Journal of Information Assurance and Security.
ISSN 1554-1010 Volume 7 (2012) pp. 296-306
© MIR Labs, www.mirlabs.net/jias/index.html

A semantic based framework to identify and protect e-health critical resources

Flora Amato, Valentina Casola, Sara Romano and Antonino Mazzeo

Universita' di Napoli Federico II, Dipartimento di Informatica e Sistemistica,
via Claudio 21, Napoli 80125, Italy
{*flora.amato, casolav, sara.romano, mazzeo*}*@unina.it*

Abstract: **The e-Health is going to change the way how patients and healthcare providers interact. The exchange of confidential and integer information is one of the major open issues for the health care sector. While it is quite easy to enforce fine grain access control policies to new, *well structured*, medical records, many eHealth systems are based on "document management systems" that manage medical records as monolithic documents. The coexistence of both structured and unstructured medical records represents a huge limitation for documents management and, for the latter, it is impossible to enforce fine grain access control rules. In this paper we propose an innovative framework for critical resource identification and protection; the framework is based on a semantic methodology that can be used to classify and structure data. We also designed a modular architecture to let this methodology be useful in many different contexts by properly tuning and expert domains feedbacks. Finally, a case study was presented to structure e-health data according to HL7 and locate the proper security rules to enforce.**

Keywords: Knowledge extraction, document transformation, semantic methodology, fine-grain access control, e-health, medical records

I. Introduction

The e-Health (Electronic Health) is going to change the way how patients and healthcare providers interact. The challenge of e-Health is to contribute to good health care by providing value-added services to the healthcare actors (patients, doctors, etc...) and, at the same time, by enhancing the efficiency and reducing the costs of complex informative systems through the use of information and communication technologies.

The e-Health term encloses many meanings and services, ranging between medicine and information technologies. Just for example, emerging services are: the telemedicine (enhancing communication between doctors and patients by means of audiovisual media), the Consumer Health Informatics (optimizing the acquisition, storage, retrieval, and use of information in health), the m-Health (health care supported by mobile devices) and the Electronical Patient Records (improving patients health information sharing). The combination of such concepts introduced new features and challenges[1] and we are very interested in security open issues that arise in these new scenarios[2].

The management of health care data has different security requirements, among the others, we think that the two primary requirements are: i) the communication and storage of private information should guarantee confidentiality and data integrity, ii) fine-grained access control policies are needed for different actors.

As illustrated in Figure 1, we should consider that a medical record is a structured data made of different parts each of these can be read and/or modified by different actors. Many of this data is private and can be viewed only by patients and their doctors, other parts are anagraphical or administrative information and should be viewed only by administrators of the hospital. We have analyzed the security requirements associated to such data and the result of this analysis has led us to state that an access control model that strongly takes in consideration the attributes of the resources to protect and the actor role should be enforced in e-health systems.

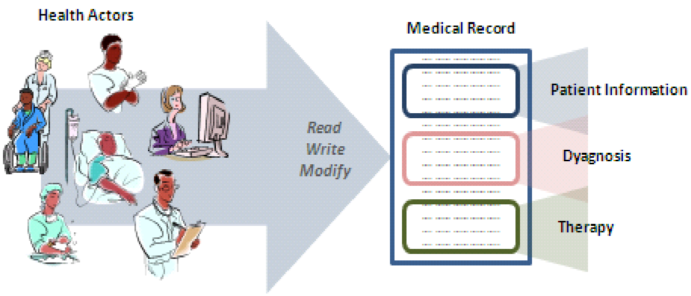


Figure. 1: Actors and medical records

Many access control models exploit the concept of data classification to protect critical resources, the Bell La Padula model [3] is a significant example of access control rules that are based on the security levels of the user-requestor and the resource-requested. Indeed, many eHealth systems are designed to enforce fine-grain access control policies and the medical records are *a-priori* well structured to properly locate the different parts of the managed complex information[4]. Many security problems occur when eHealth systems are applied in those contexts where new information systems have not been developed yet but "documental systems" are, in some way, introduced.

This means that today documental systems improperly allow users to access a digitalized version of a medical record without having previously classified the critical parts. Any document is treated as a monolithic resource.

The classification of critical elements of a not-structured documents is not easy at all; very often they contain ambiguous parts that are strongly related to the doctor activity; for example it is quite usual for a nurse to write in its portion some details of the diagnosis that is competence of the doctor or, sometimes, administrative person write down in the anagraphical parts also some information related to the patient's disease. Up to date, monolithic resources are protected at a course grain level and a permit/deny access rule can be applied to the whole document and not to the specific parts that constitute it.

A medical record contains patient's sensitive information; it is composed by several sections including patient's contact information, summary of doctor's visits, patient's diagnosis, medical and family history, list of prescriptions, health examinations, the therapy, etc. Generalizing, in many contexts as medical, juridical or humanistic, people are more used to protect their data/documents as a monolith block without understanding the risk of not structuring data.

The need to protect monolithic resources, has suggested us to propose a semantic based mechanism to automatically retrieve specific parts of a document and associate to them the proper security level.

At this aim we exploited the adoption of semantic techniques to analyze texts and automatically extract relevant information, concepts and complex relations. We proposed a methodology to classify and structure e-health data according to HL7, an available Electronic Health Records (EHR) standard [5], and we designed a reference architecture to associate the proper security level and enforce proper security policies. The documental collection is analyzed and processed by a lexical-statistical approach, with the aim of extracting the relevant terms that will be associated to the concepts of interest that will represent the resources to protect; a proper security policy is also defined to illustrate a detailed case study.

The reminder of the paper is structured as follows: in Section 2 we will illustrate the semantic methodology to extract relevant concepts from a semi-structured text to characterize resource from a security point of view and format them according to HL7. In Section 3 a reference architecture is introduced to implement the methodology and enforce access control rules on different sections of medical records. In Section 4 a detailed case study on medical records will be presented and in Section 5 some related works on the semantic methodologies and their adoption in security contexts are discussed. Finally in Section 6, some conclusion and future work are drawn.

II. A Methodology for Semantic Based Resource Characterization

The *Patient Medical Records* contains several resources that can be protected by a set of security policies. In order to properly locate and characterize resources made of text sections, we need to apply semantic text processing techniques on available data. Semantic processing of medical documents is not a easy task to be performed; it depends on many factors: the domain knowledge and interpretation given by the document author may not be the same of the reader.

The comprehension of a particular concept within a specialized domain, as the medical one, requires information about the properties characterizing it, as well as the ability to identify the set of entities the concept refers to.

