Multimedia data integration and processing for E-government

Flora Amato*, Francesco Colace†, Luca Greco†, Vincenzo Moscato* and Antonio Picariello*
*Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione. University of Naples ”Federico II”, ITALY
Email: {flora.amato,vmoscato,picus}@unina.it
†Dipartimento di Ingegneria dell'Informazione, Ingegneria Elettrica e Matematica Applicata. University of Salerno, ITALY
Email: {fcolace,lgreco}@unisa.it

***Abstract*—Knowledge management has become a challenge for almost all E-government based applications where one of the main problem is the efficient management of great amounts of data. In order to efficiently access the information embedded in very large document repositories, techniques for semantic document management are required. They ensure improvement for a large and intense process of dematerialization and aim at eliminating or at least reducing, the amount of paper documents.**

In this work, we present a novel model of digital documents for the improvement of the dematerialization effectiveness. This model represents the starting point for an information system that is able to manage the document streams in an efficient way. It takes into account E-government applications needs like the compliance with the laws and regulations in force and the adaptability to evolving technologies. At the best of our knowledge, the proposed model is one of the first attempts to give a single and unified characterization for the management of multimedia documents, pertaining to a bureaucratic domain as the E-government one, on which semantic procedures are used for the transformation of non structured documents (pertaining to specialized domain) into structured data, suitable for automatic processing.

Furthermore, an architecture for the management of documents life cycle is proposed, which provides advanced functionalities for semantic processing, such as giving formal structure to document informative content, information extraction, semantic retrieval, indexing, storage, presentation, together with long-term preservation.

***Keywords*-Ontology Learning, Ontology Population, Natural Languages Processing**

I. Introduction

E-Government processes are dedicated to the improvement of the efficiency, expensiveness and accessibility of public administration services: dematerialization activities, introduced for properly managing bureaucratic documents, are among the main tasks of the E-government works.

The core aspect related to a novel and efficient dematerialization process is the idea standing beyond the common document concept, that can be defined as the representation of acts, facts and figures directly made or by means of electronic processing, and stored onto an intelligible support[1].

In other words, a 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 documents.

During the various E-government processing phases, which differs depending on applications domains, a document is processed and eventually stored on various kinds of media, as papers and photographic films.

In order to manage documents properly, Document Management Systems (DMS) are used. They were introduced in the early 1970 for converting paper documents into electronic images stored in computers.

Nowadays DMS are becoming the basis of most business information systems, giving user control over company knowledge, providing efficient retrieval and desktop integration, reducing error rates in documents manipulation and thus improving business performance.

With the use of standards for knowledge representation, DMS are evolving, from search engine, toward systems able to integrate semantic search procedures into companies business processes. These systems, however, are limited to provide additional semantic functionalities to existent document management features. At the best of our knowledge, there are a variety of semantic-based approaches to modeling multimedia content focusing on single media (e.g. images or sounds only) but exist only a few experiments[3] for processing more complex multimedia documents such as those dealt with in this work.

The aims of this semantic-based processing are to structure input documents and to allow automatic retrieval of targeted information, based on formal representation of the domain associated to the documents, defined in a semi-automatic way, starting from the processable documents themselves.

In this work, we propose a new model of multimedia document, suitable for E-government activities, that takes into account the requirements of the E-government applications, which, depending on authorities, final users or time, produce different representations of the same multimedia contents. For describing the proposed model and system architecture, we focused on the E-Health domain. This particular domain implies a proper massive document processing. Knowledge management activities, In particular, must be performed

[1] This definition accords, for example, to the Italian civil law [1]

DOI reference number: 10.18293/DMS2015-019

in reliable, effective and error-free way. Hence, E-Health organizations needs to be supported with approaches aimed at assessing clinical guidelines, and supporting their correct and efficient execution. The reported examples, in particular, are taken from the sub-domain of the Electronic Clinical Records. According to the International Organization for Standardization (ISO) definition, an electronic clinical record means a repository of patient data in digital form, stored and exchanged securely, and accessible by multiple authorized users. It contains retrospective, concurrent, and prospective information, and its primary purpose is to set objectives and planning patient care, document the delivery of care and assess the outcomes of care [2].

