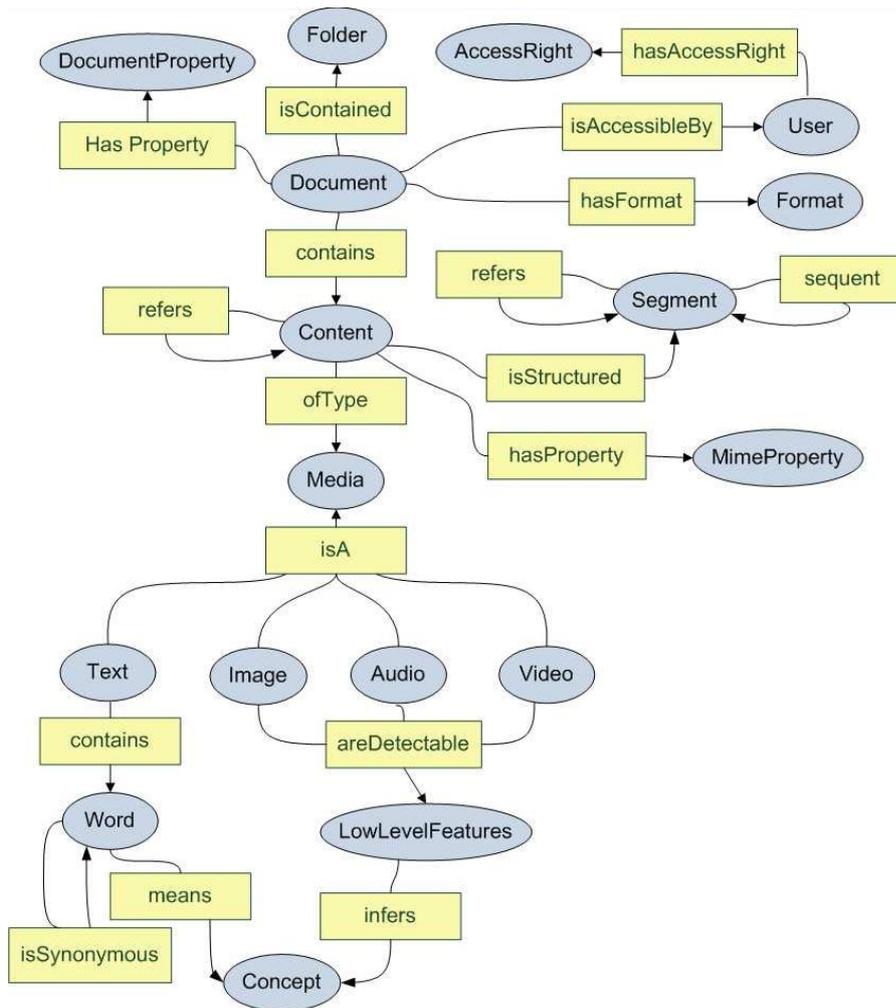


Figure 24. The Digital Document RDF Model

When the documents are submitted to the system preliminary procedure extracts the content of the examined document, such content will be organized in a ordered list of segment. Every segment constitutes a portion of the document and is of a single type of media, then it can be a sequence of words of a text delimited by punctuation mark, an image fragment or an audio stream. The relation between the elements of the same segment are modeled, on the basis of the type of media, in the data management layer. In the case of text segment, the contained words are extracted, and NLP and NER procedure are performed, in order to providing lexical, syntactical and semantical information about them. Based to the particular acception, synonymous sets are individuated for each word, and the proper concept is associated to it,giving in this way the possibility to perform, for example, semantic search operations on the documents.

For the other media, as images, audios and videos, low level features are individuated and extracted by apposite procedure realized in the data management layer, and concepts to associate to set of these feature are inferred.

The relations about different segments of the same or different media are codified in the Composition Layer, that contains informations as the reference of a segment of text to an image. In order to show how he model may be useful for e-Gov applications, let us consider again the criminal investigation example described in the introduction.

