A Semantic Web based Framework for the Interoperability and Exploitation of Clinical Models and EHR Data

María del Carmen Legaz-García[1], Catalina Martínez-Costa[2], Marcos Menárguez-Tortosa[1], Jesualdo Tomás Fernández-Breis[1]

[1] *Departamento de Informática y Sistemas, Universidad de Murcia, IMIB-Arrixaca, Spain*

[2] *Institute of Medical Informatics, Statistics, and Documentation, Medical University of Graz, Austria*

Abstract

The advent of electronic healthcare records (EHR) systems has triggered the need for their semantic interoperability, which is reinforced by the opportunities for secondary use of EHR data. The joint use of EHR standards and semantic resources has been identified as key for semantic interoperability. To date, existing tools focused on EHR standards permit to create, search, explore clinical models and to map data sources to clinical models but do not provide an appropriate support and integration of semantic resources or permit the secondary use of EHR data.

We describe a framework that leverages EHR and semantic web technologies for the exploitation of archetypes, data and ontologies and that is powered by OWL-based technologies. The framework has been implemented in the Archetype Management System (ArchMS), which enables to transform clinical models between standards, to check the correctness of clinical models or to find similar archetypes by applying semantic similarity measurements. ArchMS provides services for exploiting EHR data using OWL technologies. In particular, we focus here on the classification of patients as example of secondary use of clinical data. We also describe how ArchMS was used in a real use case, more concretely, in a study related to identifying patient

*Corresponding author:Jesualdo Tomás Fernández Breis, phone:+34868884613, fax:+34868884151, email: jfernand@um.es

cohorts in the colorectal cancer domain.

Keywords: Semantic web; medical informatics; electronic health records; ontology, semantic interoperability

1. Introduction

The increasing use of electronic health records (EHRs) in our globalized world leads to a situation where patient's health data is spread across different health systems. This situation demands semantic interoperability of clinical information, that is, their meaningful communication across EHR systems. Such interoperability has not been reached yet, what has been internationally considered as a reason for inefficiencies within the healthcare system, contributing to the waste of billions of dollars in the United States annually (see, for instance, Saleem et al. (2013)).

In Kalra et al. (2009), the SemanticHEALTH project identified EHR standards, ontologies and terminologies as key players to achieve semantic interoperability. In the last decades, many efforts have addressed the development of EHR standards and specifications, including openEHR[1], ISO 13606[2]. Such standards and specifications are based on the dual model architecture, which distinguishes two modelling levels. On the one hand, the information model provides the generic building blocks to structure the EHR information (i.e., data types and data structures). On the other hand, clinical models are used to specify clinical recording scenarios by constraining the information model structures (i.e., what needs to be recorded about the measurement of blood pressure). In both openEHR and ISO 13606, clinical models are named archetypes and according to Tapuria et al. (2013), they are a promising way of sharing clinical data in a formal and scalable way. Their interest is reinforced by the commitment of the Clinical Information Modeling Initiative (CIMI) to use archetypes[3]. HL7 specifications[4], despite not following the dual-model approach, have also evolved to include arfefacts similar to clinical models with the aim of facilitating sharing and interoperability. An example is the recent Fast Healthcare Interoperability Resources (FHIR) specification[5].

[1] http://www.openehr.org/

[2] http://www.en13606.org/

[3] http://informatics.mayo.edu/CIMI/index.php/London_2011

[4] http://www.hl7.org/

[5] http://hl7.org/fhir

The lack of appropriate tooling for applying and exploiting archetypes and archetype-based data in semantic interoperability environments is considered a barrier to the adoption of dual-model architectures by the majority of vendors. LinkEHR[6] and the tools developed by the openEHR community, like the Archetype Editor (AE)[7], ADL Workbench (AW)[8] and the Clinical Knowledge Manager (CKM)[9], are likely to be the most widely used archetype-based tools. LinkEHR permits the edition of archetypes, the normalization of legacy data using archetypes and view of EHR extracts. The AE permits the edition of archetypes, AW permits to create archetypes and templates for ISO 13606, openEHR and CIMI and to perform management tasks related to archetypes and ADL technologies, and CKM provides a repository for managing sets of archetypes. Despite the different objectives of these tools, all of them represent archetypes using the Archetype Definition Language (ADL)[10], which is a generic, formal language for representing constraint-based models, including archetypes.

The advent of EHR systems has also created new opportunities for the secondary use of data like rapid cohort identification, quality of care assessment, comparative effectiveness research, data privacy and de-/re-identification research, phenotyping methodology and predictive modelling, as discussed in Danciu et al. (2014). Some secondary uses require combining data from different systems, what requires semantic interoperability between such systems, and in works like Rea et al. (2012); Abhyankar et al. (2012) the corresponding solutions are based on standards.

As described in Martínez-Costa et al. (2009), ADL has a syntactic orientation that makes the realization of the semantic activities required in semantic interoperability environments more difficult, like (1) checking the correct semantic definition of the archetypes, (2) detecting semantically equivalent archetypes and EHR data; (3) joint exploitation of terminologies and ontologies; (4) joint exploitation of heterogeneous EHR data. The Semantic Web is described in Berners-Lee et al. (2001) as a new form of Web content meaningful to computers, and Goble and Stevens (2008) proposed the Semantic Web as a natural space for the integration and exploitation of biomedical

[6] `http://www.linkehr.com/`

[7] `http://www.openehr.org/downloads/archetypeeditor/home`

[8] `http://www.openehr.org/downloads/ADLworkbench/home`

[9] `http://www.openehr.org/ckm`

[10] `http://www.openehr.org/releases/1.0.2/architecture/am/adl.pdf`

data . There are different basic technologies for the success of the Semantic Web, among which the cornerstone technology is the ontology. An ontology represents a common, shareable and reusable view of a particular application domain (see Gruber (1993)). Semantic web technologies are meant to enable the joint exploitation of heterogeneous, distributed content because machine are able to understand the meaning of each content, which is provided in a precise way by the association between the data and the entities of ontologies.

Consequently, our hypothesis is that Semantic Web technologies provide a better support than ADL ones for performing the previously described tasks in the area of semantic interoperability. Consequently, we have developed a semantic web-based framework for the joint exploitation of archetypes, ontologies and terminologies for semantic interoperability. This framework has been implemented in the Archetype Management System (ArchMS), whose technological infrastructure of ArchMS permits to manage archetypes and EHR data from different standards using ADL/OWL technologies, for the kind of task for which each one is more appropriate. As a prototypical tool, we think that ArchMS represents a good example of how semantic web technologies can contribute to semantic interoperability environments.

2. Background

2.1. Archetypes

Archetypes are used to specify clinical recording scenarios and are meant to describe a semantic superset for the definition of medical data elements. An archetype may be used to record clinical data about a laboratory test, a blood pressure measurement, a medication order, etc. It can be defined as a specialization of another one, can include other archetypes through the slots mechanism, and can be used in combination with others by means of templates. They constitute a standardized way of capturing clinical data according to the archetype model AOM[11], and are expressed in ADL. In this work, we work with ADL/AOM1.4, where ADL archetypes include four main sections: header, description, definition and ontology. Header and description give general information about the archetype, such as name, language, author or purpose. The definition section contains the structures and constraints associated with the clinical recording scenario defined by the archetype. The

[11] `http://www.openehr.org/releases/trunk/architecture/am/aom1.5.pdf`

ontology section provides textual descriptions for each element from the definition section and bindings to other terminologies. It should be noted that the ontology section is called terminology in AOM2. Figure 1 shows a visual representation of the definition section of the openEHR blood pressure archetype. It records specific data related to the blood pressure measurement, such as systolic and diastolic blood pressure values; the protocol followed, method used, device, state of the patient in the moment of the recording, etc.

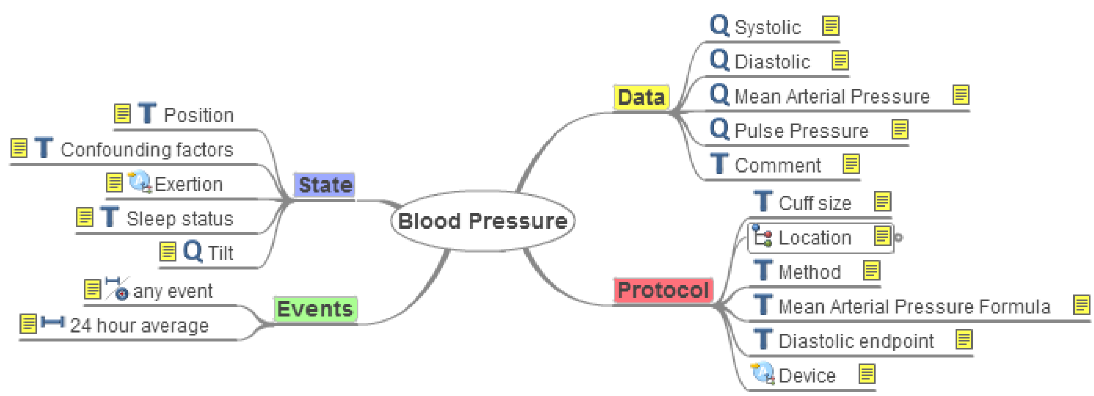

Figure 1: **Blood Pressure OpenEHR archetype mind map from CKM**

2.2. Ontologies

The study performed in Bodenreider (2008) shows the multiple applications and importance of ontologies in biomedical research. Biomedical ontologies are frequently used (1) as source vocabularies to annotate biological datasets, what improves document or data retrieval and query; (2) to facilitate information exchange in semantic interoperability and data integration scenarios; and (3) for decision support and reasoning, biomedical ontologies formally represent the biomedical domain by providing a set of axioms that define their concepts and how they relate to each other. Their formalization by using a Description Logics-based (DL) language allows performing reasoning and inferring additional information from the formalized one. Ontologies are not only fundamental for interoperability in biomedical domains, but in most application areas, as described in Maree and Belkhatir (2015).