For this reason we model presentation and informative content in a separate way, allowing the solution, among other things, of open problems related to technology evolution, different document formats and access rights. The proposed model is the starting point for an information system which integrates and processes, in the most efficient way, different multimedia data types (like as images, text, graphic objects, audio, video, composite multimedia, etc.). In particular, it allows: *i)* documents structuring *ii)* automatic information extraction from digital documents; *iii)* semantic retrieval; *vi)* semantic interpretation of the relevant information presented in the document, *v)* storing and *vi)* long term preservation.

The proposed system combines Object-Relational Database (ORDBMS) technologies, Natural Language Processing (NLP) techniques, proper domain and structural ontologies, and inference rules in order to automatically extract significant concepts instantiated to each document and to provide semantic querying facilities for users.

In the process for representation and use of domain, specific knowledge ontologies play an important role, helping to cope documents with metadata annotations for supporting the process of information structuring and retrieval.

The quality of the information retrieved is thus improved by exploiting the possibility to enrich and then refine the set of the retrieved documents by using reasoning techniques on the ontologically-defined relations[4].

The work is organized as follow: in the next section, an overview on Knowledge Modeling Methodologies is presented. In the third section the method for semantic processing implemented in the proposed system are introduced. In section 4 we report the preliminary experimental results and in section 5 present the related works together with a discussion of the original contribution of our proposal. Finally in section 6 we give our concluding remarks and we outline our future work.

II. Related Works

First, we briefly report the state of the art on the Systems developed for the Document Management and then we focused on the system managing multimedia documents.

Starting from the 1980s, a number of vendors began to develop systems to manage paper-based documents. They include the management not only of printed and published documents, but also of photos, prints, etc. Recent Document Management Systems (DMS) are dedicated to the management of digital documents. This kind of systems commonly provides facilities for document processing as storage, versioning, metadata, security, as well as indexing and retrieval capabilities. In recent years numerous DMS projects suitable for specialist domains have been realized. These systems propose features for content characterization, offering for example, templates for documents semi-automatic generation. Nowadays DMS are moving toward semantic functionality, including advanced features for contents management like semantic search.

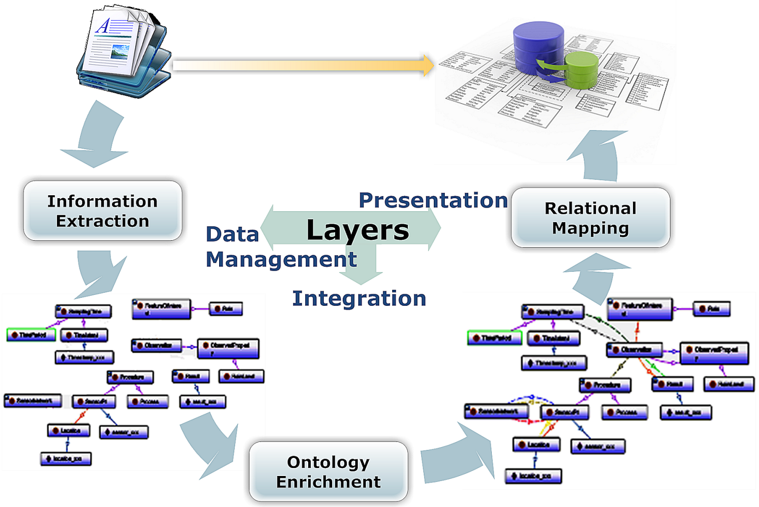

Fig. 1. General Schema of Documents Processing

In Italy, numerous projects are presented for several specialist domains as the ASTREA Project realized by the Judicial Systems Research Institute (IRSIG) for the CNR (National Research Centre) during the years 2002-2006 and TAPA project realized in 2004 for the Anti-trust Authority (AGCM). Another relevant experience to be mentioned is the ESTRELLA project (European project for Standardized Transparent Representations in order to Extend Legal Accessibility), financed by the European Union (2006-2008). For what concerns the state of art in multimedia information management system one of the main research objective is the automatic indexing of multimedia data on the basis of their content in order to make query processing easier, more effective and efficient. In the following, supported by the related state-of-the-art, we describe the major challenges in developing reliable image and text database systems. The goal of image retrieval systems is to find out images from databases while processing a query provided by a user. 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 [7] 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. Two main approaches have been proposed to cope with this deficiency: the first approach segments the image into multiple regions, and different descriptors are built for each region; the second approach exploits salient points identification techniques. Following the first approach, different systems like, SIMPLIcity and Blobworld [8] 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 [9]. Our proposal [4] 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. Finally, more recent systems, such as Cortina and ALIPR [12] have as goal the automatic classification of images on the base of low-level features and high-level human annotations.