A text, in fact, is the product of a communicative act resulting from a process of collaboration between an author and a reader: the former uses language signs to codify meanings, the latter decodes these signs and interprets their meaning by exploiting the knowledge of:

1. the *infra-textual* context, consisting in relationships at a morphological, syntactic and semantic level;
2. the *extra-textual* context and, more in general, the *encyclopedic knowledge* involving the domain of interest.

It is out of the scope of this paper to detail the lexical and semantic methodology adopted to extract information from text, the interested reader can found more details in [6].

Starting from these points, the activity of knowledge extraction from texts includes different kinds of text analysis methodologies, aiming at recreating the model of the domain the texts pertain to. In the next subsections, we will illustrate the process of extracting information from a medical record, aided with a running example.

To better explain the stages of the methodology, we will use a fragment of a psychiatric medical record as a running example. It states that, at the entrance of the hospital, a patient results quiet and cooperative, calm in the maxilla-facial expression:

```
Diagnosi di entrata la paz. e' tranquilla
e collaborante, serena nell' espr. maxillofacciale
```

We explicitly note that the following examples refer to Italian language, nevertheless the proposed approach is general enough to be applicable to other languages, too.

A. Stages of the Methodology

Term-extraction is a fundamental activity in the automatic document processing and derivation of knowledge from texts.

Terms serve to convey the fundamental concepts of a specific knowledge domain: they have their realization within texts and their relationships constitute the semantic frame of the documents and of the domain itself. The main goal is to find a series of relevant and peculiar terms in order to detect the set of concepts that allow the resource identification.

In order to extract relevant terms from text, we use an hybrid method that combines *linguistic and statistical techniques*: we employ a *linguistic filter* in order to extract a set of candidate terms and then use a *statistical method* to assign a value to each candidate term. In particular, linguistic filters are applied on the words, like as *part-of-speech tagger* (aiming at extracting the categories of interest, such as nouns and verbs), and *lemmatization* (that restore words to a dictionary form).

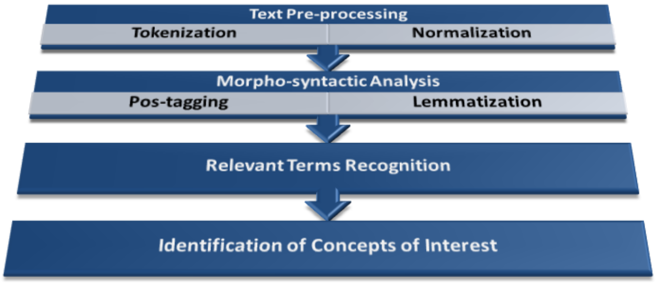

Text Pre-processing
Tokenization
Normalization
Morpho-syntactic Analysis
Pos-tagging
Lemmatization
Relevant Terms Recognition
Identification of Concepts of Interest

Figure. 2: Stages of the methodology for concepts identification

Statistical methods are based on the analysis of word occurrences within texts, in order to measure the “strength” or “weight” of a candidate term. As a matter of fact, not all words are equally useful to describe documents: some words are semantically more relevant than others, and among these words there are lexical items weighting more than others.

In order to extract relevant terms from a medical record, several steps are required; these are described in details in the following sections and illustrated in Figure 2.

1) *Text Preprocessing*

This stage aims at extracting processable plain text from the input documents, by detecting units of lexical elements that can be processed in next stages. It implements text tokenization and text normalization procedures.

Text tokenization consists in segmentation of sentences into tokens, minimal units of analysis, which constitute simple or complex lexical items, including compounds, abbreviations, acronyms and alphanumeric expressions.

Text tokenization requires, various sub-steps, as: *grapheme analysis*, to define the set of alphabetical signs used within the text collection, in order to verify possible mistakes as, for example, typing errors, misprints or format conversion; *disambiguation of punctuation marks*, aiming at token separation; *separation of continuous strings* (i.e. strings that are not separated by blank spaces) to be considered as independent tokens: for example, two terms separated by the character “ ’ ”; and *identification of separated strings* (i.e. strings that are separated by blank spaces) to be considered as complex tokens and, therefore single units of analysis.

This segmentation can be performed by means of special tools, defined *tokenizers*, including *glossaries* with well-known expressions to be regarded as medical domain tokens and *mini-grammars* containing heuristic rules regulating token combinations. The combined use of glossaries and mini-grammars ensures high level of accuracy, even in presence of texts with acronyms or abbreviations that can increase the mistakes rate. Considering our example, the output of text tokenization is:

```
Diagnosi//di//entrata//la//paz.//
e’//tranquilla//e//collaborante,//
serena//nell//’//espr.// maxillofacciale//
```

Text normalization takes variations of the same lexical expression back in a unique way; for example, *(i)* words that assume different meaning if are written in small or capital letter, *(ii)* compounds and prefixed words that can be (or not) separated by a hyphen, *(iii)* dates that can be written in different ways (“1 Gennaio 1948” or “01/01/48”), *(iv)* acronyms and abbreviations (“USA” or “U.S.A.”, “pag” or “pg”), etc.

The transformation of capital letters into small letters, is a not trivial operation: for example, a capital letter helps in identifying the beginning of a sentence and differentiating a common noun (like the flower “rosa”) from a proper name (such as “Rosa”) or even to recognize the distinction between an acronym (e.g.“USA”) and a verb (e.g. “usa”, 3rd sing. pers. of the Italian infinitive “usare”). The output of this phase is, for the running example:

```
Diagnosi//di//entrata//la//
paziente////tranquilla//e//collaborante,//
serena//nell//’//espressione//maxillo-facciale//
```

2) *Morpho-syntactic analysis*

The main goal of this stage is the extraction of word categories, both in simple and complex forms. This leads to obtain a list of candidate terms on which relevant information extraction can be performed.

Part-of-speech (POS) tagging consists of the assignment of a grammatical category (noun, verb, adjective, adverb, etc.) to each lexical unit identified within the text collection.

Morphological information about the words provides a first semantic distinction among the analyzed words. The words can be categorized in: *content words* and *functional words*. Content words represent nouns, verbs, adjectives and adverbs. In general, nouns indicates people, things and places; verbs denote actions, states, conditions and processes; adjectives indicate properties or qualities of the noun they refer to; adverbs, instead, represent modifiers of other classes (place, time, manner, etc.). Functional words are made of articles, prepositions and conjunctions; they are very common in the text.

Automatic POS tagging involves the assignment of the correct category to each word encountered within a text. But, given a sequence of words, each word can be tagged with different categories [7].