It should be noted that the ontologies available do not usually meet all the requirements of a given task. On the one hand, existing ontologies might not include all the concepts needed. On the other hand, many currently available biomedical ontologies have been designed for annotation purposes and therefore are not suitable for automated reasoning. Best practices in ontology engineering recommend to reuse existing ontologies and to create modular ontologies (see Rector et al. (2012)). Finding the appropriate concepts for reuse is not an easy task, but repositories like Bioportal and tools like Watson (see Noy et al. (2009); d'Aquin and Motta (2011)) might help to find relevant concepts and ontologies on the web. In case of different candidate ontologies for re-use, decisions have to be made based on the quality of the ontologies and their appropriateness for the task.

The Web Ontology Language (OWL)[12] is the *de facto* standard for the implementation of ontologies and enables the precise formalization of data meaning in a way that can be automatically exploited. In recent years, our research group has developed OWL-based methods for the archetypes representation, the interoperability of archetypes and of clinical data, and for checking their semantic consistency (see Martínez-Costa et al. (2009, 2010a); Costa et al. (2011); Menárguez-Tortosa and Fernández-Breis (2013)). In line with this, a recent approach presented in Martínez-Costa and Schulz (2014) uses OWL for the detection of isosemantic content in heterogeneous EHR systems, that is content with the same meaning but structurally different. Besides, hundreds of biomedical ontologies are available in OWL format in repositories like Bioportal and more and more medical terminologies are becoming available in OWL. Besides, Tao et al. (2013) uses semantic web technologies to support secondary uses of EHR data. In this context, leveraging archetypes and ontologies has been demonstrated useful for patient cohort identification in Fernández-Breis et al. (2013).

In this work, we use ontologies for representing EHR information models, clinical models, clinical data and biomedical domain knowledge. The ontologies are implemented in OWL-DL, which is the OWL subset based on Description Logics. By proceeding in this way, the domain knowledge is made explicit in a set of OWL ontologies and therefore ready to be exploited by means of automated reasoning. In the next sections, further details of how such ontologies are used in our work will be provided. In this paper, we

[12] `http://www.w3.org/TR/owl2-overview/`

will use the Manchester OWL Syntax[13] to show OWL content.

3. Methods

3.1. Clinical Models in OWL

In recent years, works like Martínez-Costa et al. (2009); Sari et al. (2012); Tao et al. (2013); Iqbal (2011); Schulz and Martínez-Costa (2013) have provided OWL representations for clinical information and clinical models of different EHR standards like ISO 13606, openEHR, HL7 or Clinical Element Models (CEM)[14]. Such representations include the information model and the constraints-based clinical models. Most of such representations use OWL classes to represent clinical models in OWL because they pursue the use of OWL reasoning for validating the models or representing data instances as individuals of such classes. However, the representation based on individuals proposed in Martínez-Costa et al. (2009) has demonstrated its usefulness in interoperability settings.

In this work we use two different OWL representations for archetypes for both ISO 13606 and openEHR, since they are the standards currently dealt with in our approach. Both representations have in common that the information model is represented in OWL, but differ in the type of OWL entity used for representing archetypes:

1. Archetypes are represented as OWL individuals of the corresponding information model ontology (i.e., ISO 13606, openEHR), as described in Martínez-Costa et al. (2009) . The information model ontologies were obtained by making a semantic interpretation of the specifications of the EHR standards.This representation has demonstrated its usefulness to support the transformation of both archetypes and EHR extracts between openEHR and ISO 13606 Costa et al. (2011). Besides, we use it for adding archetypes annotations (based on external ontologies/terminologies) which are exploited with different purposes (e.g. archetype comparison and search).
2. Archetypes are represented as OWL classes, as described in Menárguez-Tortosa and Fernández-Breis (2013). The information model ontology were obtained by applying the Ontology Definition Metamodel

[13] `http://www.w3.org/TR/owl2-manchester-syntax/`
[14] `http://www.clinicalelement.com`

(ODM)[15] specification to represent UML models in OWL. We use this representation for tasks that require performing automated reasoning over their content, like validating the correctness of specialised archetypes.

3.2. Clinical Data in OWL

Most legacy EHR systems use relational databases to store the clinical data, but recent EHR specifications like ISO 13606 and openEHR represent EHR data extracts in XML. Both relational databases and XML technologies have many limitations for the semantic processing of information and do not provide support for automatic reasoning such as OWL. In relational databases one might propose to transform tables into ontology classes, records into individuals, and columns into properties. Such a transformation is not more than a change of format, because the real meaning of the entities represented is completely ignored in such a process, which is therefore not enough to achieve semantic interoperability. There exist different tools for performing a canonical transformation of XML content or relational databases into semantic formats. Some examples of such tools are presented in Auer et al. (2009); Bizer and Cyganiak (2006); Rodrıguez-Muro et al. (2012)). The main limitation of such tools, according to our objectives, is that they perform a generic transformation of the data, that is, they do not take into account their underlying model of meaning.

In our approach, the transformation is based on defining specific mappings between the corresponding data and knowledge schemas. We use a generic approach driven by domain knowledge to transform EHR data into OWL (see Figure 2). Given that we are interested in working with archetype-based EHR standards, we are interested in transforming XML EHR extracts into OWL. The transformation process is guided by the mapping between the archetypes used to capture the EHR data and the target domain ontology. The mapping rules define how archetyped data are transformed into ontology individuals, and such rules must be defined by an expert.

We propose the use of ontology patterns (see Falbo et al. (2013)) to facilitate the definition of the mapping rules. Such patterns define the templates for the creation of OWL individuals with specific types of axioms. A pattern is created from an ontological representation of a domain, which includes for-

[15] `http://www.omg.org/spec/ODM/1.0/`

mal descriptions of types of entities and their relations. Ontology patterns are based on the target domain ontology and have both fixed and variable parts. Their variable parts are mapped to the archetype data elements that are later instantiated with data.

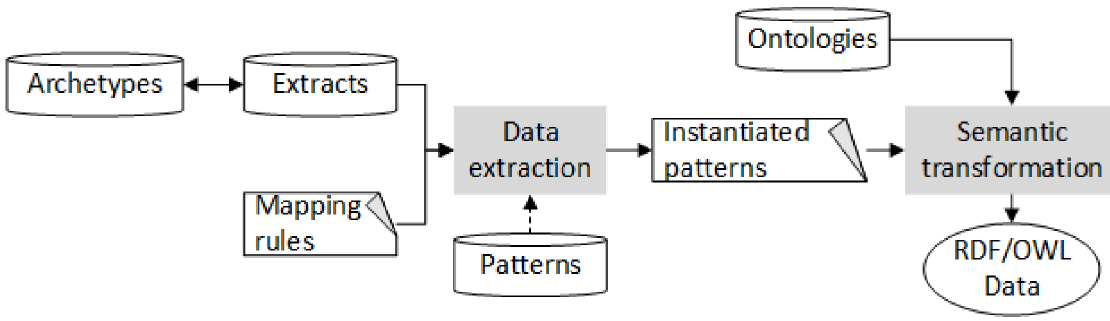

Figure 2: **Overview of the method for the transformation of archetype-based EHR data**

Once the mappings between an archetype and the target ontology are defined, they can be applied to any data extract conforming to that archetype for getting data in OWL format. Our current process uses the ontology patterns defined in the context of the European project SemanticHealthNet (SHN) [16], which pursue to ease the mapping process and to create semantically interoperable datasets in OWL. Besides, the patterns are currently implemented using the Ontology Pre-Processing Language version 2 (OPPL2)[17]. OPPL2 is a scripting language for OWL that can be used to modify the axioms of an ontology using a pattern approach, offering an API that permits the execution of the patterns while controlling the transformation processes. OPPL2 works in conjunction with reasoners, which has advantages for our purpose: (1) defining patterns that exploit inferencing; (2) ensuring the transformation of only logically consistent content of the clinical models.

An excerpt of the SHN Medication Administration pattern in OPPL2 is shown next. This pattern defines that a medical record (*?medRecord*) is a Plan 'realized by' a Medication Administration 'to a patient' of a prod-

[16] `http://www.semantichealthnet.eu/`
[17] `http://oppl2.sourceforge.net/`

uct. The prefixes btl, sct and shn refer to BioTopLite, SNOMED CT and SHN ontologies respectively. This pattern has two input variables, namely, *?medRecord* and *?product*, which are, respectively, a medical record and a subclass of *PharmaceuticalProduct*. *?medRecord* is the new individual to be created, while *?product* is the variable to be instantiated. The OWL axioms to be created upon each execution of the pattern are included between the BEGIN/END keywords. A mapping rule for this pattern, when applied to a medical record, only needs to instantiate the *?product* variable and the patterns completes the rest by itself.

```
?medRecord:INDIVIDUAL,
?product:CLASS[subClassOf sct:PharmaceuticalProduct]
BEGIN
  ADD ?medRecord Type btl:Plan and
    btl:hasRealization only (sct:MedAdmin and btl:hasPatient some ?product);
END
```

3.3. Semantic Annotation

Semantic annotation is the process of adding semantic metadata to content. Kiryakov et al. (2004) states that it is about assigning to the entities in the content links to their semantic descriptions. According to Oren et al. (2006), annotation processes can be manual, semi-automatic and automatic:

- Manual methods permit the user to access the repository of semantic resources, with the support of search methods, and to manually select the annotation term. Manual methods require the participation of domain experts, who are able to produce accurate annotations, but manual methods are also prone to errors.