The textual processing phase requires the use of different techniques from interdisciplinary fields. Both theoretical and applicative fields have to be considered: the first for defining lexical dictionaries for legal domain, the second for organizing, storing and retrieving information of interest[11].

III. Knowledge Modeling

In order to manage the composition of different multimedia data, their semantic relations and the structure imposed for bureaucratic documents, the defined document model is divided in three levels, as described in the following:

- **Data Management Layer**: describes the semantic content of each single media element (such as a text fragment or an image), providing functionalities for working on a single media; for example, information extraction and indexing over text are performed in this layer.
- **Integration layer**: describes the relations between the heterogeneous media components of the document, providing functionality for the integration of different data sources.For example the property of a text fragment of referring to an image belongs to this layer.
- **Presentation layer**: regulates the way by which the information has to be shown to users. It provides different representations of the same informative content, according to the formats, the final user's access rights and the available technology.

This approach allows the management of heterogeneous contents, by working on form and content independently, enabling solutions of open problems related to technology evolution: in order to give a concrete example, it permits to give an immutable legal validity to the content of a document even if the format of representation changes, evolving with technology. On different layers of the document, information is semi-automatically tagged according to the concepts contained in the *domain ontologies*: associations among concepts and their instances are picked out. A general schema of documents processing is depicted in Fig. 1. Different ontologies can be used for the tagging process according to the different domains of interest. **Domain Ontology** is exploited to formalize the concepts of interest in the reference domain and relationships among them.

An example of top-level fragment of ontology in the domain of E-Health is depicted in Fig. 2, showing the relevant concepts and the semantic associations among them, occurring in a medical record. Specialized Domain ontology [13] can be divided into: **Structure Ontology** that describes how information is organized within the document and models the associations between the internal sections of the document and the set of concepts that can be found in it, and **Lexical Ontology** that contains the terms of the general language and can be used to refer wide-ranging concepts presented in the documents, not enclosed in the domain of reference.

A. System Architecture

Starting from the model, we have proposed an architecture, implemented in a prototype, for the management of the medical records life cycle. As already stated, medical records contain text that can be supplied with multimedia information as pictures (e.g.in radiographies), video streaming (e.g. in echographies) and audio. It is composed of three main modules: one for the text processing, one for the processing of the other media typologies of data, one for the management of the different formats of presentation, according to the requirements of the E-government applications. The modules that compose the system architecture, depicted in fig. 3 are described in the following: **Text processing** module aims at extracting relevant information from text, associating concepts to the terms of the text and defining relationships between them. The text is processed by a series of procedures each of which produces information usable by other procedures [14]:

- **Structural Analysis:** performs text segmentation and the related classification in order to identify the different sections constituting the structure of the document.
- **Linguistic Analysis:** performs procedures of Morpho-Syntactic analysis on the text (such as text tokenization and normalization, Part-of-Speech Tagging and lemmatization, complex terms analysis) combined with statistic procedures (such as the computation of opportune indices) enabling the extraction of relevant terms from the corpus to process. These terms and the information about them, refined with the help of domain experts, constitutes a lexicon that is exploited for the building

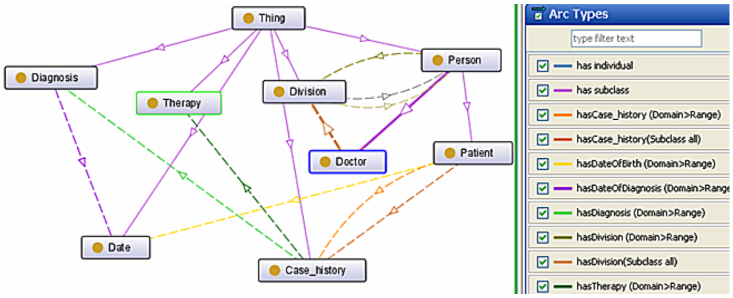
Thing
Diagnosis
Therapy
Division
Person
Doctor
Patient
Date
Case_history
Arc Types
type filter text
☑ has individual
☑ has subclass
☑ hasCase_history (Domain>Range)
☑ hasCase_history(Subclass all)
☑ hasDateOfBirth (Domain>Range)
☑ hasDateOfDiagnosis (Domain>Rang
☑ hasDiagnosis (Domain>Range)
☑ hasDivision (Domain>Range)
☑ hasDivision(Subclass all)
☑ hasTherapy (Domain>Range)

Fig. 2. A Fragment of Domain Ontology for electronic medical record

of the set of concepts used for domains formalization, performed by using ontologies.

