An Integrated Framework for Securing Semi-Structured Health Records

Flora Amato[a], Giuseppe De Pietro[b], Massimo Esposito[b*], Nicola Mazzocca[a]

[a]*Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione, University of Naples Federico II, Naples, Italy*

[b] *National Research Council of Italy – Institute for High Performance Computing and Networking (ICAR), Via P. Castellino 111, 80131 Naples, Italy*

{flora.amato, nicola.mazzocca}@unina.it {giuseppe.depietro, massimo.esposito}@na.icar.cnr.it

Abstract

In the last years, the adoption of Electronic Health Records (EHRs) have been widely promoted, with the final aim of improving care quality and patient safety. Yet, sharing patient data in a large distributed and heterogeneous context, such as the healthcare domain, has inherently introduced security and privacy risks, due to the great sensitivity and confidentiality of the patient data and the need of accessing such data by a large number of health care workers with various roles for the patient care. Even though various techniques have been developed to effectively implement fine-grained access control, which allows flexibility in specifying differential access rights of individual users, some unsolved problems can be pointed out with respect to the specification of complex policies over EHRs: i) the difficulty of forcing narrative text to assume a semi-structured coded form into EHRs in order to build access control policies also working at a section-level; ii) an overly high-level of theoretical ability required to practically use access control models and policy languages as a whole, due to a scarce integration among them; iii) the lack of tools for easily editing and upgrading access control policies over EHRs. In order to face all these open issues, this paper proposes a hybrid framework aimed at enabling and supporting the definition of fine-grained access control policies working on semi-structured EHRs. The key issues of the framework are: i) a semantic-based method that hybridizes linguistic and statistical techniques in order to give a semi-structured form to a narrative text to be inserted into EHRs, by identifying its specific sections; ii) a formal role-based authorization model, encoded as a couple of ontologies, to regulate the access to these semi-structured EHRs with respect to their sections; iii) a procedural policy language and a set of patterns to simply encode and update access control restrictions in the form of "if–then rules" built on the top of the ontological model formalized. A prototype implementation of this framework is realized in the form of a system offering simple and intuitive interfaces to the security administrators. Finally, an experimental evaluation over real documents contained into EHRs, i.e. discharge summaries, is described, showing the feasibility of the proposed framework and suggesting that its application could simply and proficiently secure the access to healthcare information contained into semi-structured EHRs and, thus, face security and privacy risks in real healthcare scenarios.

Keywords: Electronic Health Record, Role-based Access Control, Policy Language, Information Structuring, Ontology.

[*] Corresponding author. Tel.:+390816139512; fax: +390816139531; e-mail: massimo.esposito@na.icar.cnr.it.

1. Introduction

Information and communication technologies have greatly affected the delivery of health care, in the form of computerized systems, such as health information systems, clinical information systems, and picture archiving and communication systems, with the final aim of both reducing healthcare costs and improving care quality and patient safety (Steele et al., 2010). In particular, in the last years, hospitals and health care providers have increased the adoption of such electronic health care systems to manage patient health care data in the form of Electronic Health Records (hereafter, EHRs) (Martino et al., 2008). According to the International Organization for Standardization (ISO) definition, EHR 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 (Häyrinen et al., 2008). This information included in EHRs has several different functions in patient care, management and health policy, generating many benefits for clinicians, clients, and, generally, for the healthcare system (Sari et al. 2012), such as improved decision-making at the point of care (Fowler et al., 2014), safer drug administration (Phansalkar et al., 2013) better medication management, better adoption of screening programs, advanced tools and services for remote health monitoring (Goldberger et al., 2000; Rodrigues et al, 2013;Varshney, 2014), enhanced communication between a variety of healthcare professionals, improved resource utilization, and so on (Mount et al., 2000).

The amount and quality of information available to health care professionals in EHRs has a pivotal role to support continuing, efficient and quality integrated healthcare (Argüello Casteleiro et al., 2009), yet sharing patient data in a large distributed and heterogeneous context, such as the healthcare domain, inherently introduces security and privacy risks (Martino et al., 2008). In particular, due to the great sensitivity and confidentiality of the patient data and the fact that such data may need to be accessed by a large number of health care workers with various roles for the patient care, a high level of secure protection for data and data access is required. One of the most challenging aspects with respect to security and privacy for healthcare organizations, however, is the amount of power given by 'Patient consent and confidentiality' to the patients in terms of access control restrictions over their individual EHRs (Eyers et al., 2006). Even though various techniques have been developed to effectively implement fine-grained access control, which allows flexibility in specifying differential access rights for individual users, some unsolved problems can be pointed out with respect to the specification of complex policies over EHRs, where access should be granted or denied according to the right and the need of the healthcare workers to perform a particular job function on specific EHR sections.

The first point, which represents one of the largest potential obstacles, regards the difficulty of forcing narrative text to assume a semi-structured coded form into EHRs in order to build fine-grained access control policies on the top of its specific sections, coupled with the defensive attitude of physicians towards the use of computerized health information systems. Clinicians have a long tradition of using paper forms and dictation services, showing an enduring preference for narrative data (i.e. clinical text written in a natural language, such as Italian or English), due to advantages, such as familiarity, ease of use and freedom to express anything they wish (Johnson et al., 2008). Indeed, natural language provides many mechanisms that augment or enrich simple facts, for example to qualify their severity or degree, convey temporal relationships, indicate patterns of causality, provide rationale, propose hypotheses, and suggest alternatives (Johnson et al., 2008). Moreover, producing data without following structural rules can result faster in many situations. Furthermore, healthcare workers are not motivated to indicate a structured or semi-structured form for the data they produce, since scarcely aware of the possible advantages, such as automatic document processing or a more effective data protection. As a consequence of that, a fundamental requisite for efficiently and accurately securing healthcare records is to automatically transform narrative data into semi-structured EHRs, suitable for machine processing.

Secondly, with reference to existing access control models and mechanisms, many efforts have been made to meet the specific security and privacy needs of the healthcare domain. In parallel, and almost separately, general-purpose policy languages for access control have been defined, without tying to a model, in order to improve the

usability and simplicity for the designers. Indeed, existing models may not have the machinery to express all the policy details of a given system or may deliberately leave important aspects unspecified (Finin et al., 2008). To face such an issue, formal access control models and policy languages should be harmonized in the sense that advanced access control concepts should be embodied into a model as well as integrated and supported by a policy language in a natural intuitive manner (Amato et al., 2010). In particular, policy languages should allow restrictions to be described over a sharable and semantically well-defined domain model to promote common understanding among healthcare workers and support automatic reasoning about them.

Finally, since access control models and policy languages require a good expressive power, they typically suffer from a resistance amongst real users, even those entrusted with the management of access control policies, due to their highly complex syntax. Even though they are not required to be generally written for widespread use, this practical issue strengthens the convincement that different solutions should be used for encoding access control policies for healthcare systems without assuming an overly high level of theoretical ability. As a consequence of that, one prerequisite for the broad acceptance of them and their efficient application to medical settings is the guarantee of a high level of upgradability and maintainability, (i) to modify access control policies according to dynamically changing security needs, or (ii) to adapt generic, site-independent access control policies to a specific healthcare organization. Since updating policies can require a continuous intervention, it is unthinkable that it cannot be done directly by security administrators when needed. In contrast to the intensive efforts made to develop access control models and policy languages, the issue of providing solutions for easily editing and upgrading access control policies has been widely neglected thus far.

In order to face all these open issues, this paper proposes a hybrid framework aimed at enabling and supporting the definition of fine-grained access control policies working on semi-structured electronic health records. In more detail, the strong points of the proposed framework can be summarized as follows.

First, the framework provides a semantic-based method that hybridizes linguistic and statistical techniques in order to give a semi-structured form to narrative text to be inserted into an EHR, by first identifying relevant concepts in textual data and, successively, delimiting the sections composing the document depending on the concepts recognized.

Second, an authorization model for role-based access control (hereafter, RBAC) with respect to EHRs has been proposed. It extends the National Institute of Standards and Technology (hereafter, NIST) RBAC reference model to regulate the access to EHRs, working not only at a document level, but also with respect to their different sections. In other words, it is aimed at determining the authorization decisions that enable healthcare workers to carry out specific tasks on the different sections of EHRs. The framework supports this RBAC model by using Semantic Web technologies and, in particular, a high-level ontology that expresses the elements of the proposed RBAC model, and, in addition, a domain-specific ontology that captures the features of a specific application domain.

Third, this framework also includes a procedural policy language to encode access control restrictions in the form of "*if–then rules*" built on the top of the ontological formalizations of the elements belonging to the proposed RBAC model. Plus, a set of patterns has been defined for supporting the simple insertion and editing of such access control restrictions with the aim of reducing the complexity of the formalization process, by graphically guiding the definition of policies that could be functional in the context of healthcare organizations and enabling their automatic encoding into machine executable languages.

A prototype implementation of this framework has been realized in the form of a system offering simple and intuitive interfaces to the security administrators who do not have a deep technical expertise. It provides a set of facilities for structuring narrative text into an EHR and writing access control policies, by hiding the syntax constructs used in both the proposed RBAC model and policy language.

The rest of the paper is organized as follows. Section 2 introduces an overview of the state-of-the-art solutions for building access control policies on the top of semi-structured EHRs and clearly highlights the motivations for this work and its research contribution. Section 3 depicts the proposed framework. Section 4 outlines the prototype implementation, whereas Section 5 reports an experimental evaluation over real documents contained into EHRs,

i.e. discharge summaries, and shows a proof of concept application about how to use the framework for structuring these documents and inserting policies with respect to their sections. Finally, Section 6 concludes the work.

2. Background and related work

This section discusses research related to the realization of solutions for building access control policies on the top of semi-structured EHRs. In detail, first, a description of most relevant existing techniques for structuring narrative text into EHRs is given. Successively, an overview of the existing access control models and policy languages is reported. Then, some examples of semantic-based approaches integrating access control models and policy languages are described. For each of these themes treated, main characteristics and drawbacks are outlined in order to better motivate the proposed solution. Finally, the motivations for the proposed work are clearly stated and its research contribution is diffusely discussed.

2.1. Information Structuring

Generally speaking, a health record is the product of a communicative act resulting from a process of collaboration between an author (e.g. a physician), and a reader (e.g. a physician or a nurse): the former uses language signs to codify health record meanings, the latter decodes these signs and interprets their meaning by exploiting the knowledge of:

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

This implies that the comprehension of a particular concept within a specialized narrative text contained into a health record requires information about the properties characterizing it as well as the ability to identify the set of entities the concept refers to. As a result, both the subjectivity of domain knowledge and the interpretation given by the author with respect to final readers make the structuring of narrative text a thorny task to be performed. However, structuring narrative text into EHRs has many advantages, including an explicit structure, compact and lossless storage, easy maintenance, efficient retrieval and fast transmission.

Up to now, there has been extensive research on structuring narrative text (Mao et al., 2003), encompassing a wide range of interdisciplinary methodologies which combine computer science, logic, mathematics, linguistics and others. In detail, existing solutions rely on different techniques to analyze texts and automatically extract relevant information (Gomez, 1998; Wu, 2013).

Many of these techniques utilize linguistic-based approaches, embracing Natural Language Processing and Computational Linguistics (Kennedy et al., 1998), to gain an understanding of the text, others employ statistical or pattern-matching based methods (Butler et al., 2003) for analyzing specialist or sectorial narrative texts, leading to the development of specific disciplines, like Corpora Linguistics, Textual, and Lexical Statistics (De Mauro et al., 1993), or for mining information from texts and supporting document categorization (Vrusias et al., 2009).

Linguistic-based techniques involve low-level activities, such as *tokenization*, which segments sentences and identifies minimal units of text, named *tokens* (e.g. words, word particles, abbreviations, acronyms, alphanumeric expressions punctuations, etc.) and *normalization,* consisting in handling variations of the same lexical expression in order to obtain a unique representation, by harmonizing spelling and capitalization. Moreover, *Part-of-Speech* (hereafter, *POS*) *tagging* identifies the part of speech of each word within a narrative text and categorized accordingly as a *content word* (e.g. nouns, verbs, adjectives and adverbs) or *functional word* (e.g. articles, prepositions and conjunctions) (Kennedy et al., 1998). Finally, *lemmatization* identifies the lemma (i.e. a dictionary form) of all the inflected forms of individual text tokens (i.e. diagnose, diagnosing, diagnoses, and diagnosed are

all forms of diagnose). This leads to the identification of proper nouns as well as noun and verb phrases representing entities, concepts, events, and their relationships contained within a narrative text.

Standard statistical techniques use mathematical models to determine whether a word or phrase is a term that characterizes the target domain. To achieve this goal, they measure *unithood* and *termhood* as the "degree of strength or stability of syntagmatic combinations and collocations" and "degree that a linguistic unit is related to domain-specific concepts", respectively (Kageura et al., 1996). Unithood is only relevant to complex terms (i.e. multi-word terms), while termhood deals with both simple terms (i.e. single-word terms) and complex terms.

On the one hand, most of the existing techniques for measuring unithood employs conventional measures such as mutual information (Church et al., 1990) and log-likelihood (Dunning, 1994), and simply relies on the occurrence and co-occurrence frequencies from local corpora as source of evidence. Mutual information measures the co-occurrence frequencies of the constituents of complex terms to assess their dependency, whereas, log-likelihood attempts to quantify how much more likely the occurrence of one pair of words is than the other (Wong et al., 2008).

On the other hand, most of the existing approaches for evaluating termhood makes use of distributional behaviour of terms in documents and domains, and some heuristics related to the dependencies between term candidates or constituents of complex term candidates (Wong et al., 2008). Common measures for weighting terms employ frequencies of occurrences of terms in the corpus. They identify sets of terms or keywords that are collectively used to represent the content of documents, by assigning a weight for each term, which measures the importance of a term in a document. There are various implementations, but the most common one is the classical Term Frequency-Inverse Document Frequency (hereafter, TF-IDF) and its variants (Salton et al., 1988).Statistical and pattern-matching techniques have been also used to group documents into clusters or to map individual documents or parts of them to pre-defined topic categories (Decherchi et al., 2009). Algorithm types used in these methodologies include Bayesian Probability, Neural Networks, Support Vector Machines, and K-Nearest Neighbors, explanations of which are beyond the scope of this paper. Moreover, ontology learning methods, based on natural language processing, formal concept analysis and clustering, have been also developed to address the problem of automatically building conceptual structures out of large text corpora in an unsupervised process (Cimiano, 2006). In addition, a plethora of ontology learning frameworks has been developed in the last decade, such as OntoLearn (Velardi et al., 2005), OntoLT (Buitelaar et al., 2004), Terminae (Aussenac-Gilles et al., 2007) as well as TextToOnto (Maedche et al., 2001) and its successor Text2Onto (Cimiano et al. 2005), and integrated with standard ontology engineering tools. All these frameworks implement various and different ontology learning methods, but a detailed discussion and comparison is out of the scope of this paper.