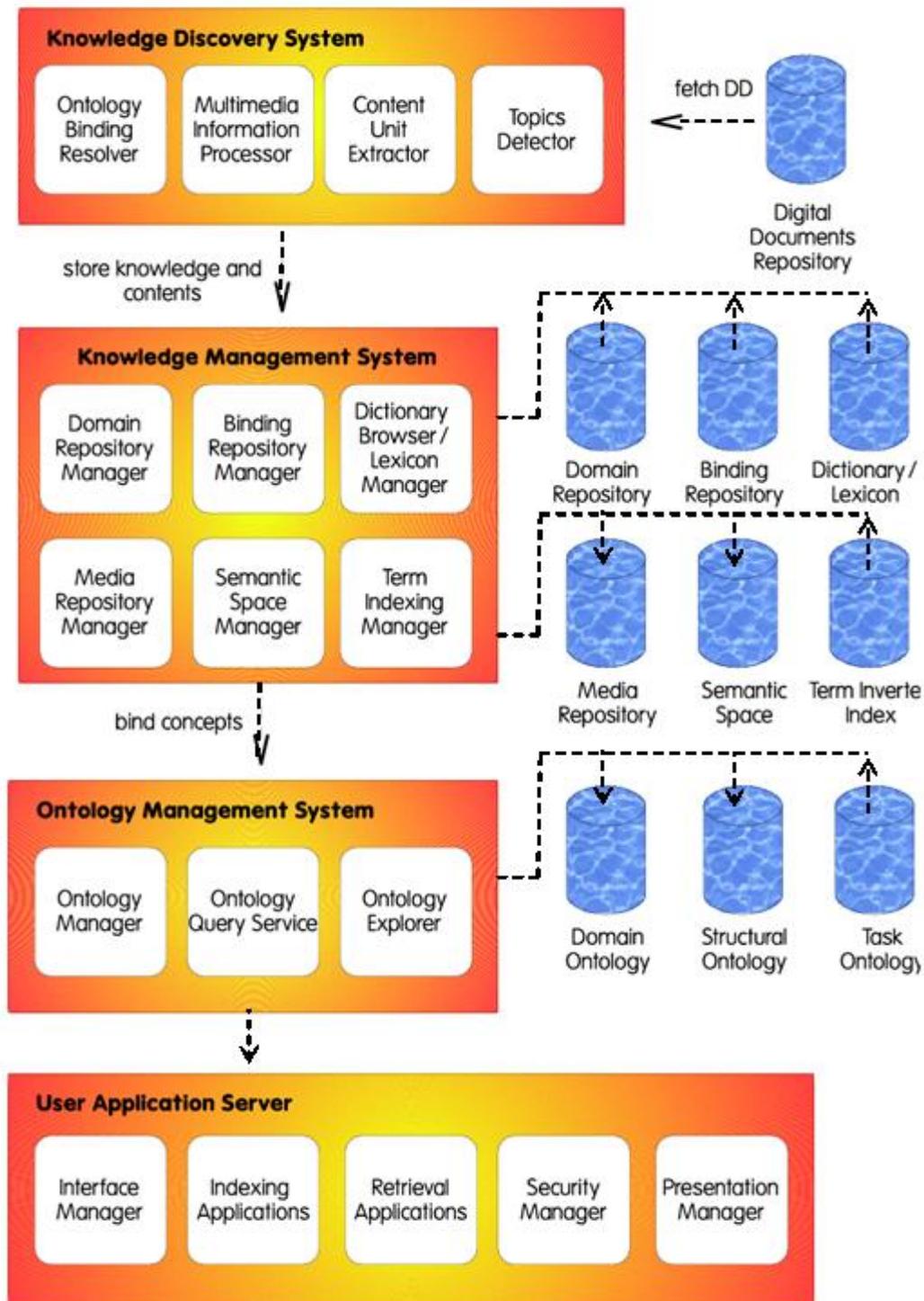


Figure 2. Model System Architecture

We note that once we submit the investigation documents to our system, the content is extracted and processed. The proper concepts are the associated to the words presented in the document, so it is possible to perform semantic search on them, for example, searching the profiling details of a person, given a name and surname in input, considering for the research the only person that have a conviction on murder charges on them. Another example is the possibility to highlight the words or the image fragment belonging to a given input concept. Once the relation of different segment are individuated, it is possible correlate them, for example indicating that a text segment is the description of a crime scene

represented in a photo, of that a text string constitutes the name of the person that speaks in a particular audio text.

### 3 SYSTEM DESCRIPTION

#### 3.1 The Architecture

The proposed Multimedia Document Management System to serve its expected purpose has the following main features:

- a unified data model that takes into account content-based and document-based characteristics;
- an ontological support for managing the semantic of data;
- a multi-layer architecture with different kinds of user interfaces;
- advanced functionalities for document indexing and semantic retrieval.