- **Semantic Analysis:** by using the information of the early analyses, it detects properties and associations among terms, defining the concepts and relationships, allowing for ontology building and documents annotation.

The **Multimedia Data Processing** module has the aim of classifying each multimedia element, associating concepts from the domain ontology. It is composed of two components implementing innovative methods that have been presented in recent works [4][10]:

- **Analyzer** identifies relevant media parts and produces a low-level description that permits to create a series of indices to help the tagging and th retrieval.
- **Classifier** uses the indexing information to deduce which concepts, from the domain ontology, are being associated to media elements.

The **Presentation** module performs the dual task of combining the information about the heterogeneous contents and managing the ways by which they are presented to different users, according to the policies of the Entity (as the Public Administration), the final user's access rights and the available technology.

This module has also the aim to map the semantic information in a standard data format. For example, in E-Health domain, the module aims to map in Standard HL7 format the data semantically enriched with information about concepts and the implicit and explicit relations among them, coded in RDF triple. The module wraps the RDF data sets, translating the file into an XML based document, according to HL7 specifications, it works applying one or more XSLT transformation, according to the HL7 data format.

The list of the XSLT transformation rules is downloadable from our Document Processing project web site[2].

The system is based on a multimedia database management system:it supports different multimedia data types (e.g. images, text, graphic objects, audio, video, composite multimedia, etc.) and, in analogy with a traditional DBMS, facilities for the indexing, storage, retrieval, and control of the multimedia data, providing a suitable environment for using and managing multimedia database information.

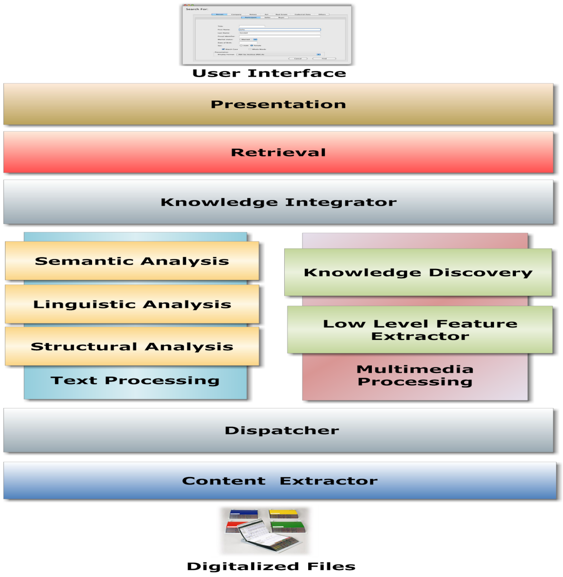

Fig. 3. System Architecture

More in details, a MMDBMS meets certain requirements that are usually divided into the following broad categories: multimedia data modelling, huge capacity storage management, information retrieval capabilities, media integration, composition and presentation, multimedia query support, multimedia interface and interactivity, multimedia indexing, high performances and distributed multimedia database management.

All document management system applications are designed on the top of a MMDBMS in order to support E-government processes in a more efficient way, in particular for those tasks regarding: automatic information extraction from documents, semantic interpretation, storing, long term preservation and retrieval of the extracted information.

The architecture of the proposed MMDBMS system, shown in Fig. 3, can be considered a particular instance of the typical MMDBMS architectural model and it is a suitable support for the management of E-government documents. The main components of the system are the modules delegated to manage the *Information Extraction and Indexing* process and those related to *Retrieval and Presentation* applications. All the knowledge associated to E-government documents is managed by proper *ontology repositories*.

In the current implementation of the system we realized three main separate subsystems that are responsible for the information extraction and the presentation tasks: one for the text processing related to e-doc, another one for processing the other kinds of multimedia information, in particular images, and the last one for presentation aims according to the requirements of public administrations.