However, the integration of some of these different techniques sufficiently general to be used in many domains as well as their customization and instantiation to face the specific application requirements pertaining the medical domain still constitute open issues. In particular, the main necessity is to design a reconfigurable solution able to handle the heterogeneity of existing health records and provide semi-automatic procedures for defining the peculiar lexicon that better represents the specific domain of interest.

2.2. Access Control Models

Generally speaking, access control in information systems consists into a set of rules determining which users are allowed to read, execute, share, modify specific elements in the system with the final aim of ensuring that only authorized access can take place (Rosero et al.,2011). Mandatory Access Control (hereafter, MAC), Discretionary Access Control (hereafter, DAC), and RBAC are well-established access control mechanisms. Each of them was designed to overcome limitations found in its predecessor.

MAC controls the access on the basis of the security classification of subjects and objects in the system, so as to result well suited to military style applications, where there is a strict ordering to both subjects and privileges. However, for many applications, including healthcare ones, MAC is simply too restrictive (Eyers et al., 2006).

DAC schemes offer much more flexibility since they restrict access to objects according to the ownership of a resource. They generally represent an access control matrix indicating, for each subject in the system, which objects can be accessed and the specific mode of access. The drawback with DAC schemes is their lack of manageability due to this splitting up of the access control matrix (Eyers et al., 2006).

Moreover, in both these models, access permissions are managed based on individual users, so introducing complexity and cost to manage a large-scale system (Le et al., 2012).

RBAC is an alternative to MAC and DAC for simplifying security administration by introducing the role abstraction. According to this model, permissions are associated with roles, whereas users are assigned to appropriate roles, so as to ensure that only authorized users are given access to certain data or resources.

For two decades, RBAC has been widely used owing to its salient features such as generalization, simplicity, and effectiveness (Le et al., 2012). RBAC mechanisms have been employed in many developments for both academia and industry in order to satisfy the unique requirements on information access management of healthcare domains. However, progress to date has not been sufficient to meet the special domain needs of a fine-grained access control mechanism able to indicate policies associated to the specific sections composing an EHR.

2.3. Policy languages

Two parallel themes in access control research are prominent in recent years. The first one has focused efforts to develop new access control models to meet the policy needs of real world application domains and has led to several successful, and now well established, models such as the ones reported in the previous sub-section (Finin et. Al, 2008). However, with reference to RBAC, for instance, it is difficult to apply that model when roles cannot be assigned in advance and it is typically not possible to change access rights of a particular entity without modifying the roles (Finin et. Al, 2008). Using policy languages allows access rights to be associated with different credentials and properties of entities, and not with roles alone.

As a consequence of that, the second research theme has developed policy languages for access control, such as Platform for Privacy Preferences (hereafter, P3P), Enterprise Privacy Authorization Language (hereafter, EPAL) and eXtensible Access Control Markup Language (hereafter, XACML). P3P policies are expressed in a standard format that can be automatically retrieved and interpreted by agents. Typically used for privacy practices of websites, they are not sufficiently fine-grained and expressive to handle the description of privacy policies at the implementation level (Martino et al. 2008). EPAL and XACML are more flexible policy languages, and, in particular, EPAL is proposed to encode enterprise's privacy-related data-handling policies and practices, which can be imported and enforced by a privacy-enforcement system, whereas XACML is a widely adopted access control model based on XML (Martino et al. 2008).

The policy languages described above have some common limitations that make them difficult to be applied to distributed and dynamic environments, such as healthcare ones. They do not: i) give some way of expressing both the elements and behaviour of an access control model leaving the designer too much freedom and no guidance; ii) allow restrictions to be described over a sharable semantic structure so as to promote common understanding among healthcare workers; iii) support automatic reasoning in order to dynamically apply new restrictions to a particular entity.

Some attempts have appeared in the last years to face these limitations, such as the framework proposed in (Croitoru et al., 2007) for expressing policy rules on the top of Layered Conceptual Graphs, i.e. a visual, structured, logic based knowledge representation formalism supporting reasoning procedures for ensuring validation, reuse and consistency. Hierarchical knowledge can be easily depicted and reasoned upon, so enabling the formalization of high level policy rules and their interdependencies. Yet, even though Layered Conceptual Graphs provide a major expressivity, the potential for depicting policy associations rigorously and employing correlated reasoning capabilities, they do not enable to specify a clear and extremely expressive semantics, if compared with other knowledge representation formalisms, such as ontology models, so leading to final policy specifications that are scarcely comprehensible and usable for non-technical experts.

2.4. Semantic-based frameworks for access control

Some approaches have been proposed to support the definition of policies on the top of ontology models, which are formal specifications of conceptualizations, described by axioms and explicit definitions and widely used for the precise semantic representation of concepts from different areas of knowledge (Herre et al., 2006). As a consequence of that, these approaches rely on the common idea of encoding basic elements composing an access control model by means of an OWL ontology.

OWL[1] is the actual W3C standard for expressing ontologies and consists into an XML-based language with a well-defined semantics grounded in Description Logics (DL). DLs are based on the notion of concepts (unary predicates, to be not confused with the concepts of the proposed RBAC model) and roles (binary relations, to be not confused with the roles of the proposed RBAC model), and are mainly characterised by constructors that allow complex concepts and roles to be built from atomic ones (Horrocks et al., 1999). The sets of concepts and roles represent the Terminological Box T (TBox) of the ontology, whereas the instances (or individuals) of the elements in the TBox represent the Assertional Box (ABox) of the ontology. Since, from a formal point of view, OWL can be seen to be equivalent to very expressive DLs, with an OWL ontology corresponding to a DL terminology, a set of sound and complete algorithms offered by DLs can be exploited for interesting inference problems such as subsumption and satisfiability of concepts (Horroks et al., 1999).

A first approach exploiting the ontology formalism is represented by KAoS (Uszok et al., 2004), which is a rich component-based framework for expressing, administrating and enforcing policies. It offers a high level ontology integrated with rules to explicitly provide constructs for modeling processes and transactions. Another framework based on ontologies is Rei (Kagal et al. 2006), which is a distributed solution to share and compose access control policies. It reuses reference policies and adapts them to its needs. However, none of these frameworks can fully reason about security policy based on the semantic relationships in the conceptual level.

Some attempts, such as the ones by (Damiani et al., 2004; Priebe et al. 2007), have been done to employ semantic relationships in their policy language. Both these approaches extend the XACML framework with ontologies to capture semantic relationships between attributes but, similar to the core XACML, they do not support automatic reasoning for policy inference.

A further solution is proposed in (Minutolo et al., 2012), which is a semantic-based system offering a set of patterns for easily editing rules built on the top of ontological vocabularies. Even if this system has been devised to guide and assist the creation and formalization of condition-action clinical recommendations for knowledge-based Decision Support Systems, it has a general basis and, thus, it can be reused in other fields of application and, in particular, also to encode access control policies on the top of ontology models. As reported in the following sections, the system proposed in (Minutolo et al., 2012), together with some aspects of the underpinning approach, has been customized and extended to realize the proposed framework for access control.

2.5. Motivations and research contribution

The analysis of the state-of-the-art solutions aforementioned gives the motivations and justifies the rationale for the approach proposed in this work, which can be summarized as follows.

As clearly highlighted in the previous subsections, to the best of our knowledge, none of the existing frameworks or tools for access control in medical settings is specifically concerned with the simplified and intuitive editing of access control policies working only on portions of EHRs, neither system-oriented researches appear to have been developed in that direction (Fernández-Alemán et al, 2013).

Indeed, firstly, existing systems enable the construction of access control policies by requiring more advanced capabilities at the price of being scarcely understandable and usable for a non-technical expert. They often speak the language of the access control models and policy formalisms supported, rather than providing some facilities

[1] *Ontology Web Language*: http://www.w3.org/TR/owl-features/

for reducing the gap between these formalisms and the language of the domain for which policies should be developed. This aspect demands a deeper insight into the underlying formalisms, and thus, highlights the lack of a good usability.

Secondly, they consider the protection object, i.e. healthcare data, as primarily document-centric rather than data-centric. Nevertheless, EHRs are composed of several portions characterized by different types of content, which often involve highly sensitive information. This point demands security mechanisms able to make only portions of EHRs, containing sensitive elements about a patient, accessible to authorized users, such as physicians or family members.

In conclusion, the need is emerged for an access control system expressly devised to guide the insertion and editing of policies built on the top of an understandable access control model. Such a model should allow regulating users' actions, in a very simple and familiar fashion, by operating also on portions of EHRs, opportunely arranged in a semi-structured form. The simplicity of usage should contextually grant the correctness of the policies edited by supporting their proper encoding in terms of admissible and well-formed structure.

The most relevant research contribution to fulfil these security requirements consists in the definition of a customized semantic-based authorization model extending RBAC to regulate the access to portions of EHRs, a rule-based language and a set of patterns to simply encode policies on the top of this model, and a tool to author and update fine-grained access control policies over semi-structured EHRs. In order to give a semi-structured form to narrative text to be inserted into EHRs, a semantic-based method has been integrated that hybridizes linguistic and statistical techniques. The combination of all these features within an integrated framework, together with a prototype implementation in the form of a system offering simple and intuitive interfaces to the security administrators, makes the whole approach flexible, extensible and powerful enough to be well-suited to the healthcare domain. A detailed explanation pertaining all the features of the proposed framework is given in the following sections.

3. The proposed hybrid framework for access control on semi-structured EHRs

This section diffusely describes the proposed hybrid framework for enabling and supporting the definition of fine-grained access control policies working at a section-level on semi-structured EHRs, preliminary constructed starting from narrative text.

In the first subsection, a fully description of the approach and tools used for constructing semi-structured EHRs is given. In the second subsection, the proposed access control model, its formalization by means of an ontology and the language used for defining policies on the top of it are outlined.

3.1. Information Structuring

The approach adopted for structuring EHRs is essentially aimed at properly locating and characterizing resources in a narrative text by recreating the domain model to which that text pertains. In detail, terms convey the fundamental concepts of a specific knowledge domain: they have their realization within texts and their relationships constitute the semantic frame of both the documents and the domain itself. For this reason, the detection of a series of relevant and peculiar terms in a text allows determining the set of concepts that can be used to define features characterizing a resource.

In order to extract relevant terms from a text, the proposed approach hybridizes linguistic and statistical techniques. In particular, linguistic filters are applied to words in order to extract a set of candidate terms, whereas a statistical method is used to calculate word occurrences within a text and, consequently, assign a value measuring the "strength" or "weight" of a candidate term. Indeed, 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 do. The whole approach is outlined in Fig. 1 and diffusely described in the next subsections.

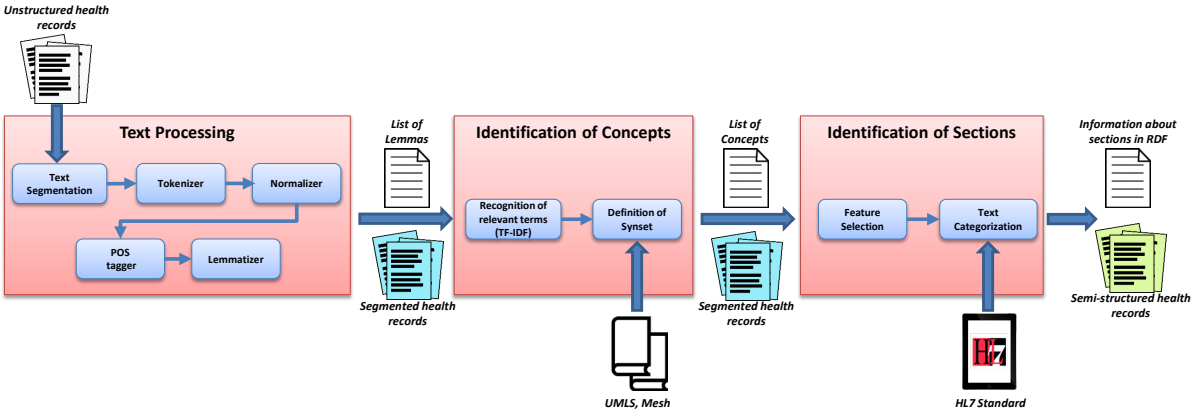

Fig. 1. The approach proposed for structuring EHRs from a narrative text.

3.1.1. Text Processing

The first stage of the proposed approach aims at segmenting, extracting and filtering text from unstructured health records in order to make it partitioned into coherent blocks and enriched with metadata specifying morphosyntactic information and citation form for each text element and, thus, suitable for automatic processing. It has been arranged in the form of a sequence of five basic techniques, namely *Text Segmentation*, *Tokenization, Normalization, POS Tagging, Lemmatization*, opportunely adapted to this specialist textual universe.

In detail, *Text Segmentation* consists in performing a complete global segmentation of an unstructured health record into distinct homogeneous regions by using features like punctuation marks or whitespaces. Specifically, narrative health records here considered are single columned documents not including graphics and photographs, and are composed of one or more blocks, each of which belongs to a coherent section of the document. One block corresponds to a set of text lines with the same typeface, a consistent line spacing and ending with known punctuations such as ".", "…", "!", "?".

The algorithm here adopted iteratively examines each line of the narrative health record in turn, from top to bottom. Lines are merged into complete blocks according to heuristic rules using their contents, typography information, or both. Details of the heuristic rules are not discussed here for the sake of brevity. As an example, in the following, a heuristic rule defined for segmenting a block is described in terms of its conditions which have to be simultaneously verified:

- **Condition 1**: The first text line in the block corresponds to either a new line or a normal text line.
- **Condition 2**: The last text line in the block is neither a new line nor a centered line.
- **Condition 3**: The last text line in the block is ended by a known punctuation.
- **Condition 4**: All text lines except for the first and last in the block are normal text lines.