As already stated, the *word-category disambiguation* involves two kinds of problems: *i)* finding the POS tag or all the possible tags for each lexical item; *ii)* choosing, among all the possible tags, the correct one. Here the vocabulary of the documents of interest is compared with an external lexical resource, whereas the procedure of disambiguation is carried out through the analysis of the words in their contexts. In this sense, an effective help comes from the *Key-Word In Context (KWIC) Analysis*, a systematic study of the local context where the various occurrences of a lexical item appear. For each concept it is possible to locate its occurrences in the text and its co-text (i.e. the textual parts before and after it).

The analysis of the co-text, then, allows detecting the role of the words in the phrase, in order to disambiguate their grammar category.

The ambiguous form is then firstly associated to the set of possible POS tags, and then disambiguated by resorting to the KWIC analysis. The set of rules defining the possible combinations of sequences of tags, proper of the language, enables the detection of the correct word category.

Consider, in the reported example, the ambiguity associated to the Italian word “entrata”: it can be a noun (“entry”) or a verb (“enter”). This ambiguity can be solved by analyzing the categories of the preceding words: rules derived by syntax of Italian language state that, if the word is preceded by an article or a preposition it is a noun, while if it is preceded by a noun it is a verb. Then, applying KWIC analysis we can derive that “entrata” is a verb.

Further morphological specifications, such as inflectional information[1], are then associated to each word. Output of this stage, for our running example, is:

```
Diagnosi          NOUN
di                PRE
entrata            NOUN
la                ART
paziente          NOUN
tranquilla        ADJ
e                  CON
collaborante       NOUN
,                 PUN
serena            ADJ
nell'             ARTPRE
espressione        NOUN
maxillo-facciale  NOUN
```

Note that, we used the following conventions: (ART) = article; (ADJ) = adjective; (ADV) = adverb; (CON) = conjunction; (NOUN) = noun; (PN) = pronoun; (PRE) = preposition; (VERB) = verb; (ARTPRE)= article + preposition.

Lemmatization is performed on the list of tagged terms, in order to reduce all the inflected forms to the respective lemma, or citation form, coinciding with the singular male/female form for nouns, the singular male form for adjectives and the infinitive form for verbs. The output of this stage, for our example is:

```
Diagnosi         NOUN   diagnosi
di               PRE   di
entrata           NOUN   entrata
la               ART   il
paziente         NOUN   paziente
tranquilla       ADJ   tranquillo
e                 CON   e
collaborante     NOUN  collaborante
,                PUN   ,
serena           ADJ   sereno
nell'            ARTPRE nel
espressione       NOUN   espressione
maxillo-facciale NOUN   maxillo-facciale
```

Note that many terms are already present in canonical form, and for this reason, in this phase, they are not converted; while the other terms, as the adjective “tranquilla” or the preposition “nell” are respectively transformed in “tranquillo” and “nel”.

3) Relevant Terms Recognition

The goal of the methodology is the identification of the relevant terms, useful to characterize the sections of interest[8]. In fact, as state above, not all words are equally useful to describe resources: some words are semantically more relevant than others, and among these words there are lexical items weighting more than other. In our approach, the semantic relevance is evaluated by the assignment of the TF-IDF index (*Term Frequency - Inverse Document Frequency*[9]), computed on the corpus vocabulary and on the base of the term frequency and the term distribution within the corpus. TFIDF index, in fact, takes into account:

- the *term frequency* (*tf*), corresponding to the number of times a given term occurs in the resource: the more a term occurs in the same section, the more it is representative of its contents. Frequent terms are then supposed to be more important. This method is used in systems to rank terms candidates generated by linguistic methods (Dagan [10]).
- the *inverse document frequency (idf)*, concerning the term distribution within all the sections of the medical records: it relies on the principle that term importance is inversely proportional to the number of documents from the corpus where the given term occurs. Thus, the more resources contain that given term, the less discriminating it is.

Therefore, $TFIDF$ enables the extraction of the most discriminating lexical items because they are frequent and concentrated on few documents. This statement is summarized in the following ratio:

$$W_{t,d} = f_{t,d} * log(N/D_t)$$

where $W_{t,d}$ is the evaluated weight of term *t* in resource *d*; $f_{t,d}$ is the frequency of term *t* in the resource *d*; *N* is the total number of occurrences within the examined corpus; D_t is the number of resources containing the term *t*.

For the running example, this phase produces the following information:

```
diagnosi          5       *
entrata           1,5
paziente          4       *
tranquillo        2,8
collaborante      3,1     *
sereno            2,5
espressione       3,8     *
maxillo-facciale  7       *
```

This information enables the selection of relevant concepts, filtering all terms that have a TF-IDF value under an established threshold. We used, as threshold, the value 3: all terms whose TF-IDF is over this threshold will be considered relevant. In the example, we mark off the relevant terms with an asterisk.

4) Identification of Concepts of Interest

Once relevant terms, belonging to the used medical subdomain, are detected, we proceed to clusterize them in **synset** *(a group of data elements that are considered semantically equivalent for the purposes of information retrieval)*, in order to associate the semantic concept that every cluster of terms refers to.

In this way it is possible referring to a concept independently from the particular term used to indicate it. Examples of the use of concepts, codified as synsets, for identifying sections of text are shown in [11] for the medical domain, and in [12] for the legal domain.

For grouping aim, we use and integrate two external resources: the medical ontology given by “Unified Medical Language System” (UMLS)[13] and “Mesh”(the Medical

[1] Inflection is the way language handles grammatical relations and relational categories such as gender (masculine/feminine) and number (singular/plural) for nouns; tense, mood, person and voice for verbs.

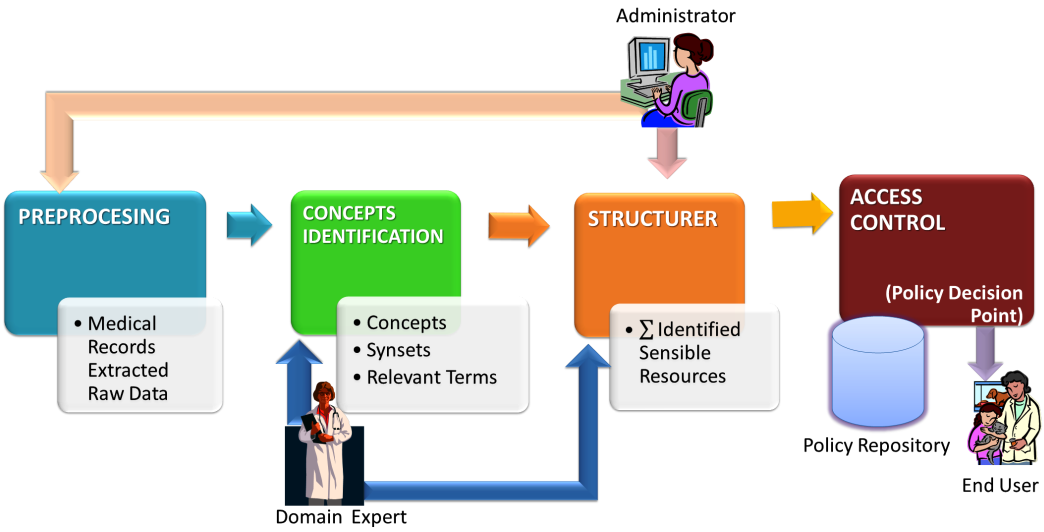


Figure. 3: Reference Architecture

Subject Headings of National Library of Medicine[14]), a thesaurus of medical terms.