The multimedia indexing and information extraction modules can be also specialized for other kinds of multimedia data like audio and video. In this case ad-hoc preprocessing

[2] http://www.unina.it/DOCProject.html

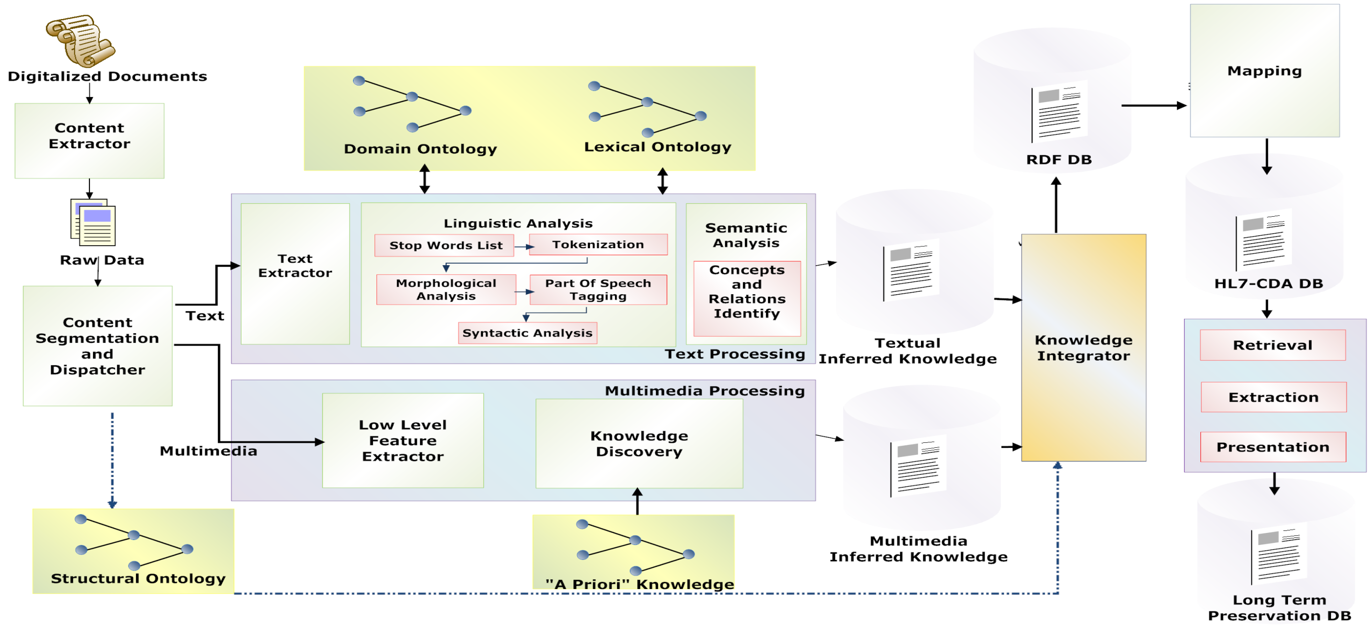

Fig. 4. Semantic Document Processing

components able to perform a *temporal segmentation* of multimedia flow are necessary to efficiently support the indexing process.

B. Proposed Multimedia Document Processing Methods

The whole process of document management, performed by the designed architecture, can be divided in Domain formalization and Final users application.

Domain formalization stage has the aim of codifying, with proper data structures (ontologies) the information of interest pertaining to the domain which the documents belong to. The information associated to contents is codified in terms of relevant concepts and relationships between them. Final users application stage implements the functionalities of document processing offered to the users in order to perform automatic operations on documents, such as searches by contents, long-term preservation and information representation according to different formats and different access policies.

1) Information Extraction and Ontology Population : Once associations between document segments and ontology fragments have been resolved, we proceed in populating concepts and relationships in the ontology fragment, by adding proper instances detected in document segments. Relevant information are then extracted, document segments are annotated and results are presented in RDF triples containing the properties identified in the segments.

Concepts and relations are extracted by exploiting an inference mechanism performed by a Rule-Based System. A generic rule is formed by a combination of token and syntactical patterns, which codifies the expert domain knowledge. In order to derive instances of relevant concepts or relationships, rules exploit:

- Named Entity Recognition (NER) functionality
- Morpho-Syntactic information obtained from NLP procedures performed in the Lexical Analysis,

eventually using subsumption on $TBox$-$Module$ for deriving more specific concepts.

The detected instances can be shown by using tools like KIM[5], that highlights the associations among detected instances and the concept defined in the domain ontology.

The extracted relevant information is presented in RDF triples.