Figure 2 shows at glance the architecture of our system. Resources in the system are Digital Documents (DD) that are managed by a dedicated component, named Digital Document Repository (DDR). Its objectives are, from one hand, to allow interoperability among the different data formats by providing import/export procedures and, from the other one, to manage security in the data access. Moreover, documents can be organized in specific folders to facilitate the management and retrieval.

In according to the introduced data model, it is possible to associate with a digital document a set of semantic concepts - retrievable by semi-automatic information extraction procedures and related to single content units of a document - and set of keywords - defined as particular properties of the entire document.

In the early stage, documents acquired by means of apposite OCR techniques are stored in the DDR and undergo the information extraction processing described in the following.

In the indexing stage, digital documents are picked up from DDR by a particular module called

Knowledge Discovery System (KDS). The KDS analyzes digital documents with the goal of obtaining useful knowledge from raw data. In particular, a Content Unit Extractor has the task of extracting (by a human-assisted process) content units from a document (and of generating an instance that can be stored in the system knowledge base), while, the Multimedia Information Processor sub-module infers knowledge in terms of semantic concepts from the different kinds of multimedia data [2],[12] (e.g. text, audio, video, image). In the opposite, a Topics Detector sub-module operates on the not-structured view of a document and aims at detecting by a natural language processing the most relevant topics for the entire document. Eventually, the Ontology Binding Resolver sub-module has the objective of creating for each discovered concept/topic a binding association with a node of domain ontology.

The extracted knowledge is then stored in the Semantic Knowledge Base (SKB) managed by a Knowledge Management System (KMS). The KMS performs indexing operations on the managed information, providing to applications functionalities for browsing and retrieval documents. The components of the SKB (and the related KMS managing modules) are described in the following.

- *Dictionary* (for each supported language) - It contains all the terms of a given language with the related possible meanings and some linguistic relationship among terms (e.g. WordNet). Each dictionary is managed by an apposite management module, called Dictionary Browser.
- *Lexicon* - It contains all the terms known by the system: dictionary terms and named entities (names of people and organizations). The is managed by an apposite module, called Lexicon Manager.
- *Term Inverted Index* - It is the data structure used for indexing terms inside documents. For each term known by the system (and contained in the lexicon) a posting list, that contains identifiers of documents and contents referring to such a term with the related frequency, is created. The inverted index is managed by an apposite Term Indexing Manager.

- *Semantic Space* - It allows the storage of the single atomic pieces of knowledge belonging to document content units, and called document segments. It is an abstraction of a shared virtual memory space (with read/write methods) by which applications can exchange multimedia data. This space is called semantic because each element is associated with a particular structural ontology that allows to relate segments of the same content unit and content units of different documents. The Semantic Space Manger provides functionalities for reading, writing, removing and searching tuples in the space.
- *Domain Repository* - It contains the description of application domain concepts and is managed by a Domain repository Manager.
- *Binding Repository* - it contains the associations between document and domain repository concepts and is managed by a Binding Repository Manager.
- *Media Repository* - it is an Object Relational DBMS able to manage the different kinds of multimedia contents. It is managed by a particular module, called Multimedia Information Manager able to support classical multimedia query for the different kinds of multimedia data - e.g. query by example/feature for images, query by content/keywords for images and text, and so on.

The semantic associated to the data contained in the knowledge base is then managed by the Ontology Management System (OMS), that contains the ontology models used by the system. In particular, we exploit three kinds of ontologies (managed by an Ontology Manager): (i) a set of domain ontologies that relate the semantic concepts in a given domain, (ii) a set of task ontologies that determine the role/meaning of a content unit in a document and (iii) a set of structural ontologies that code the relationships between contents and segments. The Ontology Explorer allows browsing of the concepts in the ontologies, while the Ontology Query Service is a component devoted to execute queries on the ontologies.

From the user point of view, the functionalities provided by the system are the indexing of a document and the semantic retrieval of information. The application interfaces are realized both as web services and desktop programs (and managed by an apposite Interface Manager). Finally, there are two modules for security and presentation management.

### **3.2 Implementation Issues**

Due to the great amount of data to deal with and security issues, we have chosen to implement the document management system prototype using ORACLE technologies (Oracle 11g DBMS, Oracle Intermedia, Oracle Text, PL/SQL Stored Procedures) for data management and repositories implementation and JAVA both for business and presentation logics.