The adoption of specialized external resources has a duplicate purpose:

- **Endogenous:** Inside the documental base, the same concepts can be referred by different terms.
- **Exogeneous:** An user, that can query the documental base with an interrogation written in natural language, can use, for indicate a certain concept, a term that is different from those used in the documental base, and then do not appear in it.

Every concept is identified by a synset (i.e. the set of synonyms), we associate each term extracted from the medical record to a synset by a unique label that represents a witness for the given synset.

This stage associates the synset, i.e. the proper concept, to each selected term of the running example, as showed in the following table:

Term	Synset	Label
diagnosi	parere, prognosi, responso, valutazione, analisi	Diagnosi
paziente	ammalato, degente, malato	Paziente
espressione	manifestazione, segno, smorfia, viso, sintomo	Sintomo
maxillo-facciale	maxillo-facciale	Maxillo-Facciale

The table shows that for each relevant term, extracted on the basis of its grammatical category and TF-IDF value, is associated a synset: a set of terms referring the same concept. This list of terms is built by exploiting the relation codified in the external domain resources: UMLS and "Mesh". In our example we obtain the concepts associated to the extracted terms: "Diagnosi" (diagnosis) , "Paziente" (patient), "Sintomo" (Symptom) and "Maxillo-Facciale" (maxillofacial).

III. Reference Architecture

In order to implement the methodology described in the previous section and enforce access control rules on medical records sections, we introduce a reference architecture as depicted in figure 3. It can be considered as an instance of a Multimedia Database Managment System (MMDBMS) architectural model [8] that accepts in input the corpus made of medical records and performs activities aiming to structure them and allowing the identification of the resources to be protected.

As shown in figure 3, the system architecture is composed of four sequential connected modules. Each module is delegated to a specific activity described as follows:

Preprocessing. This module aims at processing the input text (medical records) and produces a list of words. In order to perform this activity it implements Natural Language Processing (NLP) procedures. These procedures are language dependent and include the first stage of the semantic methodology described in section II: text preprocessing (text tokenization and text normalization).

Concepts Identification. This module aims at extracting the relevant information from the list of words produced by the Preprocessing module in order to produce a formalization of concepts belonging to the medical domain. In this module are implemented both Natural Language Processing (NLP) and Statistic procedures. As for the *Preprocessing* module the procedures involved are language dependent but, in this case, the second, third and fourth stages of the semantic methodology described in section II are involved, i.e.: the morphosyntactic analysis (part-of-speech (POS) tagging and lemmatization), the relevant term recognition (by means of grammatical category extraction and TF-IDF index) and identification of concepts of interest. Moreover this module aims at codify the semantic information (concepts) of medical domain in several levels of ontologies allowing further successive inferences that will be useful to the following modules.

Structurer. This module identifies the textual macrostructures for text sections recognition. The macrostructures identification process consists of a classification task, exploiting the concepts identified in the Concepts Identification module. At this aim a voting system

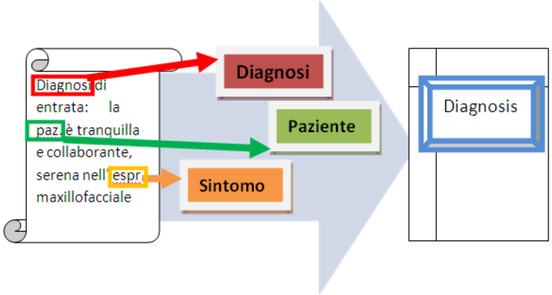


Figure. 4: Example of Process for Resource Association

is implemented, using the combination of three different kind of classifiers: Nave Bayes [15], Decision Tree [16], K-Nearest Neighbor [17]. The classification results are combined by means of a voting strategy: in case of disagreement, the assigned output class will be the one that gets the majority.

This module is able to associate a medical record section to a specific resource to be protected; this association is refined by medical domain experts that establish which concepts belong effectively to a resource (medical record section). For example, the Diagnosis resource contains the following concepts: “Diagnosi” (diagnosis), “Paziente” (patient) and “Sintomo” (manifestation). As illustrated in Figure 4, the system is able to establish that the medical record section considered in the example refers to the “Diagnosis” resource; note that, in figure, a label is associated to each synset.

The output of the structurer module can be coded in different ways, in particular we chose to formalize our examples in XML and according to the HL7 standard for medical records [18] . The result was an XML file whose main elements are reported below:

- *Patient information* Section (including anagraphycal, private doctor, personal information?),
- *Investigation and Diagnosis* Section (including the activities of investigation and detection of a disease),
- *Therapy* Section (including the therapy description with drugs and doses),
- *Patient status* Section (that can be monitored by a doctor or a nurse),
- *Analysis results* Section (including comments and analysis description by specific analysts),
- *Admission information* Section (including information on the patient when arrived in the hospital, and different aspects that can be useful to doctors for future investigations and to hospital managers to complete administrative stuffs).

These resources should be accessible only to those people having proper rights.

Access Control. This module aims at controlling information accesses, assigning the appropriate access policy to every sensible resource, identified by the Structurer module. Once the resources are identified, it is possible to apply on them a fine grain access control policy, based on users profile. The security policy is made of a set of rules structured as follows.
A rule is as a triple $\langle s_j, a_i, r_k \rangle$ where s$_j \in S$, a$_i \in A$, r$_k \in R$ and:

- $S = \{s_1, ..., s_m\}$ is the set of the actors s$_j$ that can access to the medical record,
- $R = \{r_1, ..., r_n\}$ is the set of all resources (sections) r$_i$ belonging to the medical record,
- $A = \{a_1, ..., a_h\}$ is the set of actions that can be performed by an actor s$_j \in S$ on a resource r$_i \in R$.

For each resource, a subset of rules belonging to the applicable policy set is available.