2) Information Retrieval : Once relevant information related to the domain of interest has been codified for document corpus, it is possible to execute semantic-based searches which are able to retrieve information by contents and not only by key-words.

The system we propose combines ORDBMS technologies, NLP techniques, proper domain structural ontologies management, and inference rules in order to retrieve significant concepts related to each document and to provide extended querying facilities for users. In particular, one of these facilities is the ability to perform advanced searches that overcome the limit imposed by “keyword-based” traditional queries. It also allows for a “content-based” access to documents database.

Traditional information retrieval systems, based on the comparison of sequences of characters, are in fact able to identify relevant concepts only if they are expressed with the same terms within the text: the search is always limited to the specific key-words inserted into the query and it excludes all the text parts where those keywords do not specifically appear. For instance, when searching for the word “house”, the system will ignore the documents where the words “home” or “residence” appear, even if they represent, in many contexts, the same concept. We exploit, thus, semantic characterization of the document content, in order to improve the quality of the information retrieval. The domain specific knowledge is represented by means of Ontologies, that contain concepts and relationships. Instances of such elements are indicated in the documents by means of semantic annotations, performed by information extractions procedures.

When a user submits a query, the system identifies the concepts associated to the terms used in the query. These concepts are represented by means of ontologies as *synsets*, which are the set of linguistic elements linked by a synonymy relationship, i.e. terms that can be used in the same statement without modifying its whole meaning. Furthermore, same terms can be used with different acceptations (the meaning in which a word or expression is understood). In this case, different synsets are related to different meanings. If these ambiguities are present, the system will provide features to discriminate the synset of interest in the search.

Once users have selected the desired synset (all synsets are chosen if no selection is specified) a *query expansion*[6] mechanism is used in order to perform queries on corpus where all lemmas in the selected synsets become lemmatized keywords for a text-based search.

Query expansion techniques are used for dealing with the problem of word mismatch in information retrieval: retrieval system users and authors, in fact, often use different words to describe the same concepts in documents. The adopted query expansion approach requires that the query is expanded using lemmatized terms with the same meaning of the words used in the query. Thus, words within the same synsets are used for expansion and the match is not performed between single terms but between list of terms, which concern the concept to be retrieved in documents.

The collection of all the documents retrieved from these searches constitutes the results of the semantic-based query. A ranking algorithm is used to score results depending on a similarity measure, based on Tf-Idf index evaluation.

Notice that all query words and all relevant terms present in documents (which are also used for indexing purposes), have been reduced to their lemma, in order to make the search independent from different declinations and conjugations.

IV. Implementation of the multimedia document processing

We implemented a prototypal version of the system that realizes the described data management procedures.

The proposed *Multimedia Document Management System* has the following main features:

- it exploits a unified data model that takes into account content-based and document-based characteristics;
- it uses ontological support for managing the semantics of data;
- it has a multi-layer architecture with different kinds or user interfaces;

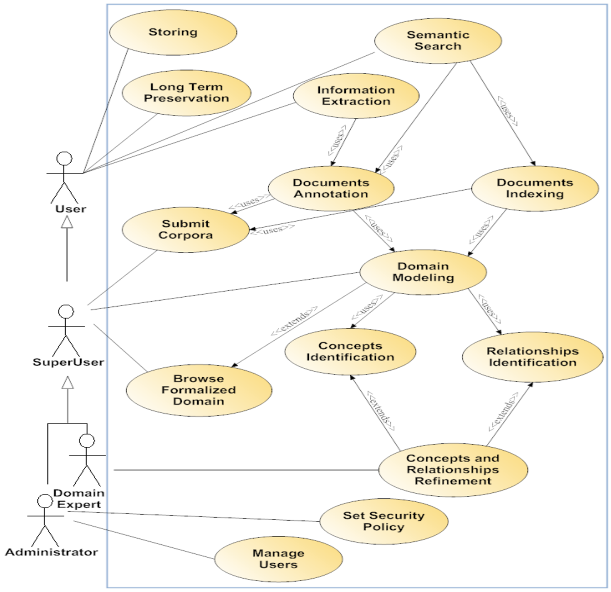

Fig. 5. Use Case Diagram of System Functionalities

- it provides advanced functionalities for document indexing and semantic retrieval. The system features are depicted in the user diagram in Fig. 5, in which the different functionalities (described in the previous paragraphs) are accessible only to the user with the appropriate privileges. Users can query the system, performing searches by content and information extraction, and use storing functionalities. Super-Users can also submit new medical documents or integrate the existent one, on which starting the process of domain modeling. Domain Experts can refine the modeled knowledge and the Administrators can manage users' proprieties and security policies too.