Successively, *Tokenization* has been applied to each block, once it is segmented, and it has been realized 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 synergic combination of glossaries and mini-grammars has been motivated by the need of a high level of accuracy, even in presence of texts with acronyms or abbreviations that can increase the error rate (Butler et al., 2003). *Tokenization* has been further partitioned into the following sequence of phases: i) *grapheme analysis*, to define the set of alphabetical signs used within a block of a segmented health record ; ii) *disambiguation of punctuation marks,* aimed at realizing the token separation; iii) *separation of continuous strings,* to recognize strings not separated by

blank spaces; iv) *identification of separated strings,* to be considered as complex tokens and, therefore, single units of analysis.

Normalization has been automatically performed by first comparing a block of a segmented health record to external lexical lists in order to recognize and standardize particular expressions (like well-known abbreviations and acronyms, toponyms, as well as grammatical phrases and specific noun phrases) and, successively, setting proper parameters in order to uniform the different forms (e.g. reduction of capital letters into small letters according to some pre-arranged conditions, such as a capital letter used after a punctuation mark identifies the beginning of a sentence).

POS Tagging has been realized by using *Key-Word In Context* (hereafter, *KWIC*) *Analysis*, i.e. a systematic study of the local context where the various occurrences of a lexical item appear (Decherchi et al., 2009), in order to provide a procedure for *word-category disambiguation.* In detail, occurrences of each concept are computed in the text and co-text (i.e. the textual parts before and after it). The analysis of the co-text allows detecting the role of the words in the phrase in order to disambiguate their grammar category. As a result, ambiguous forms are first associated to the set of possible POS tags, and then disambiguated by adopting 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. After being categorized, words are enriched with additional morphological specifications, such as inflectional information[2].

Finally, *Lemmatization* has been implemented by introducing a partitioning scheme establishing an equivalence class on the list of tagged terms in order to reduce all the inflected forms to the respective lemma coinciding with the singular male/female form for nouns, the singular male form for adjectives and the infinitive form for verbs.

3.1.2. Identification of Concepts

The second stage is aimed at identifying relevant concepts for each block of a segmented health record and organizing them in synsets, i.e. lists of terms that are considered semantically equivalent for the purposes of information retrieval (Amato et al. 2011). In more detail, preliminarily, the vocabulary of relevant terms from a block is extracted. It is worth remembering that some words are semantically more significant to describe resources, and among these words, some lexical items weight more than others do.

In the proposed approach, the semantic relevance is assessed by TF-IDF index (De Mauro 1993), computed over the corpus vocabulary based on term frequency and term distribution within the corpus. This information enables the selection of relevant terms, by filtering all terms having an associated index value under an empirically established threshold. The set of selected terms constitutes the peculiar lexicon used to define features in classification tasks.

On the extracted peculiar lexicon, lexicometric analyses are then applied in order to evaluate the rate of coverage of the extracted terms over the vocabulary of the input block (Butler et al., 2003). In the case that the coverage rate results inadequate, the whole process is reiterated by enlarging the empirically established thresholds in order to extract terms with lower associated indices.

Once relevant terms are detected, this stage proceeds to clusterize them in *synsets*, in order to associate the proper concept to the list of terms denoting it. In this way, it is possible to refer a concept independently of the particular term used to denote it. Each concept, then, is referred by a list of terms representing it, codified in a synset. The clustering has been performed by integrating two external resources: the medical ontology given by "UMLS[3]" and "Mesh"[4], a thesaurus of medical terms. The adoption of specialized external resources has a double purpose, i.e. *endogenous*, since inside the documental base, the same concepts can be referred by different terms,

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

[3] *Unified Medical Language System (UMLS)*: www.nlm.nih.gov/research/umls/

[4] *Medical Subject Headings of National Library of Medicine*: www.nlm.nih.gov/mesh/

and *exogenous*, since in a natural language query, a given concept can be denoted by using terms that are different from those occurring in the documental base.

3.1.3. Identification of Sections

The last stage is in charge of performing a text categorization aimed at assigning labels to all the blocks of the input health record depending on the presence/absence of concepts in them (Amato et al, 2013).

This categorization has been performed by using features extracted from each block as inputs to a combination of supervised linear classifiers, namely Naïve Bayes, Decision Tree and K-Nearest Neighbour (Grilheres et al. 2004). For each block of the EHR, the feature space is represented by the set of all the concepts included in the synset and appearing in the block itself, with the number of their occurrences in it as values. In other words, the bag-of-words representation is used, since each block is represented with a vector of the concept counts that appear in it.

In order to make the use of aforementioned classifiers possible, and, contextually, improve generalization accuracy and avoid "overfitting", a feature selection method based on term frequency has been preventively applied to reduce the high dimensionality of the feature spaces and select the most representative features. In particular, this method makes use of the TF-IDF index, calculated for each concept included in the synset, as evaluation metric to measure the ability of each concept to differentiate each section. With reference to the possible sections used in the text categorization, they assume values such as *personal data* or the *medical history* of a patient, the *diagnosis* of a disease or the *treatment* described in terms of drugs and doses and so on, according to the HL7 standard for EHRs (Dolin et al., 2006).

Each single classifier has been trained in order to learn the most predictive values for the features belonging to the set of the ones preliminarily selected and that can characterize a specific section. All the considered classifiers require only a small amount of labelled trained data as input and can be successively evaluated, with respect to their effectiveness, in a testing phase when previously unseen instances of data are considered. They are easy to construct and update since they require only subject knowledge and not programming or rule-writing skills.

In detail, the Naïve Bayes Classifier is constructed by using the training data to estimate the probability of each section among the set of possible ones given some feature values calculated for a new block. This probability is calculated by using the Bayes theorem, with the simplifying assumption of conditional independence, since the feature space contains more elements. In other words, the conditional probability of a concept given a section is assumed to be independent of the conditional probabilities of other concepts given that section. This classifier finally labels a block as the section with the highest probability.

The Decision Tree is constructed by using the C.4.5 algorithm among the possible learning ones. In particular, this algorithm computes a tree, where each internal node denotes a test on a feature, each branch represents an outcome of the test, and leaf nodes represent final categories, i.e. the sections. Gain ratio is used as splitting criteria, i.e. to select the set of features which best partition the different blocks into distinct sections, and, in addition, pruning is enabled to identify and remove branches reflecting noise or outliers in the training data.

Finally, K-Nearest Neighbour is used to classify a section based on the majority category amongst its k-nearest neighbours. It is based on learning by analogy, i.e. by comparing a given block with training data that are similar to it. In particular, closeness is defined in terms of a distance metric, i.e. the Euclidean distance. This algorithm computes the distances between a new block and all the previous ones already classified, sorts these distances in increasing order and selects the k blocks with the smallest Euclidean distance values. Finally, it assigns the new block to the largest section out of k selected ones.

The results produced by these different algorithms are then combined by means of a voting strategy. It is a very efficient strategy, since previous knowledge about the results that are being decided as well as a large set of data to be analysed are not required. In more detail, every vote has a fixed weight and a fixed probability of occurrence, being independent of the other types of voting. At the end of this process, the assigned output category is the one that gets the majority of votes. Its advantage is that it is almost impossible that more classifiers can produce the

same text categorization result, so almost all errors are reduced. This text categorization can be successively refined by domain experts, who can move or reallocate one or more blocks classified as belonging to a section.

The final output of this stage is a semi-structured EHR subdivided into one or more sections. The information about the different sections produced is coded in RDF[5] for being processed and secured by the framework as described in the next section.

3.2. Access Control

The approach adopted by the proposed framework to regulate the access to a semi-structured EHR with respect to the specific sections identified as just described, relies on a customized version of RBAC model. Such a model is encoded in the form of both a high-level ontology, that abstractly expresses its core elements, and a domain specific ontology that captures the features of a specific application domain. Moreover, it also includes a procedural policy language to encode access control restrictions in the form of "*if–then rules*", built on the top of the ontological formalization of the proposed RBAC model. Finally, a set of patterns has been defined for supporting the simple insertion and editing of such access control restrictions with the aim of reducing the complexity of the formalization process. The following subsections deeply describe all these features offered by the proposed framework.

3.2.1. A Customized RBAC Model

The authorization model here proposed for role-based access control with respect to the sections of semi-structured EHRs essentially relies on the existing National Institute of Standards and Technology RBAC reference model (Ferraiolo et al., 2001).

As preliminary introduced in Section 2, the key issue of RBAC relies on the capability of limiting actions or operations that a legitimate user of a computer system can perform (Sandhu et al., 1994), based on a set of authorizations indicating the user's privileges to access computer resources according to organizational roles. In other words, authorizations are not associated straight to specific users, but to roles. A role represents a job function within an organization that describes the authority and responsibility conferred on a user assigned to a role (Sandhu et al., 1996). It defines both the specific individuals allowed to access resources and the extent to which the resources are accessed (Sandhu et al., 1996) according to the "need-to-know principle", i.e. a user is allowed to access the resources required for performing a task assigned to the role he belongs to. Such a way, RBAC facilitates the management of access policies to confer on roles specific authorizations according to new requirements, since users can be simply moved from one role to another one.

The basic RBAC model has been customized to regulate the access to semi-structured EHRs at a section-level with the aim of determining the authorization decision that enables a medical user to carry out a task. In particular, this customization offers the possibility of specifying fine-grained access control mechanisms with respect to specific sections of a semi-structured health record, so as to allow a more customizable and accurate authorization policy specification, where access is granted or denied according to the right and the need of the medical user to perform a particular job function.

Fig. 2 illustrates the customized RBAC model that is composed of three main entities: S (subjects), R (roles) and P (policies). S represents the set of all the possible subjects, whereas R specifies the set of all the admitted roles. Furthermore, P indicates the set of all the policies, where each of them is described by means of one or more operations $opr \in Opr$, defined on one or more sections $sec \in Sec$ containing a list of concept $con \in Con$ of a semi-structured health record. Each policy is also coupled with the associated access rights $ar \in Ar$ indicating the conditions under which each operation can access a specific EHR section.

[5] *Resource Description Framework*: www.w3.org/RDF/

Subject-role (*SR*) and role-policy (*RP*) relations define n-to-m associations between subjects and roles, and between roles and policies, respectively. Furthermore, a role hierarchy (*RH*) is defined with a tuple $\langle R, \leq \rangle$, where R is the set of roles and $\leq$ is a partial order relation over the elements of R, devised to better fulfil different levels of authority and responsibility in a medical setting. Finally, a section hierarchy (*SecH*) is defined with a tuple $\langle Sec, \leq \rangle$, where *Sec* is the set of sections and $\leq$ is a partial order relation over the elements of *Sec*, devised to describe the different typologies of sections existing into a semi-structured EHR.

The static separation of duty (*SSoD*) constraint has been imposed during the definition of the relationships *SR* and *RP* to limit the possibilities of intentional or unintentional damage caused by excessive authorizations owned by a single subject. *SSoD* reduces the authorizations that can be applied to a subject, by means of the definition of mutually exclusive roles in *SR* and *RP* relations. In *SR*, the same subject cannot be assigned with two or more mutually exclusive roles, whereas, in *RP*, the same policy cannot be associated to mutually exclusive roles.

More formally, a *SSoD* constraint is a tuple $\langle X, C_i, C_j \rangle$ where X is the set of elements associated with the antecedent constraint set C_i and the consequent constraint set C_j. A constraint $\langle X, C_i, C_j \rangle$ is satisfied if, whenever $x \in X$ is associated with C_i, then x cannot be linked to C_j, as well. For instance, the *SSoD* constraint $\langle S, R_i, R_j \rangle$ specifies that the sets of subjects assigned to R_i and R_j must be disjoint, whereas the *SSoD* constraint $\langle P, R_i, R_j \rangle$ indicates that the sets of policies associated to R_i cannot be assigned to a mutually exclusive role R_j, as well.

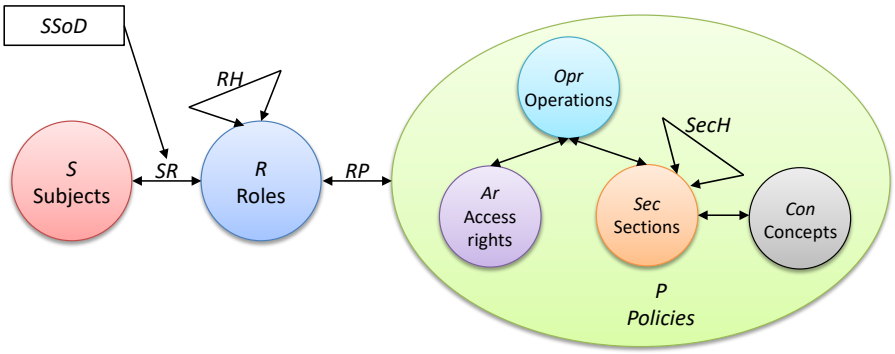

Fig. 2. The proposed RBAC Model.

On the top of this model, a set of policies can be defined in the form of a 4-tuple $\langle r, opr, sec, ar \rangle$, where $r \in R$ is the role; $opr \in Opr$ is the operation; $sec \in Sec$ specifies the section of a semi-structured health record to protect; $ar \in Ar$ specifies the access right, which can be positive or negative depending on whether or not an operation is allowed. For instance, the tuple $\langle doctor, write, diagnosis, + \rangle$ defines the following policy: a doctor is allowed to write the section pertaining the diagnosis of a semi-structured health record.

3.2.2. An ontology representation for the proposed RBAC Model

A semantic-based approach has been adopted to encode the proposed RBAC model in the form of an ontology, including the basic elements composing the model and the relationships between these elements. The choice of using an ontology is due to the possibility of defining a vocabulary semantically, by also specifying a set of modelling primitives, such as axioms about relations, which can be applied to encode the role hierarchy and the static separation of duty constraint. Moreover, this vocabulary can be also exploited to guide the policy construction by reducing the possibility of errors, as shown in the following subsection.

In more detail, the basic elements composing the proposed RBAC model have been encoded into a high-level ontology, so that the features of particular applications can be captured by means of more specific domain ontologies built on the top of the high-level one.