- Automatic methods do not only suggest terms but also select them. These methods do not require user interaction but the existence of a set of annotation rules. The quality of such rules will determine the quality of the automatic annotations.

- Semi-automatic methods are able to find related terms that are suggested to the user to select the annotation terms. In these methods, the users validate the recommendations and such confirmations are usually learned by the machine to improve the future recommendations.

In our context, the semantic descriptions are provided by biomedical terminologies/ontologies and external semantic resources and we annotate both archetypes and EHR extracts. Archetypes are already linked to terminologies by terminological bindings. However, these terminological bindings are not always defined and, depending on the specific use of the archetype, additional semantic meaning may be needed, for example to perform a personalized classification of patients. We combine both manual and semi-automatic methods in order to make easier the annotation task. Given a concrete repository of ontologies, controlled vocabularies and terminologies in OWL format, our annotation method recommends annotations based on the textual content of the archetype and permits to retrieve exact or partial matches between the content of the archetype and the terms of the entities included in the repository. Besides, the user is also provided with a search facility, which would retrieve the corresponding terms from that repository. Therefore, we ae handling both the terminological bindings and the annotations added by the user. The group of all the annotations are a representative generalization of the semantic knowledge contained in the archetype and creates what we name the semantic profile of the archetype, useful for performing new actions, such as comparison of archetypes for finding similar meaning. For this purpose, the representation of archetypes as OWL individuals is used. We represent the annotations in OWL format, so they can be jointly exploited with the content of the archetype. Both these annotations and the terminology bindings are exploited in our approach for different tasks, like archetype comparison and search. EHR data are indirectly annotated, mainly through the annotations of the archetypes used to capture the data.

3.4. Semantic Profiles

A semantic profile is defined in Bhatt et al. (2009) as a semantic description of a dataset, which constitutes a semantic interpretation of it. Semantic profiles permit efficient, effective processing without needing to use the whole information about a particular information entity, using only such semantic interpretation. Using ontologies to do such interpretation permits to make decisions and recommendations based on formal specifications of domain knowledge.

Our main information processing entities are archetypes and EHR extracts, which can be directly and indirectly annotated, respectively. Such annotations are the basic elements to build the semantic profiles. In terms of representation, a semantic profile is represented as a set of semantic annotations.

The semantic profile of an archetype A is obtained as described in Equation 1. It is the union set of all the terminological bindings of the archetype (Ti) and its external annotations (Ci).

$$Semantic\_profile(A) = \{T_1, T_2, ..., T_n\} \cup \{C_1, C_2, ..., C_n\} \quad (1)$$

The semantic profile of an EHR extract is obtained from two sources, namely, archetype and data. First, the semantic profile of the EHR data is derived from the archetype to which the archetype-base data conform (e.g. the semantic profile of an extract about blood pressure will be the semantic profile of the blood pressure archetype). Second, the EHR data permits to define a more precise profile. For example, in case of having a low value for the blood pressure, the semantic profile could include the annotation "hypotension". The application of this level requires the availability of classification rules (see Section "OWL Reasoning-based Data Classification"), which would permit to know when a patient has to be associated with "hypotension".

3.5. Semantic Similarity

The adoption of ontologies for annotation provides a means to compare entities on aspects that would otherwise not be comparable, as described in Pesquita et al. (2009). Semantic similarity measurements use a semantic structure as a context for the estimation of distances or similarities between domain entities. There are two main approaches for semantic similarity, presented in Rada et al. (1989); Lord et al. (2003); Pesquita et al. (2009):

- Edge-based approaches count the number of semantic links in the ontology between two classes (see for instance Rada et al. (1989)). In these approaches, there is no distinction between types of semantic links like taxonomic or mereology.

- Node-based approaches not only take into account the edges but also the properties of the classes involved (see for instance Lord et al. (2003)).

In our work, semantic annotations can be associated with archetypes and EHR data extracts. Besides, ontologies drive the representation in OWL of archetypes and EHR extracts as it has been described in previous sections.

Studying the similarity of archetypes is worthy because the same clinical concept can be expressed in many ways using the same or different reference information models. This makes clinical content interoperability and sharing more complex, and finding similarities can help to bridge different representations.

Figure 3 shows an overview of our archetype similarity process. The input to the similarity method is the two archetypes to be compared from the archetype repository, and the output is a score in the range [0,1]. Retrieving the archetypes from the repository enables to access the semantic profile of the archetype, which includes the semantic annotations and the terminology bindings. Besides, the ontologies of the information model, archetype model and the ones used in the semantic annotations of the semantic profile of the archetypes provide the semantic context for comparing the archetypes.

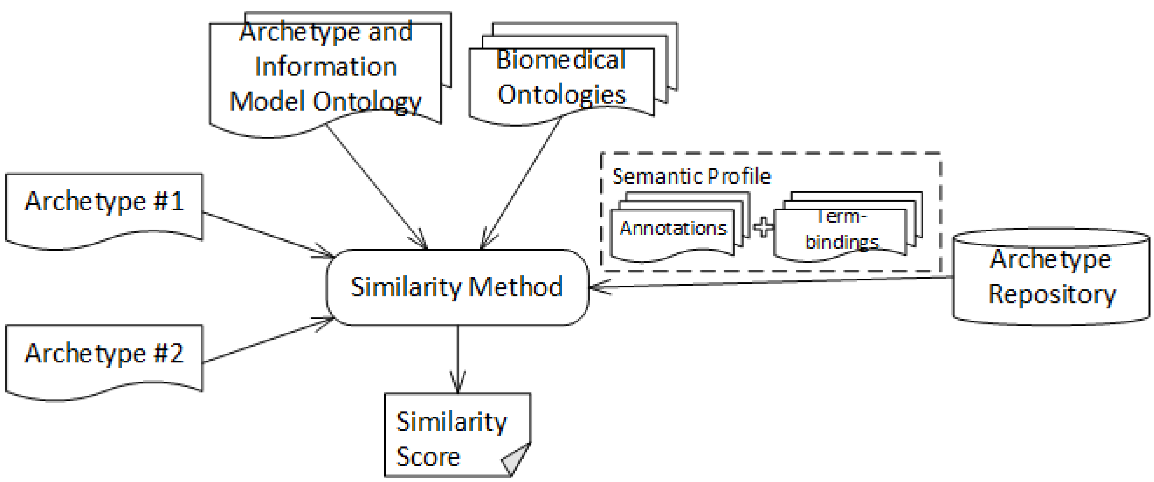

Figure 3: **Similarity process overview**

Our similarity approach is node-based because, besides the hierarchical structure of ontologies, we exploit properties like the terminological bindings and the semantic annotations. Basically, the method compares all the pairs of elements in the semantic profiles of the archetypes, obtaining a similarity score for each pair. These pairwise analyses return a set of pairs with the following steps:

- Compare all the pairs and select those with score higher than a given threshold

- Get the set of pairs that maximize the sum of the similarity scores that include only one pair per element of the semantic profile of each archetype

The pairwise similarity function for two elements of the semantic profile is based on the ones described in state of the art literature (see for instance Resnik (2011); Sánchez-Vera et al. (2012)). In particular, our similarity function uses the following factors:

- Taxonomic similarity (d): This distance measures the hierarchical distance between the classes associated with the two elements C_i and C_j, that is, through taxonomic links. This function uses the union set of ancestors of both classes and the set of common ancestors of both classes. Thus, this score is calculated as shown in Equation 2:

$$d(C_i, C_j) = 1 - \frac{|ancestors(C_i) \bigcup ancestors(C_j)| - |ancestors(C_i) \bigcap ancestors(C_j)|}{|ancestors(C_i) \bigcup ancestors(C_j)|} \tag{2}$$

  It should be noted that classes might present multiple inheritance, which would imply different taxonomic paths and, therefore, different taxonomic similarity scores. In such cases, the shortest distance is returned by the function.

- Properties similarity (ps): Similarity between the set of properties associated with the classes associated with the two elements, which is calculated as shown in Equation 3.

$$ps(C_i, C_j) = \frac{|common(C_i, C_j)|}{|common(C_i, C_j)| + y_1 * |different(C_i, C_j)| + y_2 * |different(C_j, C_i)|} \tag{3}$$

  where y_k refers to the weight given to each one of the metrics, $0 \leq y_k \leq 1$, $\sum y_k = 1$.

- Linguistic similarity (ls): A string-based calculation of the terms associated with the ontological elements compared. If we are comparing two concepts from the OWL representation of two archetypes, this calculation uses the term definition of both concepts. When comparing two concepts from a terminology, it uses labels or the local name of the concepts compared. Our current implementation uses the Levenshtein distance defined in Levenshtein (1966).

