pre-print submitted for JVLC special issue
on DMS2015

https://doi.org/10.1016/
J.JVLC.2015.10.012

DMS2015short-19: Semantic processing of multimedia data for e-government applications

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 applications where the efficient processing of large amounts of data is still a critical issue. In the last years, semantic techniques have been introduced to improve the full automatic digitalization process of documents, in order to facilitate the access to the information embedded in very large document repositories. In this paper, we present a novel model for multimedia digital documents aiming at improve effectiveness of digitalization activities within an information system supporting e-government organizations. At the best of our knowledge, the proposed model is one of the first attempts to give a single and unified characterization of multimedia documents managed by e-government applications, whereas semantic procedures and multimedia facilities are used for the transformation of unstructured documents into structured information. Furthermore, we define an architecture for the management of multimedia documents "life cycle", which provides advanced functionalities for information extraction, semantic retrieval, indexing, storage, presentation, together with long-term preservation. Preliminary experiments concerning an e-health scenario are finally presented and discussed.**

***Keywords*-Ontology Learning, Natural Languages Processing, Information Systems, Multimedia**

I. Introduction

E-government (e-gov) activities are devoted to improve efficiency, expensiveness and accessibility of public administration services: *digitalization* is surely one of the most important tasks within this kind of context.

Indeed, the core aspect related to an effective digitalization process is the idea standing beyond the common document concept: it can be defined as "the representation of acts, facts and figures directly made or by means of electronic processing, and stored into an intelligible support"[1]. In other words, a document can be seen as a set of multimedia objects (e.g. text and images) that, according to their relative positions wtihin the support, determines the shape and, consequently the structure of the document.

In addition, during the various processing phases, depending on the particular application domain, a document is computed and eventually stored into different kinds of media.

[1]This definition accords with the Italian civil law [1].

In order to properly manage documents, *Document Management Systems* (DMS) have been introduced. Initially, they were used for converting paper documents into electronic images. Nowadays, DMS are becoming the basis of the majority of information systems, giving the users an access to "company knowledge", providing efficient and effective retrieval, reducing error rates in documents manipulation and thus improving overall business performances.

With the advent of *Semantic Web* and the adoption of standards for knowledge representation, DMS evolved from simple search engines, towards more complex systems able to integrate semantic technologies for information extraction and retrieval. Such systems, however, are limited to provide additional semantic functionalities to existent document management features.

In the literature, there are a variety of semantic-based approaches to model multimedia content focusing on single type of media, but there exist only few proposals [2] for processing more complex multimedia documents as required by e-gov applications. Generally speaking, the main goal of a semantic-based processing is to structure input documents and to allow automatic retrieval of targeted information on the base of a formal representation of the related domain.

In this paper, we propose a new model of multimedia documents that meet with the specific requirements of e-gov applications, allowing a semantic processing of the related digital contents that can be exploited from various perspectives such as presentation, indexing, integration, storage, retrieval and so on. In particular, our model 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.

From a technical perspective, 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 from each document (document annotation) and to provide semantic querying facilities [2] (retrieval is improved by "enriching and then refining" the set of the retrieved documents by using reasoning techniques on the ontological relations).

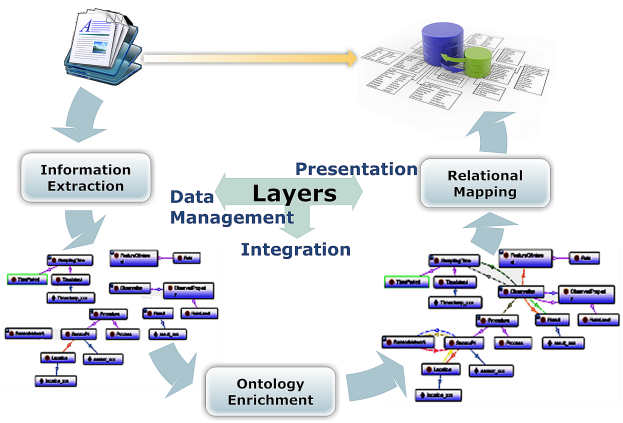

Fig. 1. General Schema of Documents Processing

We consider the *e-health* domain as a suitable case study: it implies a massive document processing that must be performed in reliable, effective and error-free way. In particular, we focus our attention on the semantic processing of *Electronic Clinical Records* [2] that, as well known, can contain several types of multimedia contents.