For the purposes of this paper, an ontology is formally defined as the tuple $O := \langle O^T, O^A \rangle$, where O^T is the TBox of the ontology and O^A is the ABox defined over the entities in O^T. Moreover, in the following, concepts and roles are denoted in **bold,** whereas individuals are indicated in *italics*. Moreover, assertions in the form x:**C** and **R**(x,y) state that "individual x is an instance of concept **C**" and "individual x is related to y by means of **R**", respectively.

3.2.2.1. The high-level ontology

The proposed high-level ontology is graphically described in Fig.3. In more detail, the set of subjects is represented by the concept **Subject**, whose individuals are the users the policies are defined on.

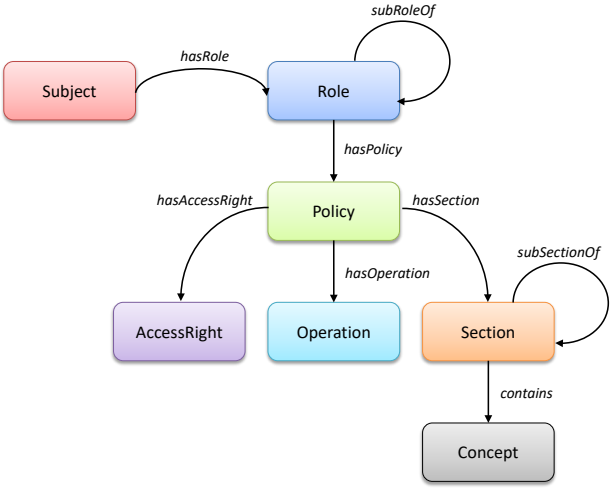

Fig. 3. Ontology for the Customized RBAC Model.

The role hierarchy defined with the tuple $\langle R, \leq \rangle$ is modelled by means of the concept **Role**, whose individuals represent all the possible roles. The $\leq$ relation is encoded by the role **subRoleOf**(**Role**, **Role**). This role is defined as transitive through the axiom **subRoleOf**$^{+}\sqsubseteq$ **subRoleOf,** which enables to infer that $\forall x,y,z \in$**Role**: **subRoleOf**(x, y), **subRoleOf**(y, z) $\Rightarrow$ **subRoleOf**(x, z).

According to the proposed model, each policy, defined as the 4-tuple $\langle r, opr, sec, ar \rangle$ is indicated by means of the concept **Policy,** whereas the concepts **Operation**, **Section** and **AccessRight,** are included to represent the sets of possible operations, sections and access rights, respectively. Finally, the concept **Concept** indicates the sets of possible concepts composing a specific section.

The concept **Policy** is linked to the concept **Role** through the role **hasPolicy**(**Role**, **Policy**). The role **isPolicyOf**(**Policy, Role**) is defined as inverse role of **hasPolicy**(**Role**, **Policy**) through the axiom **isPolicyOf** $\equiv$**hasPolicy**$^{-}$, which means that $\forall x \in$**Role,** $y \in$**Policy**: **hasPolicy**(x,y) $\Rightarrow$ **isPolicyOf**(y,x).

Moreover, the concept **Policy** is also connected to the concepts **Operation**, **Section** and **AccessRight** by three roles, namely **hasOperation**(**Policy**, **Operation**), **hasSection**(**Policy**, **Section**) and **hasAccessRight**(**Policy**, **AccessRight**).

The role **hasOperation**(**Policy**, **Operation**) is defined as functional through the axiom $\top \sqsubseteq \leq 1$**hasOperation,** whose semantics is $\forall x \in$**Policy,** y,z $\in$**Operation**:**hasOperation**(x,y), **hasOperation**(y,z) $\Rightarrow x = z$. Similarly, the role **hasSection**(**Policy**, **Section**) is defined as functional through the axiom $\top \sqsubseteq \leq 1$**hasSection,** whose semantics is $\forall x \in$**Policy,** y,z $\in$**Section**:**hasSection**(x,y), **hasSection**(y,z) $\Rightarrow x = z$. Finally, the role **hasAccessRight**(**Policy**,

AccessRight) is also defined as functional through the axiom $\top \sqsubseteq \leq 1$**hasAccessRight,** which means that $\forall x \in$**Policy,** y,z $\in$**AccessRight**:**hasAccessRight**(x,y), **hasAccessRight**(y,z) $\Rightarrow x = z$.

The section hierarchy defined with the tuple $\langle Sec, \leq \rangle$ is modelled by means of the concept **Section,** whereas the $\leq$ relation is encoded with the role **subSectionOf**(**Section**,**Section**). This role is defined as transitive through the axiom **subSectionOf**$^{+} \sqsubseteq$ **subSectionOf,** which enables to infer that $\forall x,y,z \in$**Section**: **subSectionOf** (x,y), **subSectionOf** (y, z) $\Rightarrow$**subSectionOf** (x, z).

The concept **Section** is also linked to the concept **Concept** through the role **Contains**(**Section**,**Concept**). The role **isContainedIn**(**Concept, Section**) is defined as inverse role of **Contains**(**Section**,**Concept**) through the axiom **isContainedIn** $\equiv$**Contains**$^{-}$, which means that $\forall x \in$**Section,** $y \in$**Concept**: **Contains**(x,y) $\Rightarrow$ **isContainedIn**(y,x).

Finally, *SR* and *RP* relations in the proposed RBAC model are specified with the roles **hasRole**(**Subject**, **Role**) and **hasPolicy**(**Role, Policy**), respectively. The *SSoD* constraint $\langle S, R_i, R_j \rangle$, defined on *SR* relation and indicating that the sets of subjects assigned to R_i and R_j must be disjoint, is specified via the following axiom on the role **hasRole**: ($\exists$**hasRole**.$\{R_i\}$ $\sqcap$**Subject**)$\sqsubseteq \neg$ ($\exists$**hasRole**.$\{R_j\} \sqcap$**Subject**).

On the other hand, the *SSoD* constraint $\langle P, R_i, R_j \rangle$, defined on *PR* relation and indicating that the sets of policies associated to R_i cannot be assigned also to the mutually exclusive role R_j, is encoded by means of the following axiom on the role **isPolicyOf**: ($\exists$**isPolicyOf**.$\{R_i\}$ $\sqcap$**Policy**)$\sqsubseteq \neg$ ($\exists$**isPolicyOf**.$\{R_j\} \sqcap$**Policy**).

3.2.2.2. The domain ontology for a semi-structured health record

The domain ontology, devised to model the specific elements of semi-structured EHRs, essentially populates the ABox O^A defined over the concepts and roles specified in the TBox O^T of the high-level ontology. In other words, the high-level ontology defines the TBox O^T by formulating statements about concepts and roles, while the domain ontology only specifies the Abox O^A by formulating statements about individuals, in the form of assertional axioms. In detail, the main concept assertions defined in the ontology are:

- *doctor, paramedic, clinicalResearcher*:**Role**
- *read, write*:**Operation**
- *allowed, denied*:**AccessRight**
- *hospitalAdmissionDiagnosis*, *treatmentPlan, medicalHistory*, *personalData*, *hospitalCourse*:**Section**
- *smoking, taxCode, citizenship, residence, physiology, pathology*:**Concepts**

These concept assertions have been further enriched by means of a set of more specific individuals, which are reported with respect to the concepts **Role** and **Section**:

- *cardiologist, radiologist, nurse, dietitian*:**Role**
- *familyHystory*, *historyOfPresentIllness, allergies*, *dischargeDiet, encounters*, *clinicalReminders*:**Section**

These more specific individuals have been connected to the main ones by means of role assertions so as to generate the role hierarchy $\langle R, \leq \rangle$ and section hierarchy $\langle Sec, \leq \rangle$ for the specific application domain. Some examples of role assertions are reported in the following:

- **subRoleOf**(*cardiologist, doctor),* **subRoleOf**(*radiologist, doctor)*
- **subRoleOf**(*nurse, paramedic),* **subRoleOf**(*dietitian, paramedic)*
- **subSectionOf**(*familyHystory, medicalHistory),* **subSectionOf** (*historyOfPresentIllness, medicalHistory)*
- **subSectionOf** (*allergies, medicalHistory),* **subSectionOf** (*dischargeDiet, treatmentPlan)*
- **subSectionOf** (*encounters, treatmentPlan),* **subSectionOf** (*clinicalReminders, treatmentPlan)*

Finally, according to the guidelines of the HL7 standard, individuals of the concept **Concept** have been connected to individuals of the concept **Section** by means of assertions of the role **isContainedIn,** as reported, for instance, in the following:

- **isContainedIn**(*taxCode, personalData),* **isContainedIn**(*residence, personalData),* **isContainedIn**(*citizenship, personalData);*
- **isContainedIn**(*smoking,MedicalHistory),* **isContainedIn**(*pathology, MedicalHistory),* **isContainedIn**(*physiology, MedicalHistory).*

3.2.3. The procedural policy language

The procedural policy language here proposed is aimed at encoding access control restrictions in the form of "if–then rules" built on the top of the ontological elements defined by the proposed RBAC model. Similarly to the solution proposed in (Croitoru, et al., 2008), this language allows representing different policies and protection requirements as rules. However, in addition, it enables, at the same time, to provide understandable specifications and clear semantics, since based on ontology models. Moreover, it also supports a good level of flexibility, since policies are constructed as decoupled from the proposed RBAC model.

Deeply speaking, each policy is expressed in the form *"if antecedents then consequents"* and consists in one or more statements which have to be verified with respect to some data instances, if they occur in the antecedent part of the rule, and have to be executed when the antecedent part is true, if they are placed in the consequent part. In particular, the antecedent part consists in one or more statements, composed of ontological elements of the proposed RBAC model, concatenated by conjunctive connectives. The consequent part of a rule consists in one or more statements, made of ontological elements of the proposed RBAC model, concatenated by conjunctive operators.

More formally, a rule is defined in the form:

$$\forall x_1 \ldots x_k \, (A_1 \wedge \ldots \wedge A_h) \rightarrow B$$

where A_i, B are atoms, $h \geq 0$, and $x_1 \ldots x_k$ are all the variables occurring in the formula $(A_1 \wedge \ldots \wedge A_h) \rightarrow B$.

Atoms in rules can assume the form **C**(*x*) or **R**(*x,y*), where **C** and **R** are, respectively, a concept and a role of the ontology, and *x, y* are either variables or individuals. An atom **C**(*x*) holds if *x* is an instance of the concept **C**, whereas an atom **R**(*x,y*) holds if *x* is related to *y* by the role **R**. Variables are treated as universally quantified, with their scope limited to a given rule. They are specified using the standard convention of prefixing them with a question mark (e.g., ?a). Each rule has to verify the condition referred to as "safety", that is only variables that occur in its antecedent may occur in the consequent, as well (Colantonio et al., 2012). Moreover, each individual occurring in its consequent has to be previously defined as instance of a concept in its antecedent. Such a definition has to be placed in the antecedent part since it represents a condition to be verified and not an assertion to be stated in the rule consequent.

In Fig. 4, the abstract syntax for rules has been described by means of the Extended BNF.

```
rule::= "if ("antecedent ") then (" consequent")";
antecedent::= "Antecedent(" {atom} ")";
consequent::= "Consequent(" {atom} ")";
atom::= ontologyConcept "("element")" | ontologyRole "("element ", "element")";
element::= ontologyIndividual | variable;
variable::="?"a |  "?"b | ... | "?"z;
```

Fig. 4. Abstract syntax for rules in Extended BNF.

A policy in the form of a rule asserting that a doctor is allowed to read the EHR section pertaining the hospital admission diagnosis would be written as shown in Fig. 5:

```
if (
    Subject(?a)
    hasRole(?a, Doctor) Role(Doctor)
    hasPolicy (Doctor, ?b) Policy(?b)
    hasSection(?b, hospitalAdmissionDiagnosis) Section(hospitalAdmissionDiagnosis)
    hasOperation(?b, Read) Operation(Read)
    AccessRight(Allowed)
)
then (
    hasAccessRight(?b, Allowed)
)
```

Fig. 5. An example of policy in the form of a rule.

Since the proposed rule language is extremely flexible and essentially general-purpose, a set of patterns has been defined for supporting the simple insertion and editing of the access control policies with the aim of reducing the complexity of the formalization process. In more detail, the approach proposed in (Minutolo et al., 2012) has been customized to be applied to this domain with the aim of identifying and modelling common and repetitive elements existing in the structures of some possible policies. In particular, the atoms occurring in both the antecedent and consequent parts of a rule encoding a policy and built on the top of the high-level ontology just described can be codified by exploiting a pattern named *relational.*

A *relational* pattern specifies a role between two entities and can be *self-contained*, if the entities involved are generic concepts, or *external-connected*, if one or both the entities are individuals of some concepts (Minutolo et al., 2012). Informally speaking, for instance, the statement "*a role has a policy*" is in accordance with the self-contained *relational* pattern, since both the entities involved are generic concepts, whereas the statement "*the role as doctor has a policy*" is in accordance with its external-connected form since the concept "*role*" is not generic but it is represented by the specific individual "*doctor*".

More formally, given the ontology $\mathcal{O}$, the set $\mathcal{C}$ of concepts defined in $\mathcal{O}$, the set $\mathcal{I}$ of individuals of concepts $\mathcal{C}$, the set $\mathcal{R}$ of roles defined in $\mathcal{O}$, a *relational* pattern specifies a role **R** between two concepts **C** and **D**. Moreover, it is satisfied by all the couples of individuals (*x, y*) defined in the ontology $\mathcal{O}$ and linked through the specific role **R**. Restrictions on the kinds of individuals which can belong to the domain and range of a given role have been formally specified, as well. All the forms of relational pattern used to build atoms in rules encoding policies are formally defined in the Table 1.