So, given a resource $r^* \in R$, all the possible rules, denoted as L_{r^*}, belonging to the Policy will be retrieved; by definition:

- $L_{r^*} = \{\langle s_j, a_i, r^* \rangle | r^* \in R, s_j \in S^* \subseteq S, a_i \in A^* \subseteq A\}$ is the set of all allowed combinations of (subjects,actions) on the resource r^*.

The policy will be enforced by a *policy enforced component* and a *policy decision component*.

Three kind of users are involved in the system architecture: *(i)* **Administrators**, involved in administrative tasks, like inserting new records in the documental basis and managing the user accounting; *(ii)* **Domain Experts** that should operate on the outputs of the concept identification and structurer modules in order to respectively validate the list of concepts produced, and to validate the sensible resources identified from the input medical records; *(iii)* **Final Users** that can be associated to different profiles (nurse, doctor, patient) depending on their role in the health domain.

IV. The medical record formalization: a case study

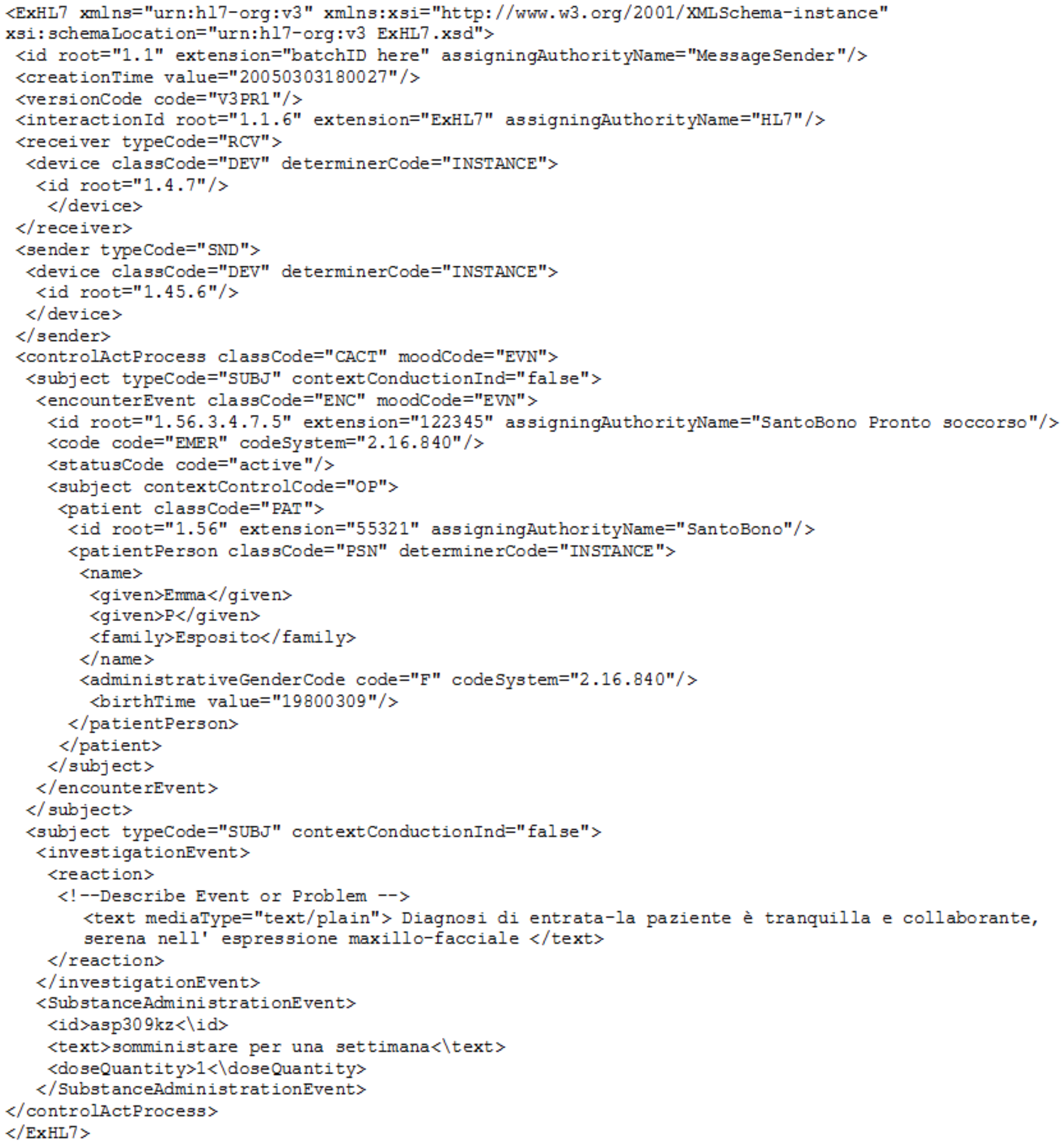
To better illustrate a case study, we can consider this complex system as made of two different phases: the first phase is the semantic processing, the second phase is the policy enforcement, as schematically illustrated in Figure 5.
As for the semantic processing phase, we consider the running example previously adopted. The whole medical record text in input to the system is:

```
Ospedale Santo Bono  reparto Pronto Soccorso
Data 03/03/2005 ore 18,00

Paziente
Nome: Emma P.
Cognome: Esposito
Nata il: 09/03/1980

Diagnosi di entrata-la paziente  tranquilla e collaborante,
serena nell’ espressione maxillo-facciale.
Somministrare per una settimana una compressa di asp309kz.
```

The details of the processing steps have already been reported in the methodology description so, in Figure 6, we just reported the resulting XML structure, coded according the HL7



```
<ExHL7 xmlns="urn:hl7-org:v3" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="urn:hl7-org:v3 ExHL7.xsd">
 <id root="1.1" extension="batchID here" assigningAuthorityName="MessageSender"/>
 <creationTime value="20050303180027"/>
 <versionCode code="V3PR1"/>
 <interactionId root="1.1.6" extension="ExHL7" assigningAuthorityName="HL7"/>
 <receiver typeCode="RCV">
  <device classCode="DEV" determinerCode="INSTANCE">
   <id root="1.4.7"/>
    </device>
 </receiver>
 <sender typeCode="SND">
  <device classCode="DEV" determinerCode="INSTANCE">
   <id root="1.45.6"/>
  </device>
 </sender>
 <controlActProcess classCode="CACT" moodCode="EVN">
  <subject typeCode="SUBJ" contextConductionInd="false">
   <encounterEvent classCode="ENC" moodCode="EVN">
    <id root="1.56.3.4.7.5" extension="122345" assigningAuthorityName="SantoBono Pronto soccorso"/>
    <code code="EMER" codeSystem="2.16.840"/>
    <statusCode code="active"/>
    <subject contextControlCode="OP">
     <patient classCode="PAT">
      <id root="1.56" extension="55321" assigningAuthorityName="SantoBono"/>
      <patientPerson classCode="PSN" determinerCode="INSTANCE">
       <name>
        <given>Emma</given>
        <given>P</given>
        <family>Esposito</family>
       </name>
       <administrativeGenderCode code="F" codeSystem="2.16.840"/>
        <birthTime value="19800309"/>
      </patientPerson>
     </patient>
    </subject>
   </encounterEvent>
  </subject>
  <subject typeCode="SUBJ" contextConductionInd="false">
   <investigationEvent>
    <reaction>
     <!--Describe Event or Problem -->
       <text mediaType="text/plain"> Diagnosi di entrata-la paziente è tranquilla e collaborante,
       serena nell' espressione maxillo-facciale </text>
    </reaction>
   </investigationEvent>
   <SubstanceAdministrationEvent>
    <id>asp309kz<\id>
    <text>somministare per una settimana<\text>
    <doseQuantity>1<\doseQuantity>
   </SubstanceAdministrationEvent>
</controlActProcess>
</ExHL7>
```