The paper is organized as in the following. Section 2 reports a brief review describing the main DMS and multimedia information management technologies. Section 3 describes the proposed framework for manage multimedia documents. Section 4 outlines some implementation details for our system, while Section 5 presents some preliminary experimental results for e-health domain. Finally, Section 6 discusses conclusions and future work.

II. Related Works

Starting from the 1980s, a number of vendors began to develop systems to manage paper-based documents. More recently, Document Management Systems (DMS) have been dedicated to the management of digital documents, providing a set facilities for document processing as storage, versioning, metadata management, security, as well as indexing and retrieval capabilities. Nowadays, DMS are evolving to integrate semantic functionalities, including advanced features for contents management like semantic search (as EUNOMOS, a knowledge management system for legal documents).

In the last years, numerous projects for document management in several specialist domains are presented, as the ASTREA Project realized by the Judicial Systems Research Institute (IRSIG), the TAPA Project realized for the Antitrust Authority (AGCM) and the ESTRELLA Project (European project for Standardized Transparent Representations in order to Extend Legal Accessibility) financed by the European Union.

They combine several NLP and machine learning techniques to extract structured information form data (text) and Semantic Web technologies to support semantic retrieval.

Concerning, 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 particular, the major challenges in developing reliable image database systems lies in the capability of such systems in extracting relevant information on the base of image visual content and semantics expressed by means of simple attributes (metadata), tags or keywords.

Traditionally, the problem of finding relevant images to the users on the base of visual content is solved using low-level image global descriptors (color, texture and shape features) for which automatic extraction methods are available, see [5] 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 [5]; the second approach exploits salient points identification techniques[6]. Finally, more recent systems [9] have as goal the automatic classification of images on the base of low-level features and high-level human annotations.

III. The Proposed Framework

A. Document Model

In order to manage the different kinds of multimedia data, their relations and the particular structure imposed by e-gov applications, the adopted document model uses three different representation layers, as described in the following.

Data Management layer: describes the semantic content of each single multimedia objects composing the document (such as a text fragment or an image), providing functionalities for managing each single media; as an example, information extraction and indexing over text and images are performed in this layer.

Integration layer: describes the relations among the heterogeneous multimedia components of the same document or belonging to different ones, providing functionality for their integration and composition. For example, the property of a text fragment of referring to a given image belongs to this layer.

Presentation layer: regulates the way by which the information has to be shown to final users. It provides different representations of the same informative content, according to the formats, the final user's access rights, user preferences and needs and the available technology and user devices.

[2] 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.

This approach allows the management of heterogeneous contents, by separating the presentation logic from the content management one. 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.

According to the different description layers of a document, information is semi-automatically extracted and tagged with respect to the concepts contained in the available *domain ontologies*: associations among concepts and their instances are picked out. A general schema of documents processing is depicted in Figure 1.

More in details, the tagging process leverages different types of ontologies. A **Domain Ontology** is exploited to formalize the concepts of interest in the reference domain and relationships among them. An example of a top-level fragment of the used 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. Some domain ontologies [10] can be further divided into: a **Structural 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 a **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.

B. Processes Overview

We proposed a general framework and the related instance 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. radiographies), video streaming (e.g. ecographies) and audio information.

The framework is composed of several processes: text processing, multimedia data processing, and the integration, retrieval, preservation and presentation tasks (see Figure 3), as described in the following.

The **Text Processing** process aims at extracting relevant information from text, associating specific concepts to the related key terms and defining relationships among them. The text is processed in according to the following pipeline [10]:

1) *Structural analysis*: performs the text segmentation and the related classification in order to identify the different sections constituting the structure of the document.
2) *Linguistic analysis*: performs a morpho-syntactic analysis of the text (i.e., text tokenization and normalization, Part-of-Speech Tagging, lemmatization and complex terms analysis) combined with statistic analysis, thus enabling the extraction of relevant terms. These terms and the information about them, refined with the help of domain experts, will constitute a *lexicon* that is exploited for the building of the set of concepts used for the domain formalization via ontologies.
3) *Semantic analysis:* by using the information of the early analysis, it detects properties and associations among terms, defining the concepts and relationships, allowing ontology building and final documents annotation.

The **Multimedia Data Processing** process has the aim of classifying the other kinds of multimedia objects, associating concepts from the domain ontology. It is composed of two components implementing innovative methods that have been presented in our recent works [2][7]:

1) *Analyzer*: identifies relevant media parts and produces a low-level description that permits to create some indices to help the tagging and retrieval tasks.
2) *Classifier*: uses the indexing information to automatically deduce which concepts, from the domain ontology, are being associated to media elements.