Table 1

Relational patterns according to an ontological perspective

Pattern	Syntax	Semantic
self-contained	$\mathbf{C} \sqcap \exists \mathbf{R}.\mathbf{D}$ where $\mathbf{C}, \mathbf{D} \in \mathcal{C}$, $\mathbf{R} \in \mathcal{R}$, $(1 \geq \mathbf{R}) \sqsubseteq \mathbf{C}, \top \sqsubseteq \forall \mathbf{R}.\mathbf{D}$	$x \mid (\exists y.\mathbf{R}(x,y) \wedge \mathbf{D}(y)) \Rightarrow \mathbf{C}(x)$
left-external-connected	$\{a\} \sqsubseteq \exists \mathbf{R}.\mathbf{D}$ where $\mathbf{C}, \mathbf{D} \in \mathcal{C}$, $\mathbf{R} \in \mathcal{R}$, $(1 \geq \mathbf{R}) \sqsubseteq \mathbf{C}, \top \sqsubseteq \forall \mathbf{R}.\mathbf{D}, a{:}\mathbf{C} \subseteq \mathcal{I}$	$x \mid (\exists y.\mathbf{R}(x,y) \wedge \mathbf{D}(y)) \Rightarrow x \in \{a\}$
right-external-connected	$\mathbf{C} \sqcap \exists \mathbf{R}.\{b\}$ where $\mathbf{C}, \mathbf{D} \in \mathcal{C}$, $\mathbf{R} \in \mathcal{R}$, $(1 \geq \mathbf{R}) \sqsubseteq \mathbf{C}, \top \sqsubseteq \forall \mathbf{R}.\mathbf{D}, b{:}\mathbf{D} \subseteq \mathcal{I}$	$x \mid (\exists y.\mathbf{R}(x,y) \wedge y \in \{b\}) \Rightarrow \mathbf{C}(x)$
external-connected	$\{a\} \sqsubseteq \exists \mathbf{R}.\{b\}$ where $\mathbf{C}, \mathbf{D} \in \mathcal{C}$, $\mathbf{R} \in \mathcal{R}$, $\mathbf{C}(1 \geq \mathbf{R}) \sqsubseteq \mathbf{C}, \top \sqsubseteq \forall \mathbf{R}.\mathbf{D}, a{:}\mathbf{C} \subseteq \mathcal{I}, b{:}\mathbf{D} \subseteq \mathcal{I}$	$x \mid (\exists y.\mathbf{R}(x,y) \wedge y \in \{b\}) \Rightarrow x \in \{a\}$

By exploiting this definition of patterns, the formalization of the atoms occurring in a rule is reduced to specify only atoms involving a role, so as to facilitate the rule encoding and minimize the possibility of errors.

Indeed, for instance, in case an atom verifying the self-contained relational pattern has to be inserted into a rule, only the role **R**, and the concepts **C** and **D** have to be indicated. As a consequence of this operation, an atom is built in the form **R**(?x,?y) and, in addition, two atoms in the form **C**(?x) and **D**(?y) are automatically generated and added to the rule if not already present.

Moreover, in case an atom verifying the left-external-connected relational pattern has to be inserted into a rule, only the role **R**, the concept **C** and an individual of it, and the concept **D** have to be indicated. The individual of **C** can be a concrete individual, defined in the ontology, or can refer to an existing variable defined in a previous atom in the rule. Thus, an atom is built in the form **R**(*a*,?y) or **R**(?v,?y), where the two notations indicate the presence of a concrete individual *a* or a reference to an existing variable *v.* In addition, two atoms in the form **C**(*a*)(or C(?v)) and **D**(?y) are automatically generated and added to the rule if not already present. It is interesting to note that, in case an atom verifying any relational pattern has to be inserted in the consequent part of a rule, two atoms, in the form **C**(*a*) or **C**(?y), depending on the type of pattern, are put in its antecedent part, since individuals or variables used in them have to be defined before being applied to generate the rule conclusion. Similar considerations can be done for the other two kinds of patterns proposed.

4. Implementation of the proposed framework

A prototypal implementation of the framework here presented has been realized by customizing and opportunely extending the system proposed in (Minutolo et al., 2012) in order to receive in input an unstructured narrative health record, identify the text sections composing it by applying the approach and tools for information structuring described above, codify them in a domain ontology for the role-based access control, by populating the ABox O^A defined over the concepts and roles specified in the TBox O^T of the high-level ontology, and finally allow to edit if-then rules, on the top of both these ontologies, for describing the policies to be verified according to the patterns just proposed by means of a graphical representation of them. The idea has been to provide a simple and intuitive tool to the clinicians, who do not have a deep technical expertise about both information structuring and ontology and rule formalisms, in order to support them in both the identification of the specific sections of an EHR and the definition of policies for access control on them.

The overall system interface is shown in Fig. 6, and it is organized as a Knowledge Tree (the left area), a Rule Editing Interface and an EHR Visualization Interface (the right area).

First, the Knowledge Tree contains concepts and roles defined in the high-level ontology and to be used in the rules. They are arranged in the form of nested trees, where each role is visualized under a concept only if the concept itself is contained in the domain restriction of the considered role. The knowledge representation formalism chosen for encoding the ontology is OWL. Specifically, the menu File allows creating/loading both the high-level and domain ontologies, loading an unstructured narrative health record, from a text file, and executing the whole process of information structuring for the identification of the sections of the document selected. The resulting semi-structured EHR is arranged in the EHR Visualization Interface as shown in Fig. 6, according to the sections identified and, in addition, its XML encoding can be visualized, as well. The menu Edit enables creation, enumeration, saving and deletion of policies in the form of if-then rules.

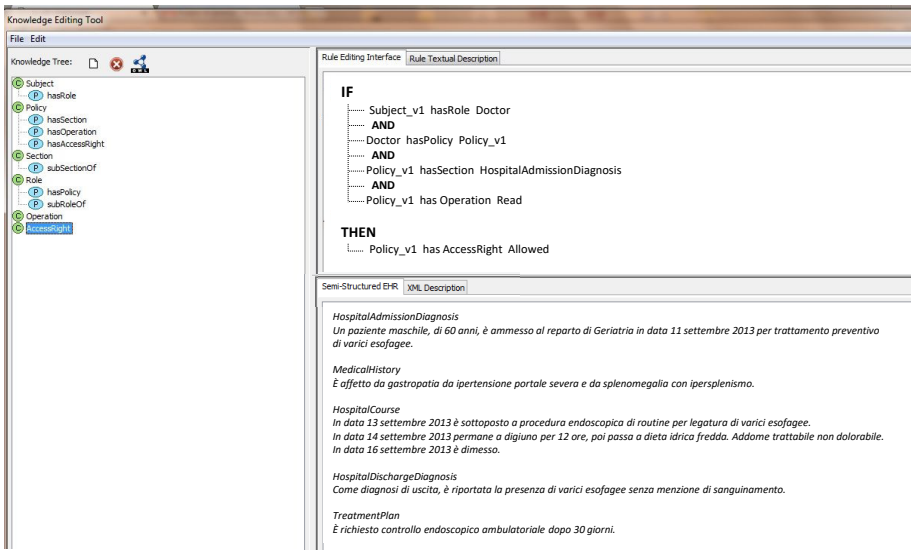

Fig. 6. The overall system interface of the prototypal implementation of the framework .

Each policy can be constructed by means of the Rule Editing Interface shown in Fig. 7. Both the antecedent and consequent parts are graphically arranged as nested trees, where the roots are respectively the nodes “IF” and “THEN” and the other nested nodes can be respectively a statement or a logical conjunctive connector. Statements added under the “IF” or “THEN” trees can be moved and placed in other positions by means of drag-and-drop operations.

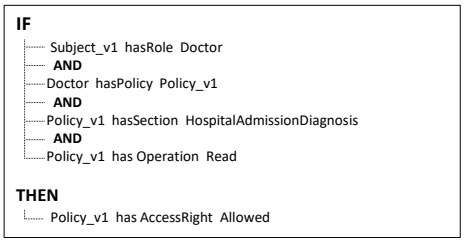

Fig. 7. Graphical representation of a policy in the form of an if-then rule.

The insertion in the rule of an atom verifying one of the relational patterns defined, where a role **R** exists between two concepts **C** and **D,** is carried out as follows. First, the role **R** in the Knowledge Tree is selected and, then, dragged and dropped in the Rule Editing Interface under the root "IF" or "THEN", depending on the fact that it has to be placed in the antecedent or consequent part of the rule. According to the kind of relational pattern, domain and range restrictions about the role **R** guide the user in the choice of the specific concepts **C** and **D,** defined in the TBox O^T of the high-level ontology and to be used in the atom. For instance, the atom *(Subject, hasRole, Role)* can be inserted by dragging and dropping the role “*hasRole*”, under the concept "*Subject*", in the Rule Editing Interface under the root "IF". As shown in Fig. 8, the atom can be easily completed by selecting the concept "*Role",* suggested by the system. Such a way, the resulting atom includes two generic variables, one for the concept "*Subject*", i.e. "*Subject_v1*" and one for the concept “*Role*", i.e. “Role*_v1*".

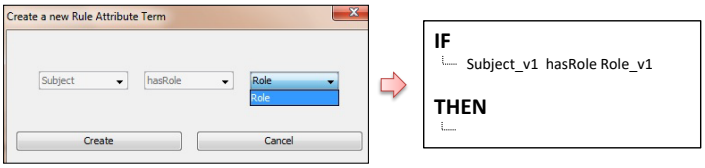

Fig. 8. The steps for creating a self-contained relational statement.

Moreover, in case one of the external connected patterns is chosen, an individual of **C**, **D** or of both of them have to be specified. The individual can be selected among the ones populating the ABox O^A and defined by means of the domain ontology or can be linked to an existing variable defined in a previous atom in the rule. For instance, the atom *(Subject, hasRole, Doctor)* can be inserted by dragging and dropping the role "*hasRole*", under the concept "*Subject*", in the Rule Editing Interface under the root "IF". As shown in Fig. 9, the atom can be easily completed by selecting the individual "*Doctor",* suggested by the system by listing all the instances of the concept *"Role"* in the ABox O^A defined by means of the domain ontology. Such a way, the resulting atom includes one generic variable for the concept "*Subject*", i.e. "*Subject_v1*" and one specific individual of the concept "*Role*", i.e. "*Doctor*".

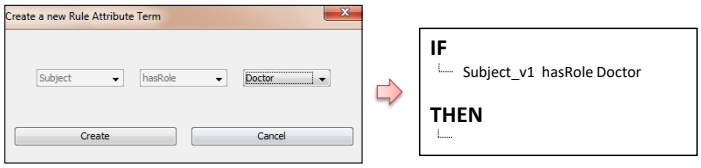

Fig. 9. The steps for creating a right-external-connected relational statement.

Finally, the system offers the possibility of simulating an actual role-based access control process on a semi-structured EHR by encoding each edited policy by means of the Jena framework[6]. In particular, such a framework offers a rule language that is practically used for implementing the policy language described above, and also a generic rule based reasoner that is adopted to test the policies encoded. In particular, example data encoded as OWL statements can be supplied to the reasoner and used to evaluate eligible policies among the ones inserted and infer all possible/not possible operations on the different sections identified on the unstructured narrative health record loaded in the system.

5. Experimental evaluation on discharge summaries

In order to concretely evaluate the proposed framework with respect to its functionalities for enabling and supporting the definition of fine-grained access control policies working on semi-structured EHRs, a set of 9890 real documents, named *Discharge Summaries* and contained in EHRs has been considered.

A discharge summary is a synopsis of a patient's admission to a hospital and provides pertinent information for the continuation of care following discharge. According to the Italian technical specifications for creating a discharge summary, the following set of mandatory or optional sections is contained in it:

- *Medical history*: history related to the patient's current and past complaints, problems, or diagnoses.
- *Hospital admission diagnosis*: primary reason for admission to a hospital facility.
- *Hospital course*: sequence of events from admission to discharge in a hospital facility.

[6] *Apache Jena*: http://jena.apache.org/

- *Relevant diagnostic tests and/or laboratory data*: the findings and interpretation of relevant diagnostic tests performed on patients in a hospital facility.
- *Hospital discharge studies summary*: results of observations generated by laboratories, imaging procedures, and other procedures.
- *Selected medicine administered during hospitalization*: relevant medications administered to the patient in a hospital facility.
- *Hospital discharge diagnosis*: relevant problems or diagnoses that are occurred during the hospitalization or that need to be followed after hospitalization.
- *Discharge medications*: medications that the patient is intended to take (or stop) after discharge.
- *Treatment plan*: data defining pending orders, interventions, encounters, services, and procedures for the patient as well as information regarding goals and clinical reminders.

The discharge summaries here considered are expressed in Italian, since they are real documents, opportunely anonymised, coming from an Italian hospital facility. However, the proposed framework is general enough to be applied also to health records formulated in other languages.

The framework has been firstly applied to that set of documents in order to assess the functionalities for structuring their contents with respect to the list of sections aforementioned. In particular, the gold standard to be used in this evaluation has been constructed as follows. The documents have been preliminarily segmented in blocks. Successively, each block has been manually classified, with the help of domain experts, as a specific section chosen among the aforementioned ones, depending on the concepts identified in it. After that, the tenfold cross validation method has been adopted for evaluating the automatic classification performed by using the proposed framework and its accuracy, calculated with respect to the gold standard, has been chosen as metric to estimate the goodness of the results achieved. This means that the whole set of discharge summaries is preliminarily divided into ten subsets of 989 documents, and ten training sessions are effected: in the i^{th} session, the i^{th} subset is kept for testing and the proposed framework is applied in the remaining 9 folds. The best solution found in each training session is then evaluated on the testing set. Finally, the average result on the testing set is computed over these ten sessions and produces a final value equal to 89,67%.

The goodness of this result can be considered as proof of feasible and efficient integration of different methods for structuring health records in terms of sections. The rate of misclassification is essentially due to the underlying ambiguity of the natural (although technical) language by which discharge summaries are written. In particular, many textual blocks contain sets of concepts that belong to more than one section and, for this reason, are wrongly classified. For instance, the whole Italian sentence "*ulcera duodenale sanguinante con anemizzazione secondaria*" can be associated as belonging to two sections, namely *Hospital admission diagnosis* and *Medical History.* Indeed, that sentence states that there is a bleeding duodenal ulcer with secondary anaemia, but it is not clearly expressed if that diagnosis is associated to a present or past critical condition affecting the patient.