The similarity between two elements of the semantic profile of the archetypes is calculated as shown in Equation 4:

$$pairwise\_similarity(C_i, C_j) = w_1 * ls(C_i, C_j) + w_2 * ps(C_i, C_j) + w_3 * d(C_i, C_j) \tag{4}$$

where w_k refers to the weight given to each one of the metrics, $0 \leq w_k \leq 1$, $\sum w_k = 1$.

The sum of all the pairwise similarities between the selected pairs of elements of the semantic profiles returned by the pairwise analysis method constitutes the similarity of the semantic profiles of the archetypes, written *profile_similarity*. In addition to this, our similarity method includes another factor that takes into account the structural types of the archetypes compared in the context of the information model ontology. This factor, written structural similarity, assumes that two archetypes of the same type COMPOSITION are more similar than two archetypes of different types, for instance, COMPOSITION and SECTION. This score is obtained by applying the taxonomic similarity function to the types of both archetypes. The final semantic similarity of two archetypes is obtained by Equation 5:

$$\begin{split} semantic\_similarity(A_1, A_2) = z_1 * structural\_similarity(A_1, A_2) + \\ z_2 * profile\_similarity(A_1, A_2) \end{split} \tag{5}$$

where z_k refers to the weight given to each one of the metrics, $0 \leq z_k \leq 1$, $\sum z_k = 1$.

3.6. OWL Reasoning-based Data Classification

A major motivation for using OWL in this work is its capability to perform sound and complete automated reasoning. OWL-DL classes have sets of axioms associated and two types of axioms are relevant for reasoning: (1) *subClassOf* axioms permit to define the necessary conditions for the members of a given OWL class; and (2) *equivalentClass* axioms permit to define which conditions would be sufficient for an OWL individual to be classified as a member of a given OWL class. *EquivalentClass* axioms are useful for defining clinical inclusion/exclusion criteria because they would enable the reasoner to automatically partition the clinical data into the groups of clinical interest defined by such criteria.

Applying OWL reasoning over data requires to have an OWL dataset based on an OWL ontological infrastructure. Such ontological infrastructure has to provide the background domain knowledge and the classification rules. In this work, we are interested in applying OWL reasoning to classify patients according to the available EHR data. Patient classification means grouping patients in different categories conforming to certain clinical criteria. For example, in our case study in colorectal cancer, the domain knowledge specifies entities like finding, adenoma, histopathology report, etc., whether the classification rules define the groups by level of risk into which we want to classify the patients. Whereas the domain knowledge is usually modelled using *subClassOf* axioms, the classification rules are specified using *equivalentClass* ones. Ideally, the classification rules are implemented in an ontology that reuses the domain ontologies previously developed, which will permit the joint exploitation of classification rules, domain knowledge and EHR data by means of automated reasoning. We call such ontology a classification ontology and it contains, at least, one class per group of interest.

Once this ontology is ready, an OWL-DL reasoner like Hermit (see Shearer et al. (2008)) can be applied over the complete semantic dataset to infer all the possible information given the data. The result of such inference process will be the resulting classifications, which can be retrieved using semantic query languages like DL-query[18] or SPARQL[19], or through a programmatic API like OWLAPI[20].

In this work we assume that categories can be specified in terms of rules expressed as OWL DL defined classes and it is assumed that the EHR data are available in RDF format conforming to OWL. Our classification method for a given patient has two inputs: the OWL ontology with the classification rules and the OWL representation of the patient EHR data. The classification rules are applied to the data by using DL reasoning, which permits to obtain the categories to which the patient belongs. This approach is generic, and the same patient data can be automatically analysed and classified by applying

[18] `http://protegewiki.stanford.edu/wiki/DLQuery`

[19] `http://www.w3.org/TR/rdf-sparql-query/`

[20] `http://owlapi.sourceforge.net/`

many different classification ontologies.

For example, a rule could be established that a patient has hypotension when the systolic blood pressure is less than 90 mm Hg or the diastolic ones is less than 60 mm Hg. This rule could be easily coded into OWL-DL as follows:

```
Hypotensive equivalentClass (Patient and
                    ((systolic some integer[<= 90 ]) or
                    (diastolic some integer [<= 60 ])))
```

Hence, if our dataset contains data about Patients with properties *systolic value 80* or *diastolic value 50*, then such patients would be classified as members of the class Hypotensive. Such classifications are also used in our approach to enriching the semantic profile of the patient, since they can be represented as new annotations associated with a given EHR extract.

4. Results

4.1. ArchMS

ArchMS[21] is a prototypical tool for the management of clinical archetypes and data that facilitates the reuse of clinical data in multiple scenarios by applying semantic web technologies. Figure 4 shows an overview of the architecture of the platform. In the figure, software modules are represented as boxes and data repositories are represented as storage boxes.

As it can be seen in Figure 4, ArchMS has different types of repositories:

- The repository of semantic resources (lower part of the figure) contains both local resources like EHR ontologies and local ontologies uploaded by the users and external resources, in this case, Bioportal ontologies. Such semantic resources are ontologies, terminologies and controlled vocabularies available in OWL format, which are used for the tasks of archetype annotation, similarity, data transformation, semantic profiling and classification.

- The repositories that link archetypes and ontologies (right part of the figure) contain (1) the mappings between the archetypes and the domain ontologies, including the semantic patterns, which are used for

[21] `http://sele.inf.um.es/archms`

transforming the data into OWL; and (2) the associations between the archetypes and the classification ontologies, which are use for classifying data instances.

- ADL, XML and OWL repositories of archetypes and clinical data (central part of the figure), which store their representation in different formats.
- The repository of semantic annotations (upper part of the figure) for both archetypes and clinical data.

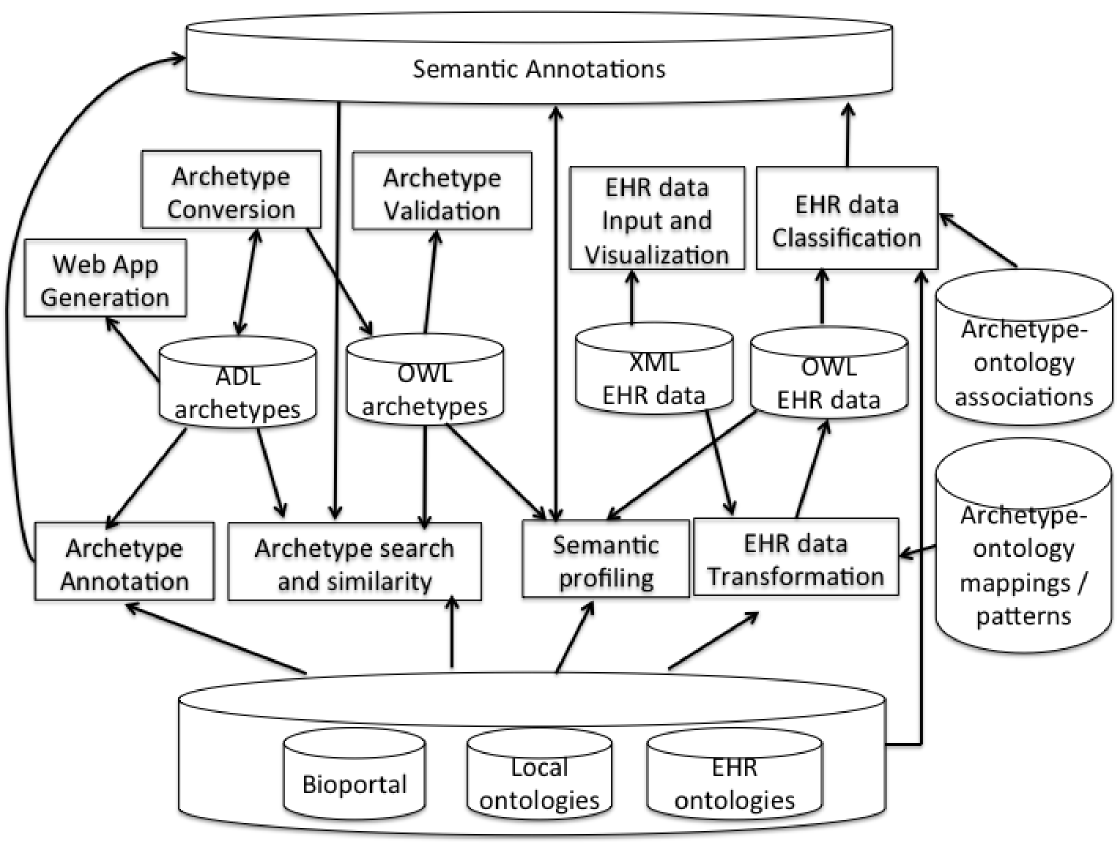

Figure 4: **ArchMS architecture overview**

4.1.1. Archetype functionality

As shown in Figure 4, the main activities that can be performed with archetypes are: conversion, validation, annotation, search for similar archetypes or for archetypes with concrete properties, and the generation of form based/GUI applications. The ArchMS user interface for working with archetypes is depicted in Figure 5. Next, we describe how the major options have been implemented and grouped in ArchMS.