Oracle Intermedia tools have been exploited, from one hand, to manage images, audio and video stored into the database with the related metadata, and from the other one, to implement the image similarity query. In particular, the oracle evaluateScore method has been used to implement an image distance

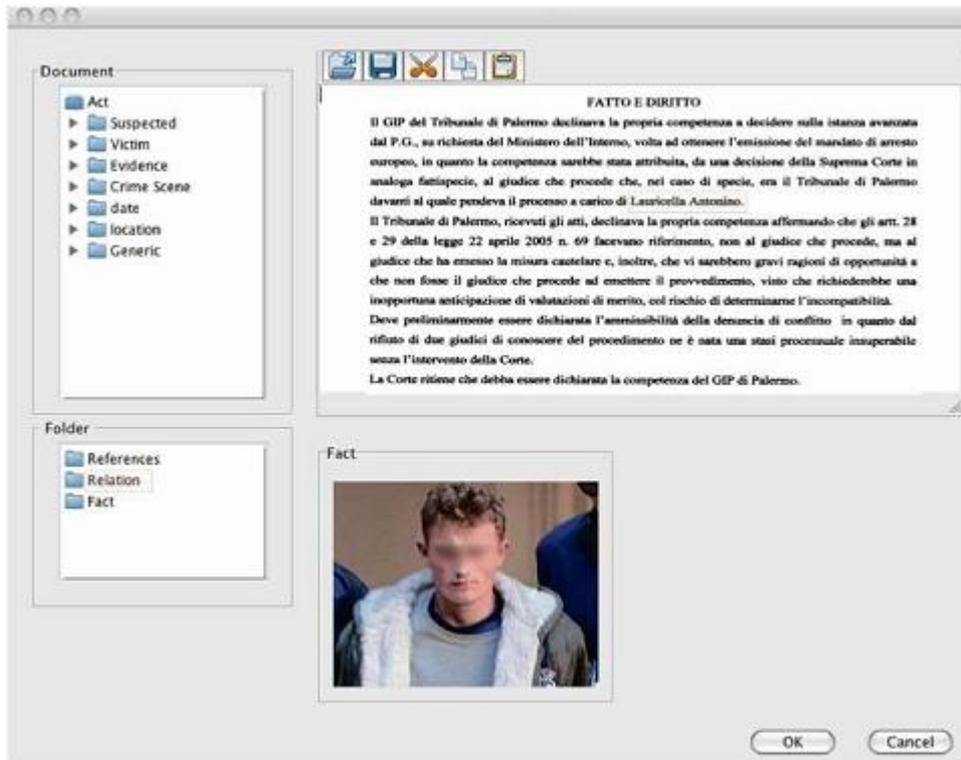


Figure 3. Interface for Information Extraction

through an apposite PL/SQL procedure. Oracle Text functionalities and ad-hoc PL/SQL procedures have been used to manage textual information and implement full-text search.

The ontologies are mapped in the oracle database and managed by the framework KAON 2, while the services of Ontology Query Service are implemented using SPARQL. Eventually, particular JAVA libraries have been exploited to implement Multimedia Information Processing module, Topics Detector, all user interfaces and the other modules. A couple of interfaces of the prototypical system are presented: in the fig 3 is reported the interface for information extraction features, in which the user is allowed to highlight the relation between a law text under analysis and an image that represent the person to which the content of the text segment references. In fig. 4, is showed the interface that allow the users to submit query to the system. The query are classified on the basis of the subject of interest, that for our domain are: the suspected, the victim, the crime scene and the evidences. In the example the user want retrieve all the acts in which the suspected is the person reported in the pictures that he inserted by the interface.

#### 4 CONCLUSIONS

In this paper we have described an e-Gov system based on a novel multimedia document model. The proposed RDF schema designed is the starting point for a variety of useful applications, in addition to the storage and retrieval facilities obtained. The RDF representation of documents is appropriate for the retrieval operations in the semantic web domain, as pointed out by a number of works in the literature. The system is designed for the management of document belonging to specialized domain. The restricted area of specialization reduces the intrinsic semantic ambiguity of the words, related at the generalist domain, allowing more accurate information extraction operations.



Figure 4. Interface for Information Retrieval

At the moment the ontologies concern to PA documents, so the ontology-driven procedure of extraction and retrieval are restricted to this kind of documents. Further works will be devoted to investigate the relationships between the model and still opened issues such as the definition of digital signature for multimedia and long term preservation.