After having tested the goodness of the functionalities for structuring EHRs, as a proof of concept, the implemented prototype has been applied to the fragment of a real discharge summary, as reported in Fig, 10, in order to deeply show how all the functionalities of the proposed framework operate on a health record.

```
Un paziente maschile, di 60 anni, è ammesso al reparto di Geriatria in data 11 settembre 2013 per trattamento
preventivo di varici esofagee. È affetto da gastropatia da ipertensione portale severa e da splenomegalia con
ipersplenismo. In data 13 settembre 2013 è sottoposto a procedura endoscopica di routine per legatura di varici
esofagee. In data 14 settembre 2013 permane a digiuno per 12 ore, poi passa a dieta idrica fredda. Addome
trattabile non dolorabile. In data 16 settembre 2013 è dimesso. Come diagnosi di uscita, è riportata la
presenza di varici esofagee senza menzione di sanguinamento. È richiesto controllo endoscopico ambulatoriale
dopo 30 giorni.
```

Fig. 10. A fragment of a discharge summary in Italian.

The considered document states that a male patient, aged sixty, was admitted to the ward of Geriatrics in date 09/11/2013 for a preventive treatment of oesophageal varices. He was affected by gastropathy due to severe portal hypertension and splenomegaly with hypersplenism. On 13/09/2013, he underwent a routine endoscopic procedure for ligation of oesophageal varices. On 14/09/2013, he remained at fasting for twelve hours, as a medical indication, and, then, he switched to a diet based on cold water. His abdomen was treatable but not tender. On 16/09/2013, he was discharged. As discharge diagnosis, the presence of oesophageal varices was indicated without any mention of bleeding. An outpatient endoscopy was required after 30 days.

As described in Section 3, the first step for structuring that narrative fragment of discharge summary is Text Processing. In more detail, *Text Segmentation* is first applied, producing eight different text blocks, and, then, *Text Tokenization* and *Normalization* are effected, producing the results reported in Table 2.

Table 2

Results of Text Tokenization and Normalization

	Text
Segmented	*Un paziente maschile, di 60 anni, è ammesso al reparto di Geriatria in data 11 settembre 2013 per trattamento preventivo di varici esofagee.*
	È affetto da gastropatia da ipertensione portale severa e da splenomegalia con ipersplenismo.
	In data 13 settembre 2013 è sottoposto a procedura endoscopica di routine per legatura di varici esofagee.
	In data 14 settembre 2013 permane a digiuno per 12 ore, poi passa a dieta idrica fredda.
	Addome trattabile non dolorabile.
	In data 16 settembre 2013 è dimesso.
	Come diagnosi di uscita, è riportata la presenza di varici esofagee senza menzione di sanguinamento.
	È richiesto controllo endoscopico ambulatoriale dopo 30 giorni.
Tokenized	*Un//paziente//maschile//,//di//60//anni//,//è//ammesso//al//reparto//di//Geriatria//in//data//11//settembre//2013//per//trattamento//preventivo//di//varici//esofagee//.//*
	È//affetto//da//gastropatia//da//ipertensione//portale//severa//e//da//splenomegalia//con//ipersplenismo//.//
	In//data//13//settembre//2013//è//sottoposto//a//procedura//endoscopica//di//routine//per//legatura//di//varici//esofagee//.//
	In//data//14//settembre//2013//permane//a//digiuno//per//12//ore//,//poi//passa//a//dieta//idrica//fredda//.//
	Addome//trattabile//non//dolorabile//.//
	In//data//16//settembre//2013//è//dimesso//.//
	Come//diagnosi//di//uscita//,//è//riportata//la//presenza//di//varici//esofagee//senza//menzione//di//sanguinamento//.//
	È//richiesto//controllo//endoscopico//ambulatoriale//dopo//30//giorni//.//
Normalized	*Un//paziente//maschile//di//60//anni//è//ammesso//al//reparto//di//Geriatria//in//data//11/09/2013//per//trattamento//preventivo//di//varici//esofagee//*
	è//affetto//da//gastropatia//da//ipertensione//portale//severa//e//da//splenomegalia//con//ipersplenismo//
	in//data//13/09/2013//è//sottoposto//a//procedura//endoscopica//di//routine//per//legatura//di//varici//esofagee//
	in//data//14/09/2013//permane//a//digiuno//per//12//ore//poi//passa//a//dieta//idrica//fredda//
	addome//trattabile//non//dolorabile//
	in//data//16/09/2013//è//dimesso//
	come//diagnosi//di//uscita//è//riportata//la//presenza//di//varici//esofagee//senza//menzione//di//sanguinamento//
	è//richiesto//controllo//endoscopico//ambulatoriale//dopo//30//giorni//

Then, *POS Tagging* is run, by identifying associations between terms and categories, as reported in Table 3.

Table 3

Associations between Terms and Categories

Term	Category	Term	Category	Term	Category	Term	Category
Un	Article	*Paziente*	Noun	*Maschile*	Adjective	*Di*	Preposition
60	Number	*Anni*	Noun	*E'*	Verb	*Ammesso*	Verb
Al	Preposition	*Reparto*	Noun	*Di*	Preposition	*Geriatria*	Noun
In	Preposition	*Data*	Noun	*11/09/2013*	Date	*Per*	Preposition
Trattamento	Noun	*Preventivo*	Adjective	*Di*	Preposition	*Varici*	Noun
Esofagee	Adjective	*È*	Verb	*Affetto*	Adjective	*Da*	Preposition
Gastropatia	Noun	*Da*	Preposition	*Ipertensione*	Noun	*Portale*	Adjective
Severa	Adjective	*E*	Conjunction	*Da*	Preposition	*Splenomegalia*	Noun
Con	Preposition	*Ipersplenismo*	Noun	*In*	Preposition	*Data*	Noun
13/09/2013	Date	*È*	Verb	*Sottoposto*	Verb	*A*	Preposition

Procedura	Noun	*Endoscopica*	Adjective	*Di*	Preposition	*Routine*	Noun
Per	Preposition	*Legatura*	Noun	*Di*	Preposition	*Varici*	Noun
Esofagee	Adjective	*In*	Preposition	*Data*	Noun	*14/09/2013*	Date
Permane	Verb	*A*	Preposition	*Digiuno*	Noun	*Per*	Preposition
12	Number	*Ore*	Noun	*Poi*	Adverb	*Passa*	Verb
A	Preposition	*Dieta*	Noun	*Idrica*	Adjective	*Fredda*	Adjective
Addome	Noun	*Trattabile*	Adjective	*Non*	Adverb	*Dolorabile*	Adjective
In	Preposition	*Data*	Noun	*16/09/2013*	Date	*È*	Verb
Dimesso	Verb	*Come*	Conjunction	*Diagnosi*	Noun	*Di*	Preposition
Uscita	Noun	*È*	Verb	*Riportata*	Verb	*La*	Article
Presenza	Noun	*Di*	Preposition	*Varici*	Noun	*Esofagee*	Adjective
Senza	Preposition	*Menzione*	Noun	*Di*	Preposition	*Sangunamento*	Noun
È	Verb	*Richiesto*	Verb	*Controllo*	Noun	*Endoscopico*	Adjective
Ambulatoriale	Adjective	*Dopo*	Preposition	*30*	Number	*Giorni*	Noun

At this point, the last stage of Text Processing, i.e. *Lemmatization*, is executed, by identifying the associations between lemmas and categories, as reported in Table 4. It is worth noting that many terms are already in canonical form, and for this reason, in this stage, they are not further converted, whereas other terms, such as the noun "*varici*" (in English *varices*, i.e. plural form of the noun "*varix*") or the verb "*permane*" (in English "*remains*", third-person singular of the present tense of the verb "*to remain*") are respectively transformed in "*varice*" (in English, *varix*) and "*permanere*" (in English, "*to remain*").

Table 4

Associations between Lemmas and Categories

Term	Category	Term	Category	Term	Category	Term	Category
Un	Article	*Paziente*	Noun	*Maschile*	Adjective	*Di*	Preposition
60	Number	*Anno*	Noun	*Essere*	Verb	*Ammettere*	Verb
A	Preposition	*Reparto*	Noun	*Di*	Preposition	*Geriatria*	Noun
In	Preposition	*Data*	Noun	*11/09/2013*	Date	*Per*	Preposition
Trattamento	Noun	*Preventivo*	Adjective	*Di*	Preposition	*Varice*	Noun
Esofageo	Adjective	*Essere*	Verb	*Affetto*	Adjective	*Da*	Preposition
Gastropatia	Noun	*Da*	Preposition	*Ipertensione*	Noun	*Portale*	Adjective
Severo	Adjective	*E*	Conjunction	*Da*	Preposition	*Splenomegalia*	Noun
Con	Preposition	*Ipersplenismo*	Noun	*In*	Preposition	*Data*	Noun
13/09/2013	Date	*Essere*	Verb	*Sottoporre*	Verb	*A*	Preposition
Procedura	Noun	*Endoscopico*	Adjective	*Di*	Preposition	*Routine*	Noun
Per	Preposition	*Legatura*	Noun	*Di*	Preposition	*Varice*	Noun
Esofageo	Adjective	*In*	Preposition	*Data*	Noun	*14/09/2013*	Date
Permanere	Verb	*A*	Preposition	*Digiuno*	Noun	*Per*	Preposition
12	Number	*Ora*	Noun	*Poi*	Adverb	*Passare*	Verb
A	Preposition	*Dieta*	Noun	*Idrico*	Adjective	*Freddo*	Adjective
Addome	Noun	*Trattabile*	Adjective	*Non*	Adverb	*Dolorabile*	Adjective
In	Preposition	*Data*	Noun	*16/09/2013*	Date	*Essere*	Verb
Dimettere	Verb	*Come*	Conjunction	*Diagnosi*	Noun	*Di*	Preposition
Uscita	Noun	*Essere*	Verb	*Riportare*	Verb	*La*	Article
Presenza	Noun	*Di*	Preposition	*Varice*	Noun	*Esofageo*	Adjective
Senza	Preposition	*Menzione*	Noun	*Di*	Preposition	*Sangunamento*	Noun
Essere	Verb	*Richiedere*	Verb	*Controllo*	Noun	*Endoscopico*	Adjective
Ambulatoriale	Adjective	*Dopo*	Preposition	*30*	Number	*Giorno*	Noun

Successively, the first stage of the Identification of Concepts, i.e. *Recognition of Relevant Terms*, is executed, by producing the values for the TF-IDF indexes as reported in Table 5. In particular, a lemma in Table 5 is

considered as relevant, and marked in bold and italic font, if its value for the TF-IDF index is greater than a given threshold. In this case, the threshold value is empirically set to 3, with the supervision of a domain expert.

Table 5

Lemmas and their values for TF-IDF index

Lemma	TF-IDF	Lemma	TF-IDF	Lemma	TF-IDF	Lemma	TF-IDF
Un	0,2	***Paziente***	5,6	***Maschile***	3,0	*Di*	0,3
60	0,8	*Anno*	2,2	*Essere*	1,8	***Ammettere***	4,0
A	0,3	***Reparto***	3,9	*Di*	0,3	***Geriatria***	4,1
In	0,3	*Data*	2,2	***11/09/2013***	6,5	*Per*	0,3
Trattamento	5,8	***Preventivo***	3,1	*Di*	0,3	***Varice***	5,1
Esofageo	6,2	*Essere*	1,8	***Affetto***	3,2	*Da*	0,4
Gastropatia	6,9	*Da*	0,4	***Ipertensione***	4,0	***Portale***	3,1
Severo	3,1	*E*	0,1	*Da*	0,4	***Splenomegalia***	6,1
Con	0,3	***Ipersplenismo***	5,8	*In*	0,3	*Data*	2,2
13/09/2013	6,8	*Essere*	1,8	***Sottoporre***	3,2	*A*	0,3
Procedura	3,1	***Endoscopico***	5,4	*Di*	0,3	***Routine***	3,2
Per	0,3	***Legatura***	4,3	*Di*	0,3	***Varice***	5,1
Esofageo	6,2	*In*	0,3	*Data*	2,2	***14/09/2013***	6,1
Permanere	2,8	*A*	0,3	***Digiuno***	3,5	*Per*	0,3
12	0,4	*Ora*	2,2	*Poi*	0,5	***Passare***	3,1
A	0,3	***Dieta***	3,4	***Idrico***	3,2	***Freddo***	3,3
Addome	4,1	***Trattabile***	3,1	*Non*	0,3	***Dolorabile***	4,0
In	0,3	*Data*	2,2	***16/09/2013***	6,2	*Essere*	1,8
Dimettere	3,2	*Come*	1,5	***Diagnosi***	3,9	*Di*	0,3
Uscita	3,1	*Essere*	1,8	***Riportare***	3,3	*La*	0,2
Presenza	3,1	*Di*	0,3	***Varice***	5,1	***Esofageo***	6,2
Senza	1,9	***Menzione***	4,1	*Di*	0,3	***Sanguinamento***	5,0
Essere	1,8	***Richiedere***	3,8	***Controllo***	4,0	***Endoscopico***	5,4
Ambulatoriale	4,5	*Dopo*	2,3	*30*	0,6	*Giorno*	2,2

The second stage, i.e. *Definition of Synset*, is then performed, by associating the synset, i.e. the proper concept, to each selected lemma, as showed in Table 6. It is worth noting that numbers and dates are obviously not considered in this step, since no concept can be associated to them. Thus, they are not taken into account for the identification of the sections.