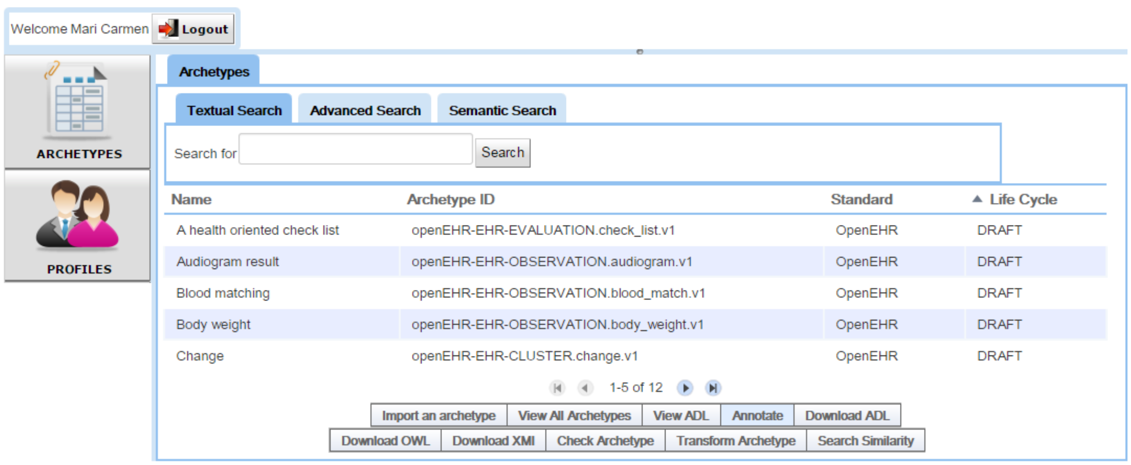

Figure 5: **Screenshot of the options available for archetypes**

- Archetype management: The system allows importing ADL1.4 archetypes for both openEHR and ISO 13606 representations. It includes the functionality provided by our previous developed tools Archeck[22] and PoseacleConverter[23]. The first one allows checking the consistency of a specialized archetype regarding its parents. The PoseacleConverter allows transforming archetypes from openEHR into ISO 13606 and viceversa and representing them in OWL. Different approaches are used to provide persistence to archetype related information. On the one hand, ArchMS uses a relational database (ADL Archetypes) to store archetype metadata properties such as name, language, or purpose,

[22] `http://miuras.inf.um.es/archeck/`
[23] `http://miuras.inf.um.es/PoseacleConverter/`

as well as the ADL file. Basically, this repository contains the non-semantic content of the archetype. On the other hand, ArchMS uses a semantic repository (OWL Archetypes) implemented using Jena[24] to store the OWL representation of the archetypes, allowing for issuing SPARQL queries. In order to speed-up queries, an archetype Lucene index is obtained by parsing the ontology and keywords sections of the archetypes. When an ADL archetype is imported into the system, the following activities are made:

1. Check the correctness using Archeck, which requires to represent the archetype in OWL as classes.
2. If the archetype is not correct, the next steps would not be executed.
3. Store this OWL representation in the OWL Archetypes repository.
4. Store the ADL content in the ADL Archetypes repository and create the Lucene index.
5. Apply PoseacleConverter to get the OWL individuals-based representation of the archetype. If the transformation is correct, then store the content in the OWL Archetypes repository.

In Menárguez-Tortosa et al. (2012) we described a generator of web applications based on archetypes, called ArchForms. ArchMS integrates it as a service, so the ArchMS administrator can generate ArchForms applications and make them available to other users for downloading and further deployment. In addition, as shown in Figure 5, the user can download the archetype in ADL and OWL formats and can perform, at any time, the validation of the archetype and the transformation of the archetype from ISO 13606 to openEHR and vice-versa.

- Archetype annotation: ArchMS implements manual and semi-automatic methods for the annotation of the archetypes. Bioportal, which contains more than three hundred biomedical ontologies, terminologies and controlled vocabularies, is the main source of annotation resources for users. In order to get recommended annotations terms from Bioportal, we use the Bioportal Web Services, presented in Whetzel et al. (2011), which provide candidate terms for a given text content. Besides, we use

[24] `https://jena.apache.org/`

a Lucene-based search method to look for exact and partial matches between the text and the content of the local semantic resources. The annotations of the archetypes confirmed or selected by the users are stored in the Semantic Annotations repository. It should be noted that, once an archetype is successfully imported, ArchMS suggests potential annotations from the semantic resources by processing the text content of the archetype.

- Archetype search and similarity: The central part of Figure 5 displays the user interface with different query options that exploit both the relational and semantic repositories:
  1. The textual search interface uses the Lucene index for finding archetypes that contain some textual description which matches the textual description of the query. It returns the archetypes that contain the search text in their textual content, including metadata properties like name, language, purpose, etc. The query “histopathology English” would return the archetypes that contain “histopathology” or “English” in any text field.
  2. The advanced search can be considered a faceted search, since it permits to find archetypes by metadata (e.g. language, archetype name, etc.) and by archetype annotations, in case they are available. The query “annotation:histopathology and language:English” would return the archetypes available in “English” that contain “histopathology” as an annotation.
  3. The semantic search executes a SPARQL query against the semantic repository of archetypes. This search facility exploits the representation of archetypes as OWL individuals. This method would permit to exploit the hierarchical structure of ontologies, terminologies and controlled vocabularies, and it would also be able to apply semantic similarity measurements to retrieve archetypes similar to the ones we are searching for. For example, if we look for “Histopathology” as a SNOMED-CT term and there are no archetypes with bindings or annotations to such concept, other ones could be suggested if they contain annotations or binding similar enough to “Histopathology”. Our semantic search method is based on the work presented in Martínez-Costa et al. (2010b) and allows formulating queries by selecting the corresponding entities from the ontologies contained in the repository, and which

are automatically translated into SPARQL. Therefore, we are providing a specific semantic query language for archetypes based on SPARQL.

On the other hand, ArchMS implements the semantic similarity function described in the Section "Semantic Similarity", which permits to find which archetypes are similar. Archetypes can be compared in ArchMS in two ways. On the one hand, a user can decide to search for similar archetypes. In such case, the user interface would permit the user to specify the similarity threshold and the weights. Such threshold can automatically get a value 1, which means, equivalence, if the strict comparison method is selected. On the other hand, once an archetype is successfully imported, ArchMS looks for similar archetypes using the default similarity threshold. Such search has two objectives: (1) checking whether an equivalent archetype already exists in the repository; (2) recommending annotations associated with similar archetypes.

4.1.2. Data functionality

As shown in Figure 4, the main activities that can be performed with EHR data are to obtain the OWL representation of XML extracts, to visualize and input EHR data, to obtain the semantic profiles of EHR extracts and to classify EHR data. The ArchMS user interface for working with data is depicted in Figure 6. The upper part of the figure shows the option of an EHR extract. The lower part of the figure lists the archetypes associated with the extract. Next, we describe how these options have been implemented in ArchMS.

- Data management: ArchMS processes XML EHR extracts from both ISO 13606 and openEHR specifications, which have to be imported into the system. ArchForms applications permit to capture data based on the archetypes used to create the applications and to export such data as XML extracts. Such extracts can be uploaded into ArchMS, but they can be generated using any other ISO 13606 or openEHR data management system. Besides, ArchMS provides a simple visualization of the content of the extract which has been reused from the interface used in ArchForms applications. Such option is available once the extract has been successfully imported, as seen in Figure 6. In the same

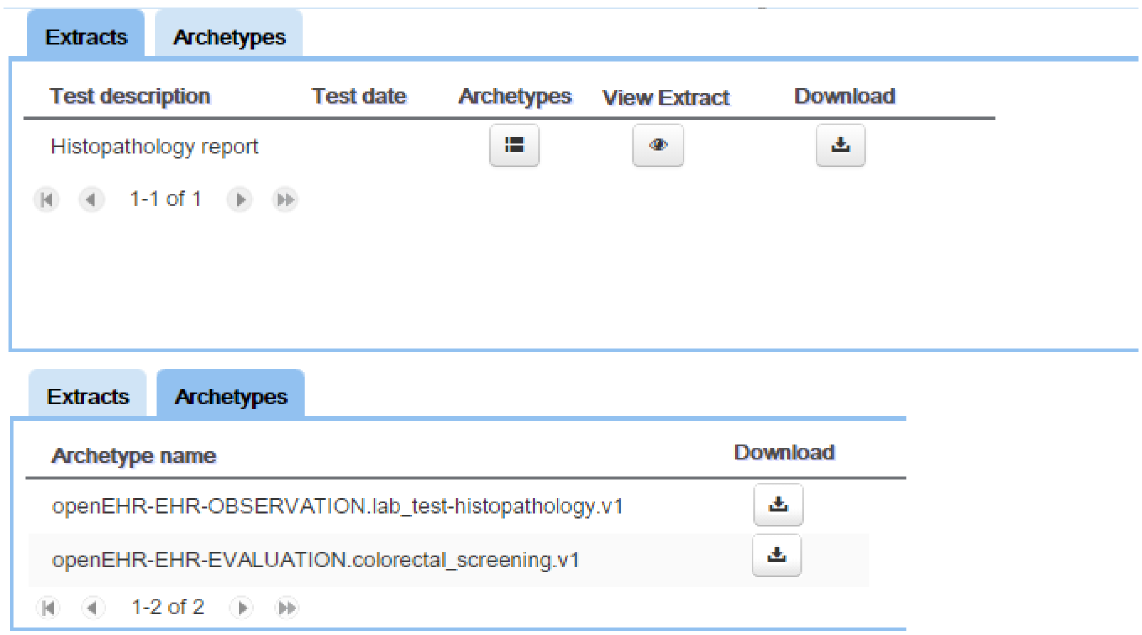

Figure 6: **Screenshot of the options available for data extracts**

figure, we can also see that we can access the archetypes used to capture the data. In order to simplify data import tasks, ArchMS permits to import XML files containing information from multiple extracts of the same patient, that is, different extracts captured using different archetypes. When this option is used, it will be stored as one extract associated with multiple archetypes.