## References

- F. Amato, A. Mazzeo, V. Moscato, A. Picariello: "Semantic Management of Multimedia Documents for E-Government Activity", Proceedings of 3rd International Conference on Complex, Intelligent and Software Intensive Systems (CISIS09), pp. 997-1002, 2009
- F. Amato, A. Mazzeo, A. Penta, A. Picariello, "Knowledge Representation and Management for E-Government Documents", Book Chapter of E-Government Ict Professionalism and Competences Service Science, pp.31-40, Springer Boston, 2008.
- F. Amato, A. Mazzeo, A. Penta, A. Picariello, "Using NLP and Ontologies for Notary Document Management Systems", in Proceedings of 19th International Conference on Database and Expert Systems Application, (DEXA), pp.67-71, 2008.
- G. Boccignone, A. Chianese, V. Moscato, and A. Picariello. "Context-sensitive queries for image retrieval in digital libraries.", Journal of Intelligent Information Systems, vol. 31, Issue 1, pp. 53-84, 2008.
- Breuker, J. "A functional ontology of law", Journal of Artificial Intelligence and Law, Vol. 7, pp.341-361. 1994
- P. Capasso, A. Chianese, V. Moscato, A. Penta, A. Picariello, "Automatic Categorization of Image Databases using Web Folksonomies", in Proceedings of IEEE International Symposium on Multimedia, Dicembre 15-17, Berkley (California, USA), 2008.
- C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blob world: image segmentation using expectation-maximization and its application to image querying", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 24, Issue 8, pp. 1026-1038, 2002.
- J.M. Corridoni, A. Del Bimbo, A., and P. Pala, "Image retrieval by color semantics", Multimedia Systems, vol. 7, n. 3, pp. 175-183, 1999.
- C. Colombo, and A. Del Bimbo, "Visible image retrieval", In L. Bergman and V. Castelli, eds., Image Databases, Search and Retrieval of Digital Imagery, Chapter 2, pp. 11-33, Wiley 2002.

- R. Datta, and D. W. J. Joshi, "Image retrieval: ideas, influence, and trends of the new age", *ACM Computing Survey*, vol. 40, n. 2, pp. 5-64, 2008.
- Deliberation of 13 dicembre 2001, n. 42, published on *Gazzetta Ufficiale della Repubblica Italianan*. 296 of 21 dicembre 2001
- J. S. Hare, and P. H. Lewis, "On image retrieval using salient regions with vector-spaces and latent semantics", *Image and Video Retrieval (CIVR 2005)*, Singapore, Springer Ed., 2005.
- J. R. Hobbs, D. Appelt, M. Tyson, J. Bear, and D. Israel, "Sri international: description of the Fastus system used for muc-4", *Fourth Message Understanding Conference*, Morgan Kaufmann, pp.143-147. 1992.
- Jacobs, P. and Rau, L. (1990) "Scisor: extracting information from on-line news", *Comm ACM*, Vol. 33, No. 11, pp.88-97.
- R. Klischewski, M. Jeenicke, "Semantic Web technologies for information management within e-Government services", *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*, vol., no., pp. 10 pp.-, 5-8 Jan. 2004.
- M. S. Lew, N. Sebe, D. Djeraba, and J. Rain, "Content-based multimedia information retrieval: State of the art and challenges", *ACM Trans. Multimedia Comput. Commun. Appl*, vol. 2, n.1, 2006 pp. 1-19.
- J. Li, and J. Z. Wang, "Real-time computerized annotation of pictures", *In Proc. ACM Int. Conf. on Multimedia*, pp. 911-920, 2006.
- B. S. Manjunath and et al. Cortina, "Searching a 10 million images database", *Technical report*, Sep 2007.
- L.T. McCarty, "A language for legal discourse i. basic features", *ICAIL '89: Proceedings of the 2nd International Conference on Artificial Intelligence and Law*, New York, NY: ACM, pp.180-189. 1989.
- S. Miike, E. Itoh, K. Ono, and K. Sumita, "A full-text retrieval system with a dynamic abstract generation function", *Proceedings SIGIR 94*, pp.152-161. 1994
- S.Santini, "Evaluation Vademecum for Visual Information Systems," *Proc. of SPIE*, vol. 3972, San Jose, USA, 2000
- A. Penta, A. Picariello, L. Tanca, "Towards a definition of an Image Ontology", in *Proceedings of DEXA Workshops*, pp. 74-78.2007.
- A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval at the End of the Early Years", in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, 2000, pp. 1349- 1379.
- Stamper, R. "The role of semantics in legal expert systems and legal reasoning", *Ratio Juris*, Vol. 4, No. 2, pp.219-244. 1991.
- L. Stojanovic, A. Abecker, N. Stojanovic, R. Studer, "On Managing Changes in the Ontology-Based E-Government" in *On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE*. 2004, pp 1080-1097.
- Tiscornia, D. "Some ontological tools to support legal regulatory compliance, with a case study", in *Workshop on Regulatory Ontologies and the Modeling of Complaint Regulations (WORM CoRe)*, Springer LNCS, November 2003.
- O. Udrea, V. S. Subrahmanian, Z. Majkic, "Probabilistic RDF" , in *Proceedings of the 2006 IEEE International Conference on Information Reuse and Integration, IRI*, pp. 172-177. 2006.
- P.R.S. Visser, T.J.M. Bench-Capon, "The formal specification of a legal ontology", in *Proceedings of JURIX-96*, Tilburg University Press, pp.15-24. 1996.
- J. Z. Wang, J. Li, and G. Wiederhold, "Simplicity: Semantics sensitive integrated matching for pictures libraries. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, n. 1, pp. 1- 16, 2001.