Figure. 6: The structured medical record

standard and containing the main elements just indicated in the previous section. These elements represent the resources to protect $R = \{PatientInformation, Diagnosis, Therapy, Patient's\ status, AnalysisResults, AdmissionInformation\}$.

To define the policies, we have to consider that several people belonging to the hospital staff can access and modify medical records. In [19], [20] several hospital actors were identified, we considered the following actors: $S = \{patients, doctors, nurse, hospitalmanagers\}$. Each actor can perform several actions on the resources belonging to the electronic medical folder. We considered the following actions: $S = \{read, write, delete, modify\}$.

All possible rules on the medical resources are defined by the security policies. In Figure 7 an instance of security policy is shown.

In this instance the actors set is $S = \{Doctor, Nurse, Hospital\ manager, Patient\}$; the set of actions is $A = \{read, write, modify, delete\}$; the set of resources is a subset of all possible resources, i.e $R = \{Patient's\ contact\ Information,\ Diagnosis\}$.

In Figure 8 an example of the system behavior is shown: the system takes as input the unstructured medical record containing several sections including diagnosis and therapy. This sections are identified by means of semantic processing; starting from the identified resources, it is possible to retrieve the security policy in which the allowed operations on resources performed by the actors are described and enforce them. The set of security policies will contain all possible rules applicable to all resources.

In conclusion, the semantic methodology application proposed in this work provides a fine-grain resource identifica-

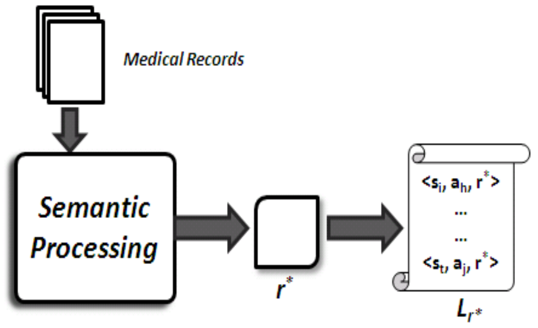


Figure. 5: System phases: semantic processing and policy enforcement

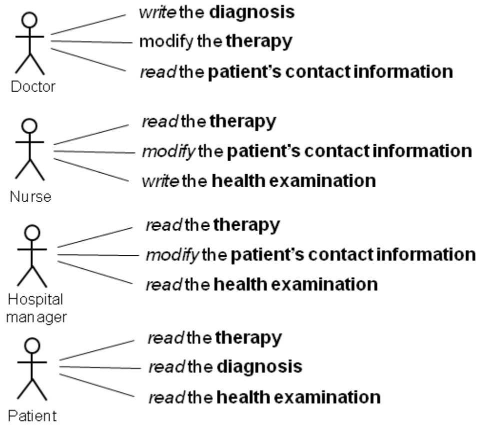

Doctor
write the **diagnosis**
modify the **therapy**
read the **patient's contact information**
Nurse
read the **therapy**
modify the **patient's contact information**
write the **health examination**
Hospital manager
read the **therapy**
modify the **patient's contact information**
read the **health examination**
Patient
read the **therapy**
read the **diagnosis**
read the **health examination**

Figure. 7: Medical record security policy. The actors are on the left, the actions are in italic and the resources are in bold

tion in unstructured texts enabling the application of role-based security policies.

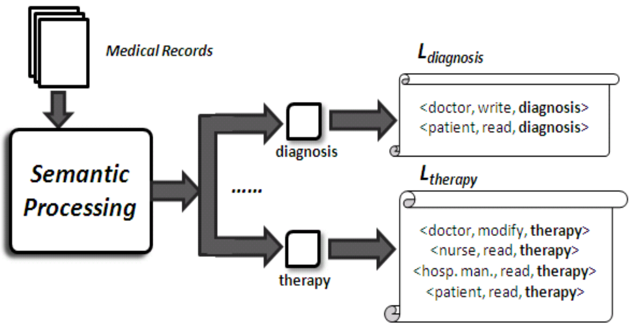


Figure. 8: System behavior with the resources *therapy* and *diagnosis*

V. Related Works

In this section we report the state of the art of the knowledge management methodologies applied in the security fields. In order to properly model sensible information in the domain of interest, in the second part of this section, standards for representation models for clinical information are elicited.

The adoption of semantic approaches in policy management and in the security research fields is quite new in the literature; in [21] the authors propose an ontology-based policy translation approach that mimics the behavior of expert administrators, to translate high level network security policies into low level enforceable ones. In [22], a text mining method has been proposed to deal with large amounts of unstructured text data in homeland-security applications. The document-clustering approach proposed in this work addresses security issues by combining an effective distance measure and an established pattern-recognition grouping algorithm to catalog different information that can be reconducted to criminal acts. Text extraction techniques are widely adopted in e-mail spam recognition [23], too.

Indeed, the activity of knowledge extraction from texts includes different kinds of text analysis methodologies. The state of the art in this field is related to techniques of NLP and to cross-disciplinary perspectives including Statistical Linguistics (De Mauro T. and Butler C.S. [24]) and Computational Linguistics (Biber [25] ; Habert B. [26]; Kennedy G. [27]), whose objective is the study and the analysis of natural language and its functioning through computational tools and models. In particular, for the analysis of limited textual universes, as well as sectorial areas, specific disciplines have been developed, like Corpora Linguistics and Textual and Lexical Statistics (La Torre M. and De Mauro [28]).