- Data transformation and profiling: ArchMS enables to perform semantic activities on EHR data by transforming EHR extracts into OWL. This function uses the mappings between the archetypes and the ontologies, which have been previously uploaded to ArchMS. Such mappings can be created using our Semantic Web Integration Tool (SWIT)[25], which implements the methods described in the Section "Clinical Data in OWL". SWIT services are invoked from ArchMS to automatically execute the data transformation once the extract has been imported

[25] `http://sele.inf.um.es/swit`

into ArchMS. The corresponding content in OWL is stored in the OWL EHR data repository. Once transformed into OWL, the semantic profile of the EHR data is automatically extracted. Such profile is updated when: (1) the annotations associated with the archetypes used to capture the data change; (2) new classifications of the patients are available (see next item).

- Data classification: ArchMS permits to classify EHR data according to rules that define the clinical status of the patients. For this purpose, the user has to select the classification ontology. In order to be effective, this classification ontology must have been implemented in OWL-DL and using *equivalentClass* axioms describing the classification groups for the patients. This requirement is due to the use of DL-reasoning in ArchMS. Our current implementation uses Hermit as reasoner. ArchMS permits to associate classification ontologies with archetypes, which means that every time EHR data captured with such archetype is classified, such ontology can be used. When selecting the classification option for a patient, the classes to which the patient belongs are shown. This means the classes according to all the classification ontologies associated with the archetypes used to capture the data. ArchMS stores such classifications as annotations of the EHR data in the Semantic Annotations repository.

4.1.3. Users

ArchMS manages three types of users, namely, administrator, patient and physician. The administrator is in charge of the maintenance of the system and will be responsible for particular tasks like assigning the physician role to the corresponding users or generating the ArchForms applications. The administrator is also in charge of defining the semantic profile of the archetypes, annotating them either using directly biomedical terminologies or recommendations from similar archetypes, and uploading the mappings between archetypes and ontologies. The result of this process is a set of repositories of archetypes that can be semantically exploited by the users. Patients have access to their clinical data and additional knowledge based on their profiles, depending on classification resources. Physicians can access to the medical histories of all their patients, get additional knowledge from the clinical data of their patients, such as their classification according to their risk of developing colorectal cancer, and use advanced process for transforming the clinical data and creating repositories suitable to their research.

4.2. ArchMS in a colorectal cancer study

In the previous section we have described the ArchMS functionality. Now, we describe some of the tool functions by their application to a colorectal cancer study. In this study we used more than 20,000 anonymized patients' records from a colorectal cancer screening program of the Region of Murcia in Spain. The objective of the study was to classify patients in risk groups by applying the European (see Atkin et al. (2012)) and American[26] colorectal cancer guidelines and to study the existence of discrepancies with the classifications recorded in the database.

The European guideline defines three levels of risk for patients (low, intermediate, and high) whereas the American one defines only low and high. In both guidelines, the group of risk is assigned depending on the number, type and size of the adenomas found during colorectal cancer screening tests. For example, both guidelines define that patients with less than 3 normal adenomas whose size is less than 10 mm are classified as low risk. Table 1 shows an example of the colorectal cancer screening data of a patient. This patient has three findings, being just one of them a Normal Adenoma of size 5 mm. This would be an example of low risk patient.

In such study, the patient cohorts were identified by combining archetypes and semantic web technologies. The complete details of the study can be found in Fernández-Breis et al. (2013). Here, we focus on how the semantic activities offered by ArchMS were useful to carry out this study.

Table 1: Colorectal cancer screening test of a patient

Finding	**Endoscopic configuration**	**Dysplasia Type**	**Anatomical pathology**	**Max Size**	**Adenoma**
1	Sessile	Unknown	Hyperplasia	2	No
2	Sessile	Unknown	Hyperplasia	2	No
3	Sessile	Low degree	Tubular adenoma	5	Normal Adenoma

[26] `https://www.nice.org.uk/guidance/cg131`

4.2.1. Selecting the archetypes

The source data for the study was a relational database of colorectal cancer screening tests. Such data have to be transformed into XML EHR extracts according to openEHR or ISO 13606 to be imported into ArchMS. Such transformation can be executed with external tools like LinkEHR, but requires a series of archetypes that model colorectal cancer data. Hence, the first step is to find such archetypes. By analysing the data we concluded that archetypes for recording histopathology and colorectal screening information were required.

The ArchMS search options can be used to find archetypes of interest for the domain of colorectal cancer screening. We used the textual search for finding the archetypes providing the best match our data (see Figure 7). For that we used some keywords extracted from the original records such as “histopathology”, “finding”, “size”, “dysplasia” or “sessile”. As a result, we retrieved an archetype suitable for the histopathology report (i.e. openEHR-EHR-OBSERVATION.lab_test-histopathology.v1), but the repository did not contain any appropriate for colorectal screening.

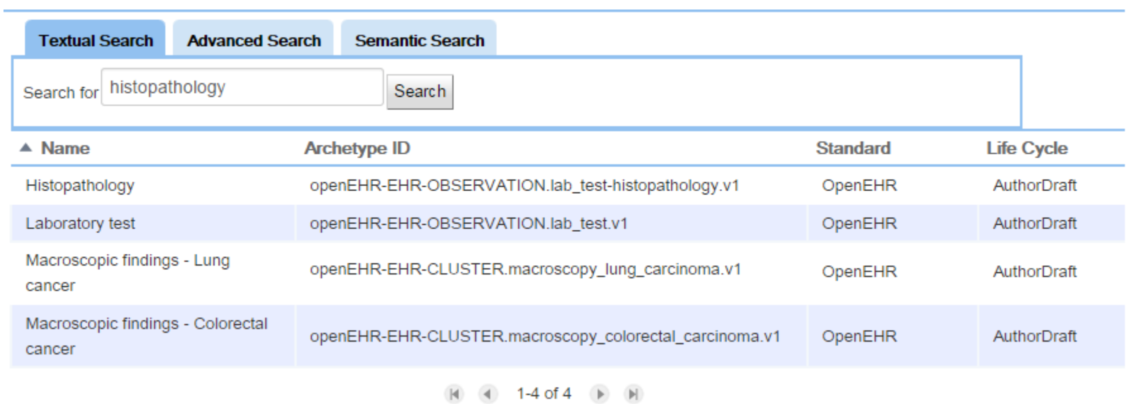

Figure 7: **Use of textual search with keyword histopathology**

Since our study required recording more information than the one provided by the histopathology archetype (e.g. adenoma findings information such as type, maximum size, dysplasia grade, sessile and/or advanced finding, etc.), we specialized the existing archetype (see left side Figure 8). Besides, we created an archetype for colorectal screening (see right side Figure 8) which allows recording study related information (e.g. maximum size of all adenomas and number of adenomas). We used the archetype editor LinkEHR for editing both archetypes.

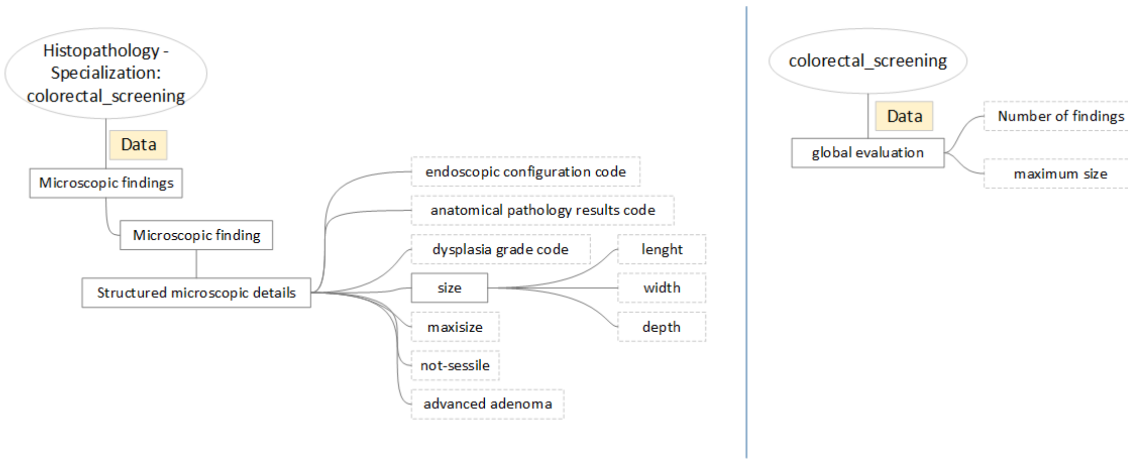

Figure 8: **(Left) Excerpt of Histopathology - Specialization colorectal screening (openEHR-EHR-OBSERVATION.lab_test-histopathology-colorectal_screening.v1); (Right) Excerpt of Colorectal Screening archetype (openEHR-EHR-EVALUATION.colorectal_screening.v1)**

4.2.2. Importing and annotating the archetypes

The two mentioned archetypes were then imported into ArchMS. This implies checking the consistency of the imported archetypes regarding its parents, if any, by using the Archeck functionality. In our case, the histopathology archetype has been specialised for recording colorectal cancer related information and, therefore, the correctness of its specialisation is checked. Once the archetypes have been successfully imported, the tool automatically looks for related archetypes in the repository in order to be able to reuse their semantic annotations. Two similar archetypes were found for openEHR-EHR-OBSERVATION.lab_test-histopathology-colorectal_screening (see Figure 9), both of them related with it through specialisation. No similar archetypes were found for openEHR-EHR-EVALUATION.colorectal_screening, what is explained by the fact that no archetype related to such domain had been found in the repository in the search process.