Fig. 6 shows at glance the Component 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 for 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 easy management and retrieval.

According to the introduced data model, it is possible to associate a digital document with 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 whole 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 analyses digital documents

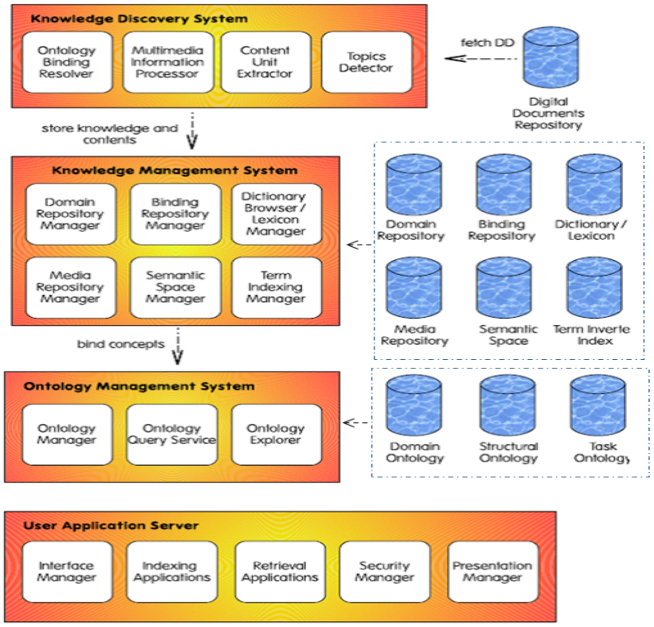

Fig. 6. Component Architecture

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[4] (e.g. text, audio, video, image). Furthermore, a *Topics Detector* sub-module operates on the not-structured view of a document and aims at detecting by natural language processing the most relevant topics for the whole 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 features for the browsing and the retrieval of the 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 (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 lexicon is managed by a proper 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 terms with the related frequency, is created. The inverted index is managed by a *Term Indexing Manager*.
- *Semantic Space* - It allows for the storage of atomic pieces of knowledge belonging to document content units, which are 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 to a particular structural ontology that allows for relating segments of the same content unit to 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 it is managed by a *Domain repository Manager*.
- *Binding Repository* - It contains the associations between document and domain repository concepts and it is managed by a *Binding Repository Manager*.
- *Media Repository* - It is an Object Relational DBMS able to manage different kinds of multimedia contents. It is managed by a particular module, called *Media Repository 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 semantics 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 features provided by the system are the *indexing* of documents and the *semantic retrieval* of information. The application interfaces are realized both as web services and desktop programs (and managed by an *Interface Manager*). Finally, two modules are provided for the *security* and the *presentation* management.

V. Preliminary Experimental Results

In this section we report some experiments we have carried out for evaluating the impact of the proposed system on enhancing user effort in indexing about 10000 medical records, properly anonymized, coming from an Italian health care organization. To set up our experimentation, we chose a sub-set of the collected data (constituted by 2000 randomly chosen documents) as training set for training the classifier used for text segmentation. The objective in this experimentation is to evaluate the system correctness (precision) in automatically discovering relevant concepts of a medical document and in particular:

(1) Personal Data, ***(2)*** Diagnosis, **(3)** Diary of significant events, ***(4)*** Hospital discharge.

Relevant concepts discovery procedures exploit a domain ontology built from scratch from the medical records dataset, with the help of domain experts. Table 1 shows the related results and in particular the num- ber of documents that has a given value of precision (100%: all the concepts have been correctly discovered, 75%: three concepts have been correctly discovered, 50% two and 25% only one concept has been correctly discovered.

In the majority of cases for which precision is 50% correct relevant concepts are the Personal Data and Hospital discharge, thus in our approach the most difficult concept to discover is that related to diary of significant events, probably due to the fact that such diaries are written in free text, also by different categories of medical users.

Eventually, table shows, on the right side, the average indexing times with respect to the document size[3].