Final information are stored as RDF assertions into a *Knoweldge Base*, that is also mapped in the HL7-CDA standard data format[3]. All the knowledge associated to a documents is in turn managed by proper *ontology repositories*.

Different processes (i.e., **Knowledge Integrator**, **Retrieval**, **Extractor**, **Presentation**) are finally devoted to realize the other discussed tasks.

We want to note that the multimedia knowledge is then managed by a *Multimedia DataBase Management System* (MMDBMS).

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. [5]

[3]RDF triples are translated into an XML based document, according to HL7 specifications, applying one or more XSLT rules. The list of the XSLT transformation rules is downloadable from ourproject web site[4].

[5]A MMDBMS meets certain requirements that are usually divided into the following broad categories: multimedia data modeling, 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 management.

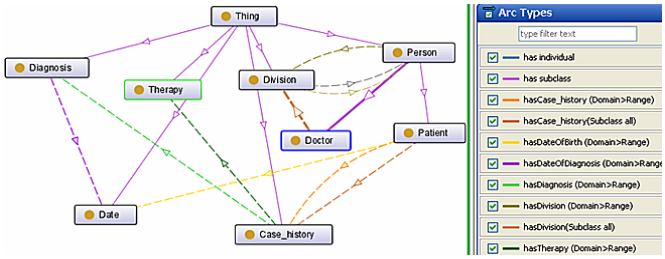

Fig. 2. A Fragment of Domain Ontology for electronic medical record

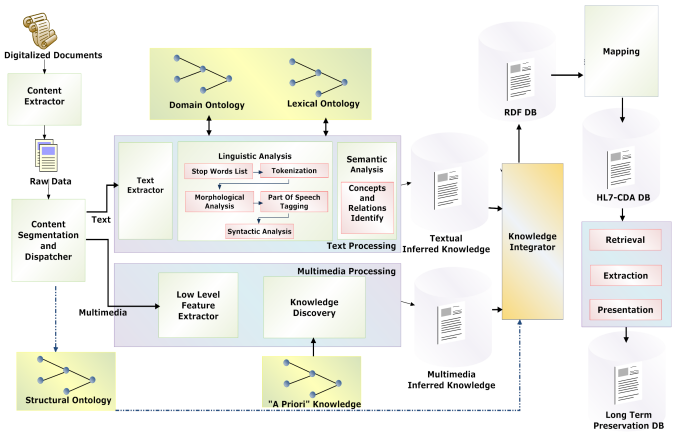

Fig. 3. Document Processing in details

C. Semantic Processing

The document semantic processing supported by the designed system needs a preliminary domain formalization stage that has the aim of codifying, with proper data structures (ontologies), the information of interest pertaining to the domain which the documents belong to. On the top of such structures, we can activate the different described tasks for semantic processing of documents.

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 by adding the detected instances. 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), eventually using subsumption on a $TBox$-$Module$ for deriving more specific concepts.

The detected instances can be shown by using tools like KIM [3], that highlights the associations among detected instances and the concept defined in the domain ontology.

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

Our system combines ORDBMS technologies, NLP techniques, proper domain and structural ontologies and inference rules in order to retrieve significant concepts related to each document and to provide semantic querying facilities for users. 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* [4] 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. 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.

IV. System Architecture and Implementation Details

A prototype version of a **Multimedia Document Management System** has been implemented, according to the following features: *(i)*: it exploits a unified data model that takes into account content-based and document-based characteristics; *(ii)* it uses ontological support for managing the semantics of data; *(iii)* it has a multi-layer architecture with different kinds or user interfaces; *(vi)* it provides advanced functionalities for document indexing and semantic retrieval.