Table 6

Synsets associated to Relevant Lemmas

Lemma	TF-IDF	Synset	Concept
Gastropatia	6,9	Gastropatia	*Gastropatia*
Esofageo	6,2	Esofageo	*Esofageo*
Splenomegalia	6,1	Splenomegalia	*Splenomegalia*
Ipersplenismo	5,8	Ipersplenismo	*Ipersplenismo*
Trattamento	5,8	Trattamento, Accoglienza, Cura, Manipolazione, Processo, Terapia	*Trattamento*
Paziente	5,6	Paziente, Ammalato, Degente, Malato	*Paziente*
Endoscopico	5,4	Endoscopico	*Endoscopico*
Varice	5,1	Varice	*Varice*

Sanguinamento	5	Sanguinamento, Emorragia, Perdita Di Sangue	*Emorragia*
Ambulatoriale	4,5	Ambulatoriale, Ambulatorio, clinica, consultorio, gabinetto, infermeria, sanatorio	*Ambulatorio*
Legatura	4,3	Legatura	*Legatura*
Geriatria	4,1	Geriatria, Gerontoiatria, Gerontologia	*Geriatria*
Menzione	4,1	Menzione, Accenno, Cenno, Citazione, Riferimento	*Riferimento*
Addome	4,1	Addome, Pancia, Ventre	*Addome*
Ammettere	4	Ammettere, Autorizzare, Accettare, Accogliere, Associare, Iscrivere	*Ammettere*
Controllo	4	Controllo, Condizionamento, Direzione, Dominio, Governo, Ispezione, Padronanza, Revisione, Vigilanza	*Controllo*
Dolorabile	4	Dolorabile	*Dolorabile*
Ipertensione	4	Ipertensione	*Ipertensione*
Diagnosi	3,9	Diagnosi, Parere, Prognosi, Responso, Valutazione	*Diagnosi*
Reparto	3,9	Reparto, Compartimento, Dipartimento, Sezione, Suddivisione	*Reparto*
Richiedere	3,8	Richiedere, Esigere, Imporre, Pretendere, Sollecitare, Volere	*Richiedere*
Digiuno	3,5	Digiuno, Astinenza, Fame, Inedia	*Digiuno*
Dieta	3,4	Dieta, Astinenza, Nutrizione, Regime Dietetico	*Dieta*
Freddo	3,3	Freddo, Esanime, Gelo, Impassibile, Inerte	*Freddo*
Riportare	3,3	Riportare, Citare, Restituire, Ricondurre, Ridare, Riferire, Ripetere	*Riportare*
Affetto	3,2	Affetto	*Affetto*
Dimettere	3,2	Dimettere, congedare, esonerare, licenziare, rinunciare	*Dimettere*
Idrico	3,2	Idrico	*Idrico*
Routine	3,2	Routine, Abitudine	*Routine*
Sottoporre	3,2	Sottoporre, Assoggettare	*Sottoporre*
Passare	3,1	Passare, Circolare, Estendersi, Percorrere, Snodarsi, Toccare, Transitare	*Passare*
Portale	3,1	Portale, Porta	*Portale*
Presenza	3,1	Presenza, Apparenza, Cospetto, Esistenza, Figura, Parvenza, Prontezza, Vista	*Presenza*
Preventivo	3,1	Preventivo, Cautelativo, Difensivo, Precauzionale, Prudenziale, Valutazione	*Preventivo*
Procedura	3,1	Procedura, Causa, Metodo, Norma, Prassi, Processo, Sistema	*Procedura*
Severo	3,1	Severo, Rigido, Grave, Oneroso	*Severo*
Trattabile	3,1	Trattabile, Curabile, Usabile	*Trattabile*
Uscita	3,1	Uscita, Apertura, Battuta, Deflusso, Edizione, Esalazione, Fuga, Fuoriuscita, Passaggio, Porta, Sbocco, Sortita Varco, Via	*Uscita*
Maschile	3	Maschile, Da Uomo, Mascolino, Virile	*Maschile*

This table shows for each relevant lemma, extracted on the basis of its grammatical category and TF-IDF value, the synset associated, i.e. a set of terms referring the same concept. This list of terms is built by exploiting the relations codified in “Mesh”. In the example, a proper synset is associated to each extracted term. For instance: “Gastropatia” (Gastropathy), “Esofageo” (Oesophageal), “Splenomegalia” (splenomegaly), “Ipersplenismo” (Hypersplenism), “Trattamento” (Treatment), “Paziente” (Patient), “Endoscopico” (Endoscopic), “Varice” (Varix), “Emorragia” (Hemorrhage).

Finally, the last phase, i.e. *Identification of Sections*, is carried out, which classifies as a section each of the blocks of the fragment of discharge summary considered, depending on the presence of the identified concepts in

them. Table 7 reports, for each block, the set of concepts recognized as significant for the classification as well as all the concepts present in it.

Table 7

Classification of Text Blocks in Sections

Block		Concepts	Section
1	Significant	Paziente, Maschile, Ammettere, Reparto, Geriatria, Trattamento, Preventivo	Hospital admission diagnosis
	All	Paziente, Maschile, Ammettere, Reparto, Geriatria, Trattamento, Preventivo, Varice, Esofageo	
2	Significant	Affetto, Gastropatia, Ipertensione, Splenomegalia, Ipersplenismo	Medical history
	All	Affetto, Gastropatia, Ipertensione, Portale, Severo, Splenomegalia, Ipersplenismo	
3	Significant	Sottoporre, Procedura, Endoscopico, Routine, Legatura	Hospital course
	All	Sottoporre, Procedura, Endoscopico, Routine, Legatura, Varice, Esofageo	
4	Significant	Digiuno, Dieta, Idrico, Freddo	Hospital course
	All	Digiuno, Passare, Dieta, Idrico, Freddo	
5	Significant	Addome, Trattabile, Dolorabile	Hospital course
	All	Addome, Trattabile, Dolorabile	
6	Significant	Dimettere	Hospital course
	All	Dimettere	
7	Significant	Diagnosi, Uscita	Hospital discharge diagnosis
	All	Diagnosi, Uscita, Riportare, Presenza, Varice, Esofageo, Emorragia	
8	Significant	Richiedere, Controllo, Ambulatorio	Treatment plan
	All	Richiedere, Controllo, Endoscopico, Ambulatorio	

The concepts and sections identified in Table 7 are modelled in the domain ontology as individuals of the concepts **Concept** and **Section.** Moreover, they are opportunely linked coherently with the results shown in Table 7 by means of assertions of the role **isContainedIn.** Some examples are reported in the following:

- **isContainedIn**(*Paziente, HospitalAdmissionDiagnosis),* **isContainedIn**(*Ammettere, HospitalAdmissionDiagnosis);*
- **isContainedIn**(*Gastropatia, MedicalHistory),* **isContainedIn**(*Splenomegalia, MedicalHistory),*
- **isContainedIn**(*Controllo,TreatmentPlan);*
- **isContainedIn**(*Legatura, HospitalCourse);*
- **isContainedIn**(*Uscita, HospitalDischargeDiagnosis).*

It is important to note that some concepts, such as Varice (Varix), Endoscopico (Endoscopic), Esofageo (Oesophageal), are finally associated to different sections, since they appear simultaneously in them. In the example, this issue has been not relevant for the final classification, since the combination of other concepts present in the different blocks has been used to discriminate among the different possible sections. Nevertheless, in general, the presence of concepts in more than one text block can generate ambiguities and, thus, possible misclassifications.

Summarizing, coherently with the results shown in Table 7, the initial fragment of discharge summary is structured with respect to its sections as highlighted in Fig. 11, where all the text blocks belonging to the sections *Hospital admission diagnosis*, *Medical history, Hospital course, Hospital discharge diagnosis* and *Treatment plan* are marked by proper tags of sections. This information is codified in RDF.

```
<rdf:Description about=" Un paziente maschile, di 60 anni, è ammesso al reparto di Geriatria in data 11
                         settembre 2013 per trattamento preventivo di varici esofagee.">
      <prop:Section> Hospital admission diagnosis </prop:Section>
</rdf:Description>
<rdf:Description about=" È affetto da gastropatia da ipertensione portale severa e da splenomegalia con
                         ipersplenismo.">
      <prop:Section> Medical history </prop:Section>
</rdf:Description>
<rdf:Description about=" In data 13 settembre 2013 è sottoposto a procedura endoscopica di routine per
                         legatura di varici esofagee. In data 14 settembre 2013 permane a digiuno per 12
                         ore, poi passa a dieta idrica fredda. Addome trattabile non dolorabile. In data
                         16 settembre 2013 è dimesso.”>
      <prop:Section> Hospital course </prop:Section>
</rdf:Description>
<rdf:Description about=" Come diagnosi di uscita, è riportata la presenza di varici esofagee senza
                         menzione di sanguinamento.”>
      <prop:Section> Hospital discharge diagnosis </prop:Section>
</rdf:Description>
<rdf:Description about=" È richiesto controllo endoscopico ambulatoriale dopo 30 giorni.”>
      <prop:Section> Treatment plan </prop:Section>
</rdf:Description>
```

Fig. 11. The fragment of discharge summary partitioned in terms of the sections identified.

Finally, with respect to the sections identified, the prototype system is then used to specify, by exploiting the patterns defined, a set of policies for the following roles, *Doctors, Administrative Managers and Nurses.* Such policies are schematically reported in Table 8.

Table 8

Policies defined for the sections identified

Role	Section	Operation	Access Right
Doctor	*Hospital admission diagnosis*	*Read/Write*	*Allowed/Allowed*
Nurse	*Hospital admission diagnosis*	*Read/Write*	*Allowed/NotAllowed*
Administrative Manager	*Hospital admission diagnosis*	*Read/Write*	*NotAllowed/NotAllowed*
Doctor	*Medical history*	*Read/Write*	*Allowed/Allowed*
Nurse	*Medical history*	*Read/Write*	*Allowed/NotAllowed*
Administrative Manager	*Medical history*	*Read/Write*	*NotAllowed/NotAllowed*
Doctor	*Hospital course*	*Read/Write*	*Allowed/Allowed*
Nurse	*Hospital course*	*Read/Write*	*Allowed/Allowed*
Administrative Manager	*Hospital course*	*Read/Write*	*NotAllowed/NotAllowed*
Doctor	*Hospital discharge diagnosis*	*Read/Write*	*Allowed/Allowed*
Nurse	*Hospital discharge diagnosis*	*Read/Write*	*Allowed/ NotAllowed*
Administrative Manager	*Hospital discharge diagnosis*	*Read/Write*	*NotAllowed/NotAllowed*
Doctor	*Treatment plan*	*Read/Write*	*Allowed/Allowed*
Nurse	*Treatment plan*	*Read/Write*	*Allowed/ NotAllowed*
Administrative Manager	*Treatment plan*	*Read/Write*	*NotAllowed/NotAllowed*

In particular, the procedure for building the part of the first policy shown in Table 8 pertaining the operation *Read* is deeply described in Fig. 12, by highlighting, in a stepwise manner, the different kinds of relational patterns exploited.

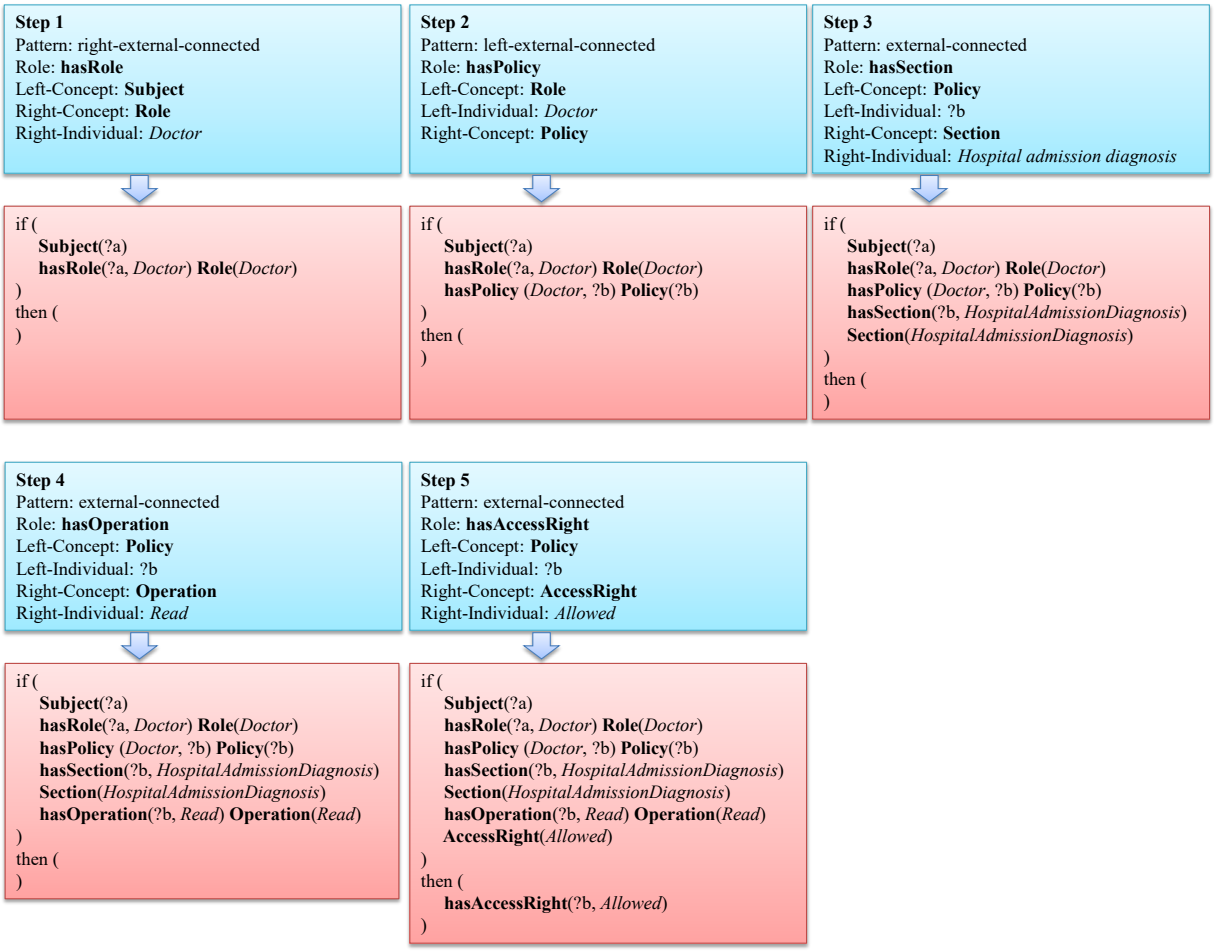

Fig. 12. An example of application of the relational patterns to build an access control policy.

6. Conclusions

With the recent advances in healthcare technologies and systems, healthcare services are becoming more and more efficiently and ubiquitously accessible and, contextually, the amount of clinical data electronically available is significantly increased, arising new expectations of fine-grained access control over complex objects such as EHRs, where various protection requirements should be met. Indeed, EHRs are composed of different portions characterized by different types of content, such as personal data, diagnosis reports, medical history and so on, which often involve highly sensitive information. As a result, access control systems are required to consider the protection object, i.e. healthcare data, as primarily data-centric rather than document-centric.