Precision	**Documents**
100%	984
75%	2024
50%	3713
25%	2498
0%	781

Doc. size	**Indexing Time**
<150K	1, 2 s
150K ∼ 300K	1, 8 s
300K ∼ 500K	2, 5 s
500K ∼ 1000K	2, 9 s
>1000K	4, 8 s

TABLE I
INDEXING PRECISION AND INDEXING TIMES

VI. CONCLUDING REMARKS

In this work, we have defined a novel system for automatic processing of documents, based on semantic technologies. The realized semantic-based functionalities, as well as search by contents and information extraction, are based on the modeling of the relevant information of the domain of interest, codified by ontologies. Even if it is possible to provide as input data structures containing significant information, for example in form of lexicon for refinement purpose, the proposed system is able to define a formal representation for the domain of interest, in terms of concepts and relationships. The domain representation is built on the basis of the documental corpus, analysed in the early domain formalization phase. The formalization procedure is semi-automatic, because domain expertise can be exploited in order to refine ontologies, automatically built in a previous stage. The system, intended to be the core of an E-government information system, exploits the use of Linguistic and Semantic Analysis in order to transform unstructured (or semi-structured) documents into structured, automatically processable records, codified by RDF triples. The system is designed for the management of documents belonging to specialized domains; the restricted area of specialization reduces the intrinsic semantic ambiguity of the words, related to the generalist domain, allowing more accurate information management operations. In order to perform semantic based document processing, we have defined a model for multimedia digital document, particularly suitable for processing data from E-government activities. The model is a starting point of a general framework for structuring, presenting and retrieving relevant information for a a specialized domain. Experimental results (not reported for brevity) have shown encouraging results. Future direction will be devoted to improve the interoperability among the available procedures.

[3] All experiments presented in this Section were conducted on a Linux Cluster of 3 machines, each one mounting a 2GHz Intel Core i7 processor with a 8 GB, 1600 MHz DDR3

REFERENCES

[1] Deliberation of 13 dicembre 2001, n. 42, published on Gazzetta Ufficiale della Repubblica Italiana n. 296 of 21 dicembre 2001

[2] Colantonio, S., Esposito, M., Martinelli, M., De Pietro, G., Salvetti, O. (2012).“A knowledge editing service for multisource data management in remote health monitoring”. IEEE Transactions on Information Technology in Biomedicine, 16(6), 1096-1104.

[3] Khoo, Michael, et al. ”Towards digital repository interoperability: The document indexing and semantic tagging interface for libraries (distil).” Theory and Practice of Digital Libraries. Springer Berlin Heidelberg, 2012. 439-444.

[4] Colace, F., De Santo, M., Greco, L., Moscato, V., Picariello, A. (2015). “A collaborative user-centered framework for recommending items in Online Social Networks”. Journal of Computers in Human Behavior. 2015. Doi : http://dx.doi.org/10.1016/j.chb.2014.12.011.

[5] B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, M. Goranov. “KIM – Semantic Annotation Platform”. Book Chapter of The SemanticWeb - ISWC (2003). pp. 834 – 849. ISBN 978–3–540–20362–9–. Springer Berlin / Heidelberg.

[6] Z. Jiuling , D. Beixing ,L. Xing , Concept Based Query Expansion Using WordNet, pp. 52-55, 2009 International e-Conference on Advanced Science and Technology, 2009.

[7] 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.

[8] 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.

[9] 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.

[10] Chianese A. and Piccialli F. “Designing a smart museum: when Cultural Heritage joins IoT.” Next Generation Mobile Apps, Services and Technologies (NGMAST), 2014 Eighth International Conference on. IEEE, 2014.

[11] Chianese, Angelo, and Francesco Piccialli. “SmaCH: A Framework for Smart Cultural Heritage Spaces.” Signal-Image Technology and Internet-Based Systems (SITIS), 2014 Tenth International Conference on. IEEE, 2014.

[12] B. S. Manjunath and et al. Cortina, “Searching a 10 million images database”, Technical report, Sep 2007.

[13] F. Amato, A. Mazzeo, V.Moscato, A. Picariello. “A System for Semantic Retrieval and Long Term Preservation of Multimedia Documents in E-Government Domain”. To Appear in International Journal of Web and Grid Services, Vol. 5, No. 4, Inderscience Publishers, pp. 323.338(16), 2009.

[14] F. Amato, A. Mazzeo, A. Penta, A. Picariello, “A semantic document management system for legal applications”, International Journal of Web and Grid Services, Vol. 4, No. 3, Inderscience Publishers, pp. 251–266(16), 2008.