## Appendix A: RDF Data Model

```

<!DOCTYPE rdf:RDF [
  <!ENTITY owl "http://www.w3.org/2002/07/owl#" >
  <!ENTITY xsd "http://www.w3.org/2001/XMLSchema#" >
  <!ENTITY owl2xml "http://www.w3.org/2006/12/owl2-xml#" >
  <!ENTITY rdfs "http://www.w3.org/2000/01/rdf-schema#" >
  <!ENTITY rdf "http://www.w3.org/1999/02/22-rdf-syntax-ns#" >
  <!ENTITY Ontology1239005609406
    "http://www.semanticweb.org/ontologies/2009/3/Ontology1239005609406.owl#" >
]>

<rdf:RDF xmlns="http://www.semanticweb.org/ontologies/2009/3/Ontology1239005609406.owl#"
  xmlns:base="http://www.semanticweb.org/ontologies/2009/3/Ontology1239005609406.owl"
  xmlns:owl2xml="http://www.w3.org/2006/12/owl2-xml#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"

  xmlns:Ontology1239005609406="http://www.semanticweb.org/ontologies/2009/3/Ontology1239005609406.owl#"
  xmlns:owl="http://www.w3.org/2002/07/owl#">
  <owl:Ontology rdf:about=""/>

  <!--      // Object Properties      -->

  <owl:ObjectProperty rdf:about="#Contains">
    <rdfs:range rdf:resource="#Content"/>
    <rdfs:domain rdf:resource="#Document"/>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:about="#Equivalent">
    <rdf:type rdf:resource="&owl;ReflexiveProperty"/>
    <rdf:type rdf:resource="&owl;SymmetricProperty"/>
    <rdf:type rdf:resource="&owl;TransitiveProperty"/>
    <owl:equivalentProperty rdf:resource="#Synonymous"/>
    <rdfs:range rdf:resource="#Word"/>
    <rdfs:domain rdf:resource="#Word"/>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:about="#Hyperonym">
    <rdf:type rdf:resource="&owl;TransitiveProperty"/>
    <rdfs:range rdf:resource="#Concept"/>
    <rdfs:domain rdf:resource="#Concept"/>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:about="#Mean">
    <rdfs:range rdf:resource="#Concept"/>
    <rdfs:domain rdf:resource="#Word"/>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:about="#Refers">
    <rdfs:range rdf:resource="#Segment"/>
    <rdfs:domain rdf:resource="#Segment"/>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:about="#Synonymous">
    <rdf:type rdf:resource="&owl;ReflexiveProperty"/>
    <rdf:type rdf:resource="&owl;SymmetricProperty"/>
    <rdf:type rdf:resource="&owl;TransitiveProperty"/>
    <rdfs:range rdf:resource="#Word"/>
    <rdfs:domain rdf:resource="#Word"/>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:about="#areDetectable">
    <rdfs:domain rdf:resource="#Audio"/>
    <rdfs:domain rdf:resource="#Image"/>
    <rdfs:range rdf:resource="#LowLevelFeatures"/>
    <rdfs:domain rdf:resource="#Video"/>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:about="#canAccess">
    <rdfs:range rdf:resource="#Document"/>
    <rdfs:domain rdf:resource="#User"/>
    <owl:inverseOf rdf:resource="#isAccessibleBy"/>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:about="#hasAccessRight">
    <rdfs:range rdf:resource="#AccessRight"/>