Despite efforts to find a common standard for medical documents structuring and to facilitate the interoperability of information, many goals remain unfulfilled. The representation models of clinical information do not yet have a theoretical base strong enough to ensure information interoperability and computability. A model for the EHR should satisfy a large set of requirements including: computational efficiency, maintainability, scalability and extensibility requirements of the system for health information privacy and security. To meet these needs, in openEHR a new aspect was introduced: ontologies[29]. Ontologies are a formal way to describe aspects of a domain. These are used primarily for two reasons: a) people and machines can agree on the "facts" of the domain and b) inferences can be performed, usually based on the classification of "facts" in individual medical categories (eg, the patient has a chronically high blood pressure means that the person is hypertensive) and alert classes (patient A has a high risk of stroke). As regards the first aspect (a) POMR Ontology (Problem-Oriented Record Ontology)[30] considers that a medical record is a repository of medical information and is the means of communication. POMR Ontology is an ontology that describes the medical records so that there is a unique vocabulary for electronic health records. As regards the second aspect (b) Beale and Heard [31]propose a model for clinical information based on health care ontological analysis seen like a problem-solving process. According to their point of view, medical records contain a list of events, situations, etc. that are interpreted by professionals. The implication is that any model for the health information representation should be, in some way, the "cognitive" communication process of health professionals. To achieve this, the authors propose an ontology whose main purpose is to codify some types of information such as medical advice and observations from which the system is able to automatically identify actions that should be undertaken on the patient. In order to address security and access control for EHR systems, several solutions have been proposed [32]. Although these solutions utilize role based access control for security management none of these took into account the structure and the semantics of EHRs. A first step in this direction was made in [33]. This approach focuses on identifying and organizing EHRs by means of semantic interpretation of internal data so that access control policies can be specified to authorize EHRs portions data sharing.

In the field of data modeling, several standards have been developed in order to *(i)* support interoperability and *(ii)* meet the data structures heterogeneity: Health Level 7 (HL7) Clinical Document Architecture (CDA) [18], CEN EN 13606 EHRcom [34] and openEHR Community[35]. These standards aim to structure medical record contents for data exchange improvement. The IHE (Integratine the healthcare Enterprise) initiative[36] specify the Cross-Enterprise Document Sharing standard to manage documents sharing between several healthcare organizations. The IHE Cross-Enterprise Document Sharing basic idea is to preserve the health document in a XML-based format in order to facilitate the sharing. A medical record may also contain images for example from X-rays; DICOM (Digital Imaging and Communication in Medicine) [37] has become the de-facto standard for communication of medical images. This standard defines the data structures to facilitate the exchange of medical images and attached information. Of course there are also proposals to convert one standard to another. For example, the HL7 consortium proposes a mapping between DICOM SR "Basic Diagnostic Imaging Report" in HL7 CDA Release2 "Diagnostic Imaging Report" Mapping. Another important initiative, born in the 90's, is GEHR (Good Electronic Health Record)/openEHR that introduces an additional concept in the context of electronic health records: the archetype. An archetype is a formal expression of a single concept such as, for example, "blood pressure", "laboratory results", "clinical exams" that are expressed as constraints on data whose instances conform to a reference model [38]. Systems based on archetypes specify standards for access to medical information exchange protocols and thus promoting information interoperability and accessibility. In order to meet future requirements, this standard has been designed so that it can be easy to expand it. In this way, the information contained in systems based on archetypes can be used across several institutions both at present and in the future.

VI. Conclusion

In the last years, the eHealth systems are considerably improving the quality and performance of services that an hospital is able to provide to its patients and his workers. Up to date, many systems are based on document management systems and cannot benefit of new system design techniques to structure data and enforce fine-grain access control policies. Indeed, the medical records, especially the old ones, are just digitalized and made available to users. Being a monolithic resource, it is difficult to enforce proper security rules to guarantee privacy and confidentiality of data. In this paper we have analyzed the security requirements of medical records and proposed a semantic approach to analyze the text, retrieve information from specific parts of the document that can be useful to classify them from a security point of view and, finally, associate a set of security rules that can be enforced on those parts. We have illustrated the adoption of the methodology on a simple case study to put in evidence the potentiality of the proposed methodology. We think that the adoption of the semantic analysis on data that are already available and that cannot be structured a-posteriori, is very promising and can strongly help in facing security issues that arise once data are made available for new potential applications.

References

[1] NR Hardiker, S Bakken, W Goossen, D Hoy, and A Casey. International standards to support better information management. In C Weaver, C Delaney, P Weber, and R Carr, editors, *Nursing and Informatics for the 21st Century: An International Look at Practice, Education and EHR Trends, Second Edition*, pages 253–261. HIMSS, Chicago, 2010.

[2] Ajit Appari and M. Eric Johnson. Information security and privacy in healthcare: current state of research. *International Journal of Internet and Enterprise Management*, 6(4):279– 314, 2010.

[3] D. Elliot Bell and Leonard J. LaPadula. Secure computer systems: Mathematical foundations and model, 1973.

[4] Ji-Won Byun, Elisa Bertino, and Ninghui Li. Purpose based access control of complex data for privacy protection. In *Proceedings of the tenth ACM symposium on Access control models and technologies*, SACMAT '05, pages 102–110, New York, NY, USA, 2005. ACM.

[5] Ilias Iakovidis. Towards personal health record: current situation, obstacles and trends in implementation of electronic healthcare record in europe. *International Journal of Medical Informatics*, 52(1-3):105–115, October 1998.

[6] Katerina T. Frantzi, Sophia Ananiadou, and Junichi Tsujii. The c-value/nc-value domain independent method for multi-word term extraction. In *Natural Language Processing*, 1999.

[7] Tamburini Fabio. Annotazione grammaticale e lemmatizzazione di corpora in italiano, 2000.

[8] Flora Amato, Antonino Mazzeo, Vincenzo Moscato, and Antonio Picariello. A system for semantic retrieval and long-term preservation of multimedia documents in the e-government domain. *Int. J. Web Grid Serv.*, 5:323–338, December 2009.

[9] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. In *Information processing and management*, pages 513–523, 1988.

[10] Ido Dagan and Ken Church. Termight: identifying and translating technical terminology. In *Proceedings of the fourth conference on Applied natural language processing*, ANLC '94, pages 34–40, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

[11] Flora Amato, Valentina Casola, Antonino Mazzeo, and Sara Romano. A semantic based methodology to classify and protect sensitive data in medical records. In *Information Assurance and Security (IAS), 2010 Sixth International Conference on*, pages 240 –246, aug. 2010.