In addition to the annotations retrieved from similar archetypes (see Figure 9), we added new annotations from SNOMED CT and MeSH. This

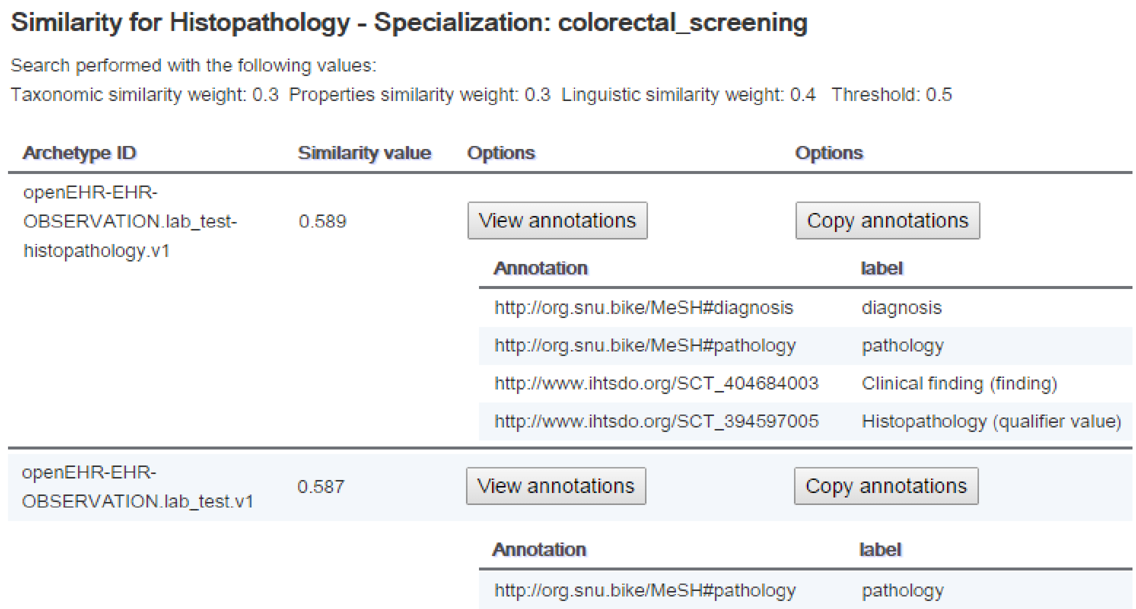

Figure 9: **Similar archetypes to openEHR-EHR-OBSERVATION.lab_test-histopathology-colorectal_screening.v1 and their annotations**

annotation is supported by the tool by suggesting possibly related terms based on the archetype textual content in their keywords and ontology sections. Figure 10 shows part of the recommendations for openEHR-EHR-OBSERVATION.lab_test-histopathology-colorectal_screening.v1 archetype. The archetype contains in its keywords and ontology sections terms such as "accession", "adenoma", "colorectal", etc. The system looks for matches for those terms in the selected annotation resource (MesH in this case) and shows the selected terms to the user. For this example, "adenoma", "adenomatoid_tumour" or "adenomatous_polyp" are the MesH terms recommended for the query "adenoma". Tables 2 and 3 show the final set of annotations added to the histopathology and colorectal cancer archetypes respectively.

4.2.3. Getting the EHR data in OWL

In our study, the patient data were provided in a relational database. We transformed them into XML openEHR data extracts by using LinkEHR, and we used the SWIT services in ArchMS to define the mappings between the archetypes and the ontology and to transform the data into OWL.

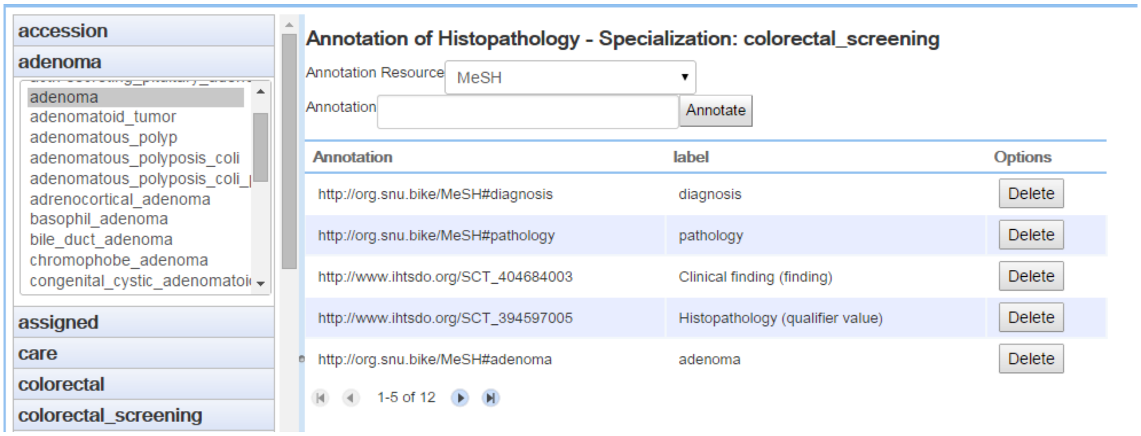

Figure 10: **Recommendation of annotation for openEHR-EHR-OBSERVATION.lab_test-histopathology-colorectal_screening.v1**

Table 2: SNOMED-CT and MESH annotations for the histopathology colorectal screening archetype

SNOMED-CT code	Label	MeSH code	Label
25723000	Dysplasia	D000236	Adenoma
394597005	Histopathology	Q000175	Diagnosis
264267007	Colorectal	D003106	Colon
148322003	Screening	D010336	Pathology
404684003	Clinical Finding	D012007	Rectum
32048006	Adenoma	D008403	Mass Screening

Table 3: SNOMED-CT and MESH annotations for the colorectal screening archetype

SNOMED-CT code	Label	MeSH code	Label
264267007	Colorectal	Q000175	Diagnosis
148322003	Screening	D008403	Mass Screening

Figure 11 describes part of the mapping for the data transformation process between the archetypes and the domain ontology for colorectal cancer[27]. In this case we used a pattern that defines a histopathology report according to the domain ontology, which contains a set of findings (*hasFinding* property), records the total number of adenomas found (*number* property) and the size of its biggest adenoma (*maxsize* property). The pattern in OPPL2 is shown next:

```
?histopathologyReport:INDIVIDUAL,
?finding:INDIVIDUAL,
?size:CONSTANT,
?number:CONSTANT

BEGIN
ADD ?histopathologyReport instanceOf HistopathologyReport,
ADD ?histopathologyReport hasFinding ?finding,
ADD ?histopathologyReport number ?number,
ADD ?histopathologyReport maxsize ?size
END;
```

The variables of the pattern (preceded by ?) represent the parts that are instantiated to the source clinical data and, therefore, they are linked to the relevant elements of the archetype. The relations between the variables do not need to be established for each data instance since they are already defined in the pattern, so the user does not need a deep knowledge of the definition of individuals in the ontology since it is included in the pattern. Each extract captured using openEHR-EHR-OBSERVATION.lab_test-histopathology-colorectal_screening.v1 archetype corresponds to a histopathology report, so we link the root of the archetype with the *?histopathologyReport* variable. The "Microscopic finding" element represents the findings reported, so we link that element with the variable *?finding*, whereas the values for the variables *?size* and *?number* are obtained from openEHR-EHR-EVALUATION.colorectal_screening.v1 (see Figure 11).

4.2.4. Classification of patients

A major goal in this study was to classify patients according to their risk of developing colorectal cancer. The rule for classifying patients with low