To address this challenge, this work has proposed a semantic-based framework offering an innovative and valuable way to enable and support the definition of fine-grained access control policies over semi-structured EHRs. The key features of the framework are: i) a hybrid approach that combines linguistic and statistical techniques to structure a narrative health record semantically, by first recognizing relevant concepts in blocks of

text and, successively, identifying the sections composing the document; ii) a customized RBAC model to regulate the access to portions of semi-structured EHRs; iii) a high-level ontology that expresses the elements of the proposed RBAC model, and, in addition, a domain specific ontology that captures the features of a specific application domain; iv) a procedural policy language to encode access control restrictions in the form of "*if–then rules*" built on the top of the ontological formalization of the proposed RBAC model; v) a set of patterns for supporting the simple insertion and editing of such access control restrictions with the aim of reducing the complexity of the formalization process.

Existing access control systems possess some of these features; it is however the combination of all these features within an integrated framework, together with a prototype implementation in the form of a system offering simple and intuitive interfaces to the security administrators, that makes the whole approach flexible, extensible and powerful enough to be well-suited to the healthcare domain.

The principal advantages of using this framework rely on the ability to support and guide security administrators in authoring and updating access control policies according to dynamically changing security needs or in customizing them to a specific healthcare organization.

Indeed, the proposed RBAC model allows statically and dynamically regulating users' actions through the establishment and definition of roles, role hierarchies, relationships, and constraints in a large-scale heterogeneous distributed environment, that is the healthcare domain, where users could need to share their data with previously unknown ones. Moreover, this RBAC model also allows for the specification and enforcement of a variety of fine-grained access control policies, also over portions of EHRs, opportunely arranged in a semi-structured form. This issue represents a major advancement in flexibility since it ensures that sensitive elements pertaining a patient and contained, for instance, in different sections of an EHR can be properly protected and made accessible only to authorized users, such as physicians or family members.

Furthermore, the method of representing access control policies in terms of a combination of more knowledge representation formalisms, i.e. if-then rules built on the top of ontological entities, has been chosen as the most suitable to the healthcare domain, since that hybridization is easily usable and understandable also by a non-technical staff. In addition, to reduce the complexity of the formalization process, a set of patterns has been defined for guiding and facilitating the editing of policies by means of a set of simpler and handier objects.

Finally, the facilities for directly working on unstructured health records convey another relevant advantage: by supporting the structuring of health records in terms of sections, users could be encouraged to define fine-grained access control policies on them, especially because directly generated starting from both their expertise and the peculiar needs associated to their healthcare organization.

As concluding remark, it is worth noting that the encouraging results given by the experimental evaluation suggest that the framework could be simply and proficiently utilized by security administrators to author and update fine-grained access control policies also over portions of semi-structured EHRs. Even if the evaluation has regarded discharge summaries expressed in Italian, the proposed framework has a general basis, and, thus, can be undoubtedly used to operate with other types of health records, also expressed in different languages. Such a way, the remarkable aim of securing healthcare information on electronic documents assuming a semi-structured form could be achieved, so as to improve the overall information security and privacy in large distributed and heterogeneous medical settings.

References

Amato, F., Casola, V., Mazzocca, N., Romano, S. A semantic approach for fine-grain access control of e-health documents. Logic Journal of IGPL, vol. 2, no. 4, pp. 692-701, 2013.

Amato, F., Casola, V., Mazzocca, N., Romano, S. A semantic-based document processing framework: a security perspective. In 2011 International Conference on Complex, Intelligent and Software Intensive Systems (CISIS), pp. 197-202, 2011.

Amato, F., Casola, V., Mazzeo, A., Romano, S. A semantic based methodology to classify and protect sensitive data in medical records. In International Conference on Information Assurance and Security (IAS 2010), pp. 240-246, 2010.

Argüello Casteleiro, M., Des, J., Prieto, M. J. F., Perez, R., Paniagua, H. Executing medical guidelines on the web: Towards next generation healthcare. Knowledge-Based Systems, vol. 22, no. 7, pp. 545-551, 2009.

Aussenac-Gilles, N., Despres, S., Szulman, S. The TERMINAE method and platform for ontology engineering from text. In Proceedings of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge, P. Buitelaar and P. Cimiano (Eds.). IOS Press, pp. 199-223, 2008.

Biber, D., Conrad, S., Reppen, R. Corpus linguistics: Investigating language structure and use. Cambridge University Press, 1998.

Butler, C. S. Structure and Function–A Guide to Three Major Structural-Functional Theories: Part 2: From clause to discourse and beyond. vol. 64. John Benjamins Publishing, 2003.

Buitelaar, P., Olejnik, D., Sintek, M. A Protégé plug-in for ontology extraction from text based on linguistic analysis. In Proceedings of the 1st European Semantic Web Symposium (ESWS), pp. 31–44, 2004.

Church, K., Hanks, P. Word association norms, mutual information, and lexicography. Computational Linguistics, vol. 16, no. 1, pp. 22-29, 1990.

Cimiano, P. Völker, J. Text2onto – A framework for ontology learning and data-driven change discovery. In Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB), E. Métais, A. Montoyo, and R. Munoz, (Eds.), vol. 3513 of Lecture Notes in Computer Science, pp. 227–238, 2005.

Cimiano, P. Ontology Learning and Population from Text: Algorithms, Evaluation and Applications, Springer, New York, NY, 2006.

Colantonio, S., Esposito, M., Martinelli, M., De Pietro, G., Salvetti, O. A Knowledge Editing Service for Multisource Data Management in Remote Health Monitoring. IEEE Transactions on Information Technology in Biomedicine, vol. 16, no. 6, pp. 1096–1104, 2012.

Crampton, J. Specifying and enforcing constraints in role-based access control. In 8th ACM Symposium on Access Control Models and Technologies, pp. 43–50, 2003.

Croitoru, M., Xiao, L., Dupplaw, D., Lewis, P. Expressive security policy rules using layered conceptual graphs, Knowledge-Based Systems vol. 21, no. 3, pp. 209–216, 2008.

Damiani, E., di Vimercati, S. D. C., Fugazza, C., Samarati, P. Extending policy languages to the semantic web. In Web Engineering, pp. 330-343, 2004.

De Mauro, T., Mancini, F., Vedovelli, M., Voghera, M. Lessico di frequenza dell'italiano parlato. Etas libri, Rome, IT, 1993.

Decherchi, S., Gastaldo, P., Redi, J., Zunino, R. A text clustering framework for information retrieval. Journal of information Assurance and Security, vol. 4, pp. 174-182, 2009.

Dolin, R. H., Alschuler, L., Boyer, S., Beebe, C., Behlen, F. M., Biron, P. V., Shvo, A. S. HL7 clinical document architecture, release 2. Journal of the American Medical Informatics Association, pp. 13, no.1, pp.30-39, 2006.

Dunning, T. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, vol. 19, no. 1, pp. 61-74, 1994.

Eyers, D. M., Bacon, J., Moody, K. OASIS role-based access control for electronic health records. IEE Proceedings-Software, vol. 153, no. 1, pp. 16-23, 2006.

Fernández-Alemán, J. L., Señor, I. C., Lozoya, P. Á. O., Toval, A. Security and privacy in electronic health records: A systematic literature review. Journal of biomedical informatics, vol. 46, no. 3, pp. 541-562, 2013.

Ferraiolo, D. F., Sandhu, R., Gavrila, S., Kuhn, D. R., Chandramouli, R. Proposed NIST standard for role-based access control. ACM Trans. Info. Syst. Security, vol. 4, pp. 224–274, 2001.

Finin, T., Joshi, A., Kagal, L., Niu, J., Sandhu, R., Winsborough, W., & Thuraisingham, B. ROWLBAC: representing role based access control in OWL. In 13th ACM symposium on Access control models and technologies, pp.73-82, 2008.

Fowler, S. A., Yaeger, L. H., Yu, F., Doerhoff, D., Schoening, P., & Kelly, B. Electronic health record: integrating evidence-based information at the point of clinical decision making. Journal of the Medical Library Association: JMLA, vol. 102, no. 1, pp. 52-55, 2014.

Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation, vol. 101, no. 23, pp.e215-e220, 2000.

Gomez, F. A representation of complex events and processes for the acquisition of knowledge from texts, Knowledge-Based Systems, vol. 10, no. 4, pp. 237-251, 1998.

Grilheres, B., Brunessaux, S., Leray, P. Combining classifiers for harmful document filtering. In RIAO, pp. 173-185, 2004.

Häyrinen, K., Saranto, K., Nykänen, P., Definition, structure, content, use and impacts of electronic health records: A review of the research literature, International Journal of Medical Informatics, vol. 77, no. 5, pp. 291-304, 2008.

Herre, H., Heller, B. Semantic foundations of medical information systems based on top-level ontologies. Knowledge-Based Systems, vol. 19, no. 2, pp. 107-115, 2006.

Horrocks, I., Sattler, U., Tobies, S. Practical Reasoning for Expressive Description Logics. In Proceedings of the 6th International Conference on Logic Programming and Automated Reasoning, pp. 161-180, 1999.

Kagal, L., Berners-Lee, T., Connolly, D., Weitzner, D. Using semantic web technologies for policy management on the web. In The National Conference on Artificial Intelligence,vol. 21, no. 2, pp. 1337-1344, 2006.

Kageura, K., Umino, B. Methods of automatic term recognition: A review. Terminology, vol. 3, no. 2, pp. 259–289, 1996.

Kennedy, G. D. An introduction to corpus linguistics. Longman, 1998.

Johnson, S. B., Bakken, S., Dine, D., Hyun, S., Mendonça, E., Morrison, F., Bright, T., Vleck, T.V., Wrenn, J., Stetson, P., An Electronic Health Record Based on Structured Narrative, Journal of the American Medical Informatics Association, vol. 15, no. 1, pp. 54-64, 2008.

Le, X. H., Doll, T., Barbosu, M., Luque, A., Wang, D. An enhancement of the Role-Based Access Control model to facilitate information access management in context of team collaboration and workflow. Journal of Biomedical Informatics, vol. 45, no. 6, pp. 1084-1107, 2012.

Maedche, A. , Volz, R. The TEXT-TO-ONTO ontology extraction and maintenance system, ICDM-Workshop on Integrating Data Mining and Knowledge Management, pp. 11, 2001.

Mao, S., Rosenfeld, A., Kanungo, T. Document structure analysis algorithms: a literature survey. In Electronic Imaging 2003, (International Society for Optics and Photonics, pp. 197-207, 2003.

Martino, L. D., Ni, Q., Lin, D., Bertino, E. Multi-domain and privacy-aware role based access control in ehealth. In Second International Conference on Pervasive Computing Technologies for Healthcare, pp. 131-134, 2008.

Minutolo, A., Esposito, M., De Pietro, G. A pattern-based knowledge editing system for building clinical Decision Support Systems. Knowledge-Based Systems, vol. 35, pp. 120-131, 2012.

Mount, C.D., Kelman, C.W., Smith, L.R., Douglas R.M. An integrated electronic health record and information system for Australia? Medical Journal of Australia, vol. 172, pp. 25-27, 2000.

Phansalkar, S., Desai, A., Choksi, A., Yoshida, E., Doole, J., Czochanski, M., Tucker, A.D., Middleton, B., Bell, D., Bates, D.W. Criteria for assessing high-priority drug-drug interactions for clinical decision support in electronic health records. BMC medical informatics and decision making, vol. 13, no. 65, pp. 1-11, 2013.

Priebe, T., Dobmeier, W., Schlager, C., Kamprath, N. Supporting attribute-based access control in authorization and authentication infrastructures with ontologies. Journal of Software, vol. 2, no. 1, pp. 27-38, 2007.

Rodrigues, J.J., Pedro, L.M., Vardasca, T., de la Torre-Díez, I., Martins, H.M. Mobile health platform for pressure ulcer monitoring with electronic health record integration. Journal of Health Informatics, vol. 19, no. 4, pp. 300-311, 2013.

Rosero, L., Aranda, J., Riguidel, M., Gidoin, D. A Fine-Grained Document-based Access Control Model. International Journal of Machine Learning and Computing, vol. 1, no. 3, pp. 317-324, 2011.

Salton, G., Buckley, C. Term-weighting approaches in automatic text retrieval. Information Processing & Management, vol. 24, no. 5, pp. 513-523, 1988.

Sari, A. K., Rahayu, W., Bhatt, M. Archetype sub-ontology: Improving constraint-based clinical knowledge model in electronic health records. Knowledge-Based Systems, vol. 26, pp. 75-85, 2012.

Sandhu, R. S., Samarati, P. Access control: principles and practice, IEEE Commun. Mag., vol. 32, pp. 40–48, 1994.

Sandhu, R. S., Coyne, E. J., Youman, C. E. Role-based access control models. IEEE Comput, vol. 29, pp. 38–47, 1996.

Steele, R., Min, K. HealthPass: Fine-grained access control to portable personal health records. In 24th IEEE International Conference on Advanced Information Networking and Applications, pp. 1012-1019, 2010.

Uszok, A., Bradshaw, J. M., Johnson, M., Jeffers, R., Tate, A., Dalton, J., Aitken, S. KAoS policy management for semantic web services. IEEE Intelligent Systems, vol. 19, no. 4, pp. 32– 41, 2004.

Varshney, U. Mobile health: Four emerging themes of research. Decision Support Systems, vol. 66, pp. 20-35, 2014.

Vrusias, B. L., Golledge, I. Online Self-Organised Map Classifiers as Text Filters for Spam Email Detection. Journal of Information Assurance and Security, vol. 4, no. 2, pp.151-160, 2009.

Velardi, P., Navigli, R., Cuchiarelli, A., Neri. F. Evaluation of OntoLearn, a methodology for automatic population of domain ontologies. In Ontology Learning from Text: Methods, Applications and Evaluation, P. Buitelaar, P. Cimiano, B. Magnini, (Eds.), IOS Press, pp. 92–106, 2005.

Wong, W., Liu, W., Bennamoun, M. Determination of unithood and termhood for term recognition. Handbook of research on text and web mining technologies, M. Song and Y. Wu (Eds.), IGI Global, 2008.

Wu, Y.C. Integrating statistical and lexical information for recognizing textual entailments in text, Knowledge-Based Systems, vol. 40, pp. 27-35, 2013.