Figure 4 shows at glance the component architecture of our system. The *Digital Documents* (DD) 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 to a set of *semantic concepts*

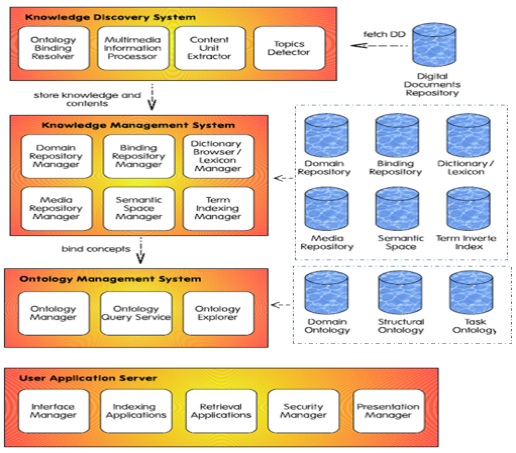

Fig. 4. Component Architecture

– retrievable by semi-automatic information extraction procedures and related to single content units of a document – and a set of *keywords* – defined as particular properties of the whole document.

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 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] (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 using NLP techniques 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 the 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 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 - contains the description of application domain concepts and it is managed by a *Domain repository Manager*.

Binding Repository - contains the associations between document and domain repository concepts and it is managed by a *Binding Repository Manager*.

Media Repository - 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 different modules provised *security* and *presentation* management.

V. Experimental Results

In this section we report some experiments we have carried out for evaluating the impact of the proposed system on enhancing the retrieval perfomances over 30000 medical records, properly anonymized, coming from several Italian health care organizations. To set up our experimentation, we select a subset of the collected data (constituted by 6000 randomly chosen documents) as training set for the different classification tasks (e.g. text segmentation, information extraction, etc.). The objective is to evaluate the system correctness in automatically discovering relevant concepts within a medical record and in particular:

1) Patient Personal Data and Clinical Events (e.g., Age, Sex, Date of Admission, Medical Department, etc.);
2) Diagnosis (Family History, Physiological anamnesis, Medical history, etc.);
3) Diary of Significant Events (e.g., Treatments, Therapeutical Plans, Diagnostic Procedures, etc.);
4) Hospital discharge (Diagnosis at Discharge).

Relevant concepts discovery procedures exploit a domain ontology built from scratch from the medical records dataset, with the help of domain experts.

The comparison between the proposed system semantic indexying output and the ground truth relies on the well

know $recall$ and $precision$, evaluated on the the test set (24000 documents).

$Recall$ measures the ratio between the relevant documents retrieved by our system and the total relevant ones (in the ground truth) with respect to a set given concept, while $precision$ measures the ratio between the relevant retrieved documents and the retrieved ones by our system.

The obtained results are summarized in Table 1, showing the average precision and recall (with the related $F\text{-}Measure$) varying the number and sets of considered concepts.

# Concepts	Average Preciision	Average Recall	$F\text{-}measure$
1	0,837	0,944	0,887285794
2	0,845	0,915	0,878607955
3	0,846	0,892	0,868391254
5	0,849	0,862	0,855450614
7	0,854	0,843	0,848464349
8	0,861	0,814	0,836840597
10	0,865	0,795	0,828524096
12	0,908	0,774	0,835662307
15	0,921	0,765	0,835782918
18	0,956	0,756	0,844317757
20	0,973	0,744	0,843228888

TABLE I
AVERAGE PERFORMANCE

The proposed method achieves an average recall value of 94.4% in finding a single concept for recall with respect to an average precision value of 83.7%, a 79.5% recall rate with a 86.8% precision rate in finding ten different concepts and, finally, an average recall of 74.4% with a precision of 97.3% in finding 20 concepts.

In order to provide an idea about the quality of this results, we also provide the $F\text{-}measure$ metric values, being $F\text{-}measure = 2 \times precision \times recall/(precision + recall)$, which is commonly used to combine $precision$ and $recall$ scores. In the majority of cases, the retrieval achieves the best perfomances with respect to diagnosis concepts. In turn, the most difficult concepts to discover are those related to the 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, the average indexing times with respect to the document size is also reported[6]. For document size (ds) of about $150K$ the indexing time is $1.3s$, for $150K \leq ds < 300K$ is $2.7s$, for $300K \leq ds \leq 500K$ is $3.1s$ and for document of size $\geq 1000K$ indexing time is of $5.2s$.

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, analyzed 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-gov 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.

[6]All the experiments 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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] A Chianese, F Marulli, V Moscato, F Piccialli. SmARTweet: A Location-based smart application for Exhibits and Museums. Signal Image Technology and Internet-Based Systems(SITIS), 2013 International Conference on Signal-Image Technology & Internet-Based Systems.

[8] S.Santini, "Evaluation Vademecum for Visual Information Systems," Proc. of SPIE, vol. 3972, San Jose, USA, 2000

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

[10] 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.