```

```

<rdfs:domain rdf:resource="#User"/>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:about="#hasFormat">
    <rdfs:domain rdf:resource="#Document"/>
    <rdfs:range rdf:resource="#Format"/>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:about="#hasMimeType">
    <rdfs:domain rdf:resource="#Content"/>
    <rdfs:range rdf:resource="#MimeType"/>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:about="#hasProperty">
    <rdfs:domain rdf:resource="#Document"/>
    <rdfs:range rdf:resource="#DocumentProperty"/>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:about="#infers">
    <rdfs:range rdf:resource="#Concept"/>
    <rdfs:domain rdf:resource="#LowLevelFeatures"/>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:about="#isAccessibleBy">
    <rdfs:domain rdf:resource="#Document"/>
    <rdfs:range rdf:resource="#User"/>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:about="#isContained">
    <rdfs:domain rdf:resource="#Document"/>
    <rdfs:range rdf:resource="#Folder"/>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:about="#isStructured">
    <rdfs:domain rdf:resource="#Content"/>
    <rdfs:range rdf:resource="#Segment"/>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:about="#ofType">
    <rdfs:domain rdf:resource="#Content"/>
    <rdfs:range rdf:resource="#Media"/>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:about="#sequent">
    <rdfs:range rdf:resource="#Segment"/>
    <rdfs:domain rdf:resource="#Segment"/>
  </owl:ObjectProperty>

  <owl:ObjectProperty rdf:about="#textContain">
    <rdfs:domain rdf:resource="#Text"/>
    <rdfs:range rdf:resource="#Word"/>
  </owl:ObjectProperty>

<!-- // Classes -->

<owl:Class rdf:about="#AccessRight">
  <rdfs:subClassOf rdf:resource="#owl:Thing"/>
</owl:Class>
<owl:Class rdf:about="#Audio">
  <rdfs:subClassOf rdf:resource="#Media"/>
</owl:Class>
<owl:Class rdf:about="#Concept">
  <rdfs:subClassOf rdf:resource="#owl:Thing"/>
</owl:Class>
<owl:Class rdf:about="#Content">
  <rdfs:subClassOf rdf:resource="#owl:Thing"/>
</owl:Class>
<owl:Class rdf:about="#Document">
  <rdfs:subClassOf rdf:resource="#owl:Thing"/>
</owl:Class>
<owl:Class rdf:about="#DocumentProperty">
  <rdfs:subClassOf rdf:resource="#Property"/>
</owl:Class>

```

```
<owl:Class rdf:about="#Folder">
  <rdfs:subClassOf rdf:resource="#owl:Thing"/>
</owl:Class>
<owl:Class rdf:about="#Format">
  <rdfs:subClassOf rdf:resource="#Property"/>
</owl:Class>
<owl:Class rdf:about="#Image">
  <rdfs:subClassOf rdf:resource="#Media"/>
</owl:Class>
<owl:Class rdf:about="#LowLevelFeatures">
  <rdfs:subClassOf rdf:resource="#owl:Thing"/>
</owl:Class>
<owl:Class rdf:about="#Media">
  <rdfs:subClassOf rdf:resource="#owl:Thing"/>
</owl:Class>
<owl:Class rdf:about="#MimeProperty">
  <rdfs:subClassOf rdf:resource="#Property"/>
</owl:Class>
<owl:Class rdf:about="#Property">
  <rdfs:subClassOf rdf:resource="#owl:Thing"/>
</owl:Class>
<owl:Class rdf:about="#Segment">
  <rdfs:subClassOf rdf:resource="#Content"/>
</owl:Class>
<owl:Class rdf:about="#Text">
  <rdfs:subClassOf rdf:resource="#Media"/>
</owl:Class>
<owl:Class rdf:about="#User">
  <rdfs:subClassOf rdf:resource="#owl:Thing"/>
</owl:Class>
<owl:Class rdf:about="#Video">
  <rdfs:subClassOf rdf:resource="#Media"/>
</owl:Class>
<owl:Class rdf:about="#Word">
  <rdfs:subClassOf rdf:resource="#owl:Thing"/>
</owl:Class>
<owl:Class rdf:about="#owl:Thing"/>
</rdf:RDF>
```