[12] Flora Amato, Antonino Mazzeo, Antonio Penta, and Antonio Picariello. Knowledge representation and management for e-government documents. In Antonino Mazzeo, Roberto Bellini, and Gianmario Motta, editors, *E-Government Ict Professionalism and Competences Service Science*, volume 280 of *IFIP International Federation for Information Processing*, pages 31–40. Springer Boston, 2008. 10.1007/978-0-387-09712-1_4.

[13] D. Lindberg, B. Humphreys, and A. McCray. The unified medical language system. *Methods of Information in Medicine*, 32(4):281–291, 1993.

[14] Fact sheetmedical subject headings (mesh), 2011.

[15] K. Aas and L. Eikvil. Text categorisation: A survey, 1999.

[16] Greg Ridgeway, David Madigan, and Thomas Richardson. Interpretable boosted nave bayes classification. In *4th International Conference on Knowledge Discovery and Data Mining*, pages 101–104, 1998.

[17] Pingpeng Yuan, Yuqin Chen, Hai Jin, and Li Huang. Msvm-knn: Combining svm and k-nn for multi-class text classification. *Semantic Computing and Systems, IEEE International Workshop on*, 0:133–140, 2008.

[18] Robert H. Dolin, Liora Alschuler, Sandy Boyer, Calvin Beebe, Fred M. Behlen, Paul V. Biron, and Amnon Shabo Shvo. Hl7 clinical document architecture, release 2. *Journal of the American Medical Informatics Association*, 13(1):30–39, 2006.

[19] Rossilawati Sulaiman, Dharmendra Sharma, Wanli Ma, and Dat Tran. A multi-agent security framework for e-health services. In *Knowledge-Based Intelligent Information and Engineering Systems and the XVII Italian Workshop on Neural Networks on Proceedings of the 11th International Conference*, KES '07, pages 547–554, Berlin, Heidelberg, 2007. Springer-Verlag.

[20] Rossilawati Sulaiman, Dharmendra Sharma, Wanli Ma, and Dat Tran. A security architecture for e-health services. In *Advanced Communication Technology, 2008. ICACT 2008. 10th International Conference on*, volume 2, pages 999 –1004, feb. 2008.

[21] Cataldo Basile, Antonio Lioy, Marco Vallini, and Salvatore Scozzi. Ontology-based security policy translation. *Journal of Information Assurance and Security*, 5:437–445, 2010.

[22] Sergio Decherchi, Paolo Gastaldo, Judith Redi, and Rodolfo Zunino. A text clustering framework for information retrieval. *Journal of Information Assurance and Security*, 4:174–182, 2009.

[23] Bogdan Vrusias and Ian Golledge. Online self-organised map classifiers as text filters for spam email detection. *Journal of Information Assurance and Security*, 4:151–160, 2009.

[24] Christopher S. Butler. *Statistics in Linguistics*. Blackwell, Oxford, 1985.

[25] Douglas Biber, Randi Reppen, and Susan Conrad. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press, 1998.

[26] Benoet Habert, Adeline Nazarenko, and Andre Salem. *Les linguistiques de corpus*. Armand Colin, 1997.

[27] Graeme D. Kennedy. *An introduction to corpus linguistics*. Longman, 1998.

[28] Tullio De Mauro, Federico Mancini, Massimo Vedovelli, and Miriam Voghera. *Lessico di frequenza dell' italiano parlato*. Etas libri, Rome, IT, 1993.

[29] Al Rector, Md Phd, R Qamar Msc, and T Marley Msc. Binding ontologies and coding systems to electronic health records and messages. In *Proc of the Second International Workshop on Formal Biomedical Knowledge Representation (KR-MED 2006); 2006; 2006. p. 11-9. The SAGE Guideline Model Page 30 of 35*, pages 11–19, 2006.

[30] Rose Dieng-Kuntz, David Minier, Marek Rzicka, Frdric Corby, Olivier Corby, and Laurent Alamarguy. Building and using a medical ontology for knowledge management and cooperative work in a health care network. *Computers in Biology and Medicine*, 36(7-8):871–892, 2006.

[31] T Beale and S Heard. The openehr ehr service model. *Revision 02 openEHR Reference Model the openEHR foundation*, 2003.

[32] David M. Eyers, Jean Bacon, and Ken Moody. Oasis role-based access control for electronic health records. *IEE Proceedings - Software*, 153(1):16–23, 2006.

[33] Jing Jin, Gail-Joon Ahn, Hongxin Hu, Michael J. Covington, and Xinwen Zhang. Patient-centric authorization framework for electronic healthcare services. *Computers and Security*, 30(2-3):116 – 127, 2011. Special Issue on Access Control Methods and Technologies.

[34] Health informatics - Electronic health record communication – Part 1: Reference model, 2008.

[35] T Beale. The openehr archetype system. *BMC Medical Informatics and Decision Making*, 1(1):1–19, 2007.

[36] Wei-hua Yao, Xu-yang Zhu, and Ni Duan. Integrating the healthcare enterprise: an overview. *Di 1 jun yi da xue xue bao Academic journal of the first medical college of PLA*, 23(12):1334–1337, 2003.

[37] Health informaticsDigital imaging and communication in medicine (DICOM), 2006.

[38] Thomas Beale and Sam Heard. Archetype definitions and principles. *Revision 06 March*, pages 1–15, 2007.

Author Biographies

Valentina Casola is currently an Assistant Professor at the Computer and System Engineering Department of the University of Napoli Federico II, Italy. She received the MSc degree in Electronic Engineering from the University of Naples Federico II with honors, in 2001. She got a Ph.D. in Computer Engineering from the Second University of Napoli in 2004. Her research activities include both theoretical and experimental issues, in the areas of security of information systems and policy based approach for security evaluation and management, these activities are documented by many publications, in national and international journals and conference proceedings.

Flora Amato received her PhD in Computer Science and Engineering in 2009 from the University of Naples Federico II. She attends her research and teaching activities in the Department of Computer Science and Systems of the University of Naples Federico II. Her current research interests lie in information retrieval, knowledge management and extraction. These activities are documented by many publications in national and international journals and conference proceedings.

Sara Romano is currently a Ph.D. student at the Computer and System Engineering Department of the University of Napoli Federico II, Italy with the supervision of Prof. A. Mazzeo. She received the MSc degree in Computer Science from the University of Naples Federico II 2009. Her research activities are about semantic knowledge management and engineering of documents with experimental issues on data belonging to the medical and juridical domain. These activities are documented by publications in national and international conference proceedings.

Antonino Mazzeo is Full Professor at the University of Naples Federico II, Italy. He has led research projects partially supported by the Ministry, CNR, ASI and EC. He is involved in scientific collaborations with international research agencies and universities. He has a wide experience in the field of information systems, applied to the e-Government domain.