[27] `http://miuras.inf.um.es/ontologies/colorectal-domain.owl`

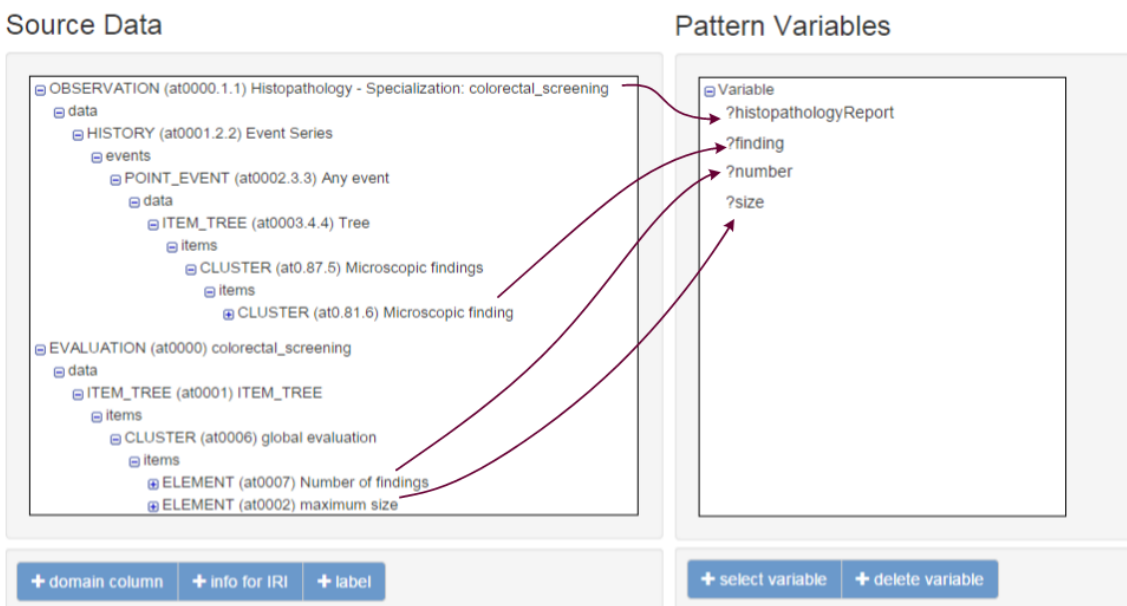

Figure 11: **Mapping between the two archetypes**

risk of developing colorectal cancer has the same definition in both European and American guidelines: A patient has low risk if the histopathology report contains less than three adenomas, all of them are normal and their sizes are lower than 10mm. Next, its implementation in OWL-DL is shown:

```
HistopathologyReport
 and (hasFinding only NormalAdenoma)
 and (max_size some integer[< 10]) and (number some integer[< 3])
```

These rules were implemented in a classification ontology which consists of a set of defined classes that formalise the patients classification rules and which import the domain ontology[28]. Both ontologies are the result of our previous work in Fernández-Breis et al. (2013). The application of the mapping to the patient data in Table 1 results in the data shown in Figure 12. This Figure corresponds to the OWL representation and classification of a histopathology report as viewed in Protégé[29]. The histopathology report has

[28]http://miuras.inf.um.es/ontologies/colorectalscreening-rules.owl
[29]http://protege.stanford.edu/

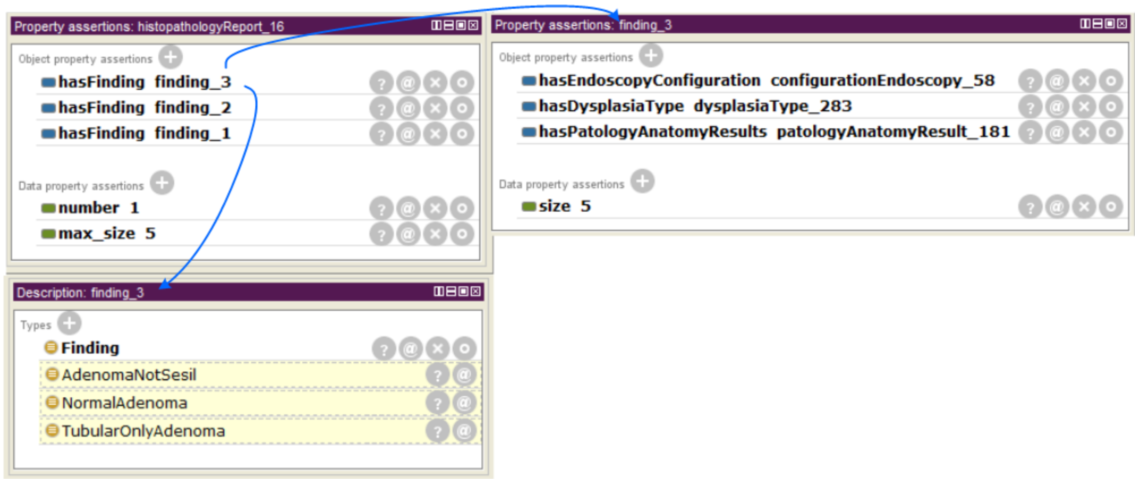

Figure 12: **OWL representation of the findings for patient in Table 1**

three findings, but only one is an adenoma (finding 3), and the largest adenoma has size 5. The upper-left side shows that the report contains three findings, one adenoma and the max size of the adenomas of the report is 5. The right side expands the properties of finding_3. The classification of a finding depends its properties, so the rule for classifying findings is defined in the domain ontology. We can see the properties hasEndoscopyConfiguration, hasDysplasiaType, hasPathologyAnatomyResults. and the size of the finding. These properties are used in order to classify the finding as a normal adenoma because it has low degree of dysplasia (value 283) it is not sessile and it is a tubular adenoma (see the types of finding_3 in the lower part of the figure). Therefore, the patient is classified as low risk according to the American and European guidelines.

5. Discussion and Conclusions

The achievement of semantic interoperability of EHR systems should be facilitated by the existence of appropriate tools for managing EHR-related information and knowledge. In this paper, we have presented a semantic web-based, integrated solution for managing archetypes and EHR extracts. ArchMS is an evolution of the functionality offered by the ResearchEHR platform described in Maldonado et al. (2012), offering a set of tools for semantic enrichment, standardization and interoperability of clinical data and archetypes. ArchMS adds new methods for supporting semantic interoperability, such as similarity-based searching methods and new querying methods. In addition to this, ArchMS includes a range of functions associated with the semantic representation and exploitation of clinical data (e.g., classification of patients, recommendation of learning contents).

ArchMS is a prototypical tool, and in this paper we have explained how it can be applied to support real studies. The evaluation of some individual modules included in ArchMS have been reported in previous papers. Performing a complete evaluation would require give access to real patients and health professionals, which is difficult taking into account the limited implantation of dual model EHR architectures in legacy systems to date. However, our case study has shown some of the great benefits of using semantic technologies in biomedical research: (1) we have been able to represent patient data, annotations about the archetypes and classification rules in the same formalism, which has permitted a joint exploitation by means of automated reasoning; (2) we have been able to reuse and exploit the content from existing archetypes and ontologies; (3) the semantic content generated and managed in ArchMS can also be reused by third parties because ArchMS follows the Semantic Web principles.

5.1. Semantic web infrastructure

ArchMS makes use of ontologies in different ways: controlled vocabulary, knowledge schema, consistent search, classifying instances, reuse and inferencing, all these uses being among the major applications of ontologies according to Stevens and Lord (2009). One major use of ontologies in biomedical domains is annotation, the Gene Ontology (see Ashburner et al. (2000)) being the most important one. In this use, ontologies are exploited as controlled vocabularies, since the ontology classes are mainly the annotation entities. Archetype terminology bindings should not be confused with the annotations provided by our system. These are usually added to the archetype data elements or terms during its building, and ArchMS does not intend to support the design and development of archetypes. The annotations provided by our system should be understood as archetype metadata since they are associated with the archetype as a whole and not with their individual terms. ArchMS is able to suggest archetype annotations in two different ways: (1) textual search; (2) archetype similarity. The textual description of the archetype is processed and issued against the BioPortal annotation recommendation service. Despite this approach has been helpful for our research projects, there are some specific archetype annotation automatic methodslike the ones presented in Yu et al. (2012); Qamar and Rector (2006), whose integration into ArchMS should be studied. The ontological infrastructure used for representing and annotating archetypes support the execution of consistent search processes, since such ontologies share both the knowledge schema and the semantic context for performing semantic search. The archetype search methods combine the semantic representation of both archetypes and annotations, which are jointly exploited for retrieving those archetypes that meet the query constraints.

The annotations suggested by ArchMS are based on the semantic similarity of archetypes calculated by applying state of the art semantic similarity functions. Such measurement is an example of the use of ontologies as domain schema since the classes and properties from the information and archetype model ontologies (the ontologies used by the PoseacleConverter) are used for the calculation. However, this measurement does not require using automated reasoning. The semantic similarity functions can be customised by the users by specifying the values to the threshold and the weights. There is no standard or automatic way to determine the best values for the weights, so an analysis has been carried out in order to suggest their potentially best range values. Additional research should be made to learn optimal sets of parameters depending on the properties of the archetypes compared and the size of the ontologies used in the annotations of the archetypes. Generally speaking, a higher value of a weight means that we are providing more importance to that factor among, on the one hand, distance vs properties vs linguistic similarity and, on the other hand, profile vs structural similarity. We consider that the weight for the linguistic similarity should be the smallest one because it does not really provide information about the particular structure or meaning of the knowledge entity. For the rest of parameters, local decisions should be made due to the local nature of their meaning. Provided that we are comparing classes in ontologies, the taxonomic distance should be considered the most important. This mechanism based on weights and thresholds permits each group of ArchMS users to obtain results adjusted to their notion of similarity.

To the best of our knowledge, OWL is not currently being exploited by archetype tools. ArchMS uses two representations in OWL for archetypes given the different purpose of the tasks and for which the representations have demonstrated to be effective. It should be noted that we are not proposing any of our OWL archetypes representations as standard ones, but they constitute appropriate technological decisions for the different semantic activities performed in our system. In addition to this, it should be noted that our work does not propose to replace ADL by OWL for any task, but to use the most appropriate formalism for each task, trying to minimize the implementation effort while maximizing the results obtained and the reuse of existing semantic resources and frameworks. In summary, we pursue leveraging archetype and ontology technologies.

It is commonly said that the Semantic Web provides a natural space for the integration and exploitation of biomedical data (see, for instance, Goble and Stevens (2008)). Among current Semantic Web initiatives, Linked Open Data effort[30] pursues the publication and sharing of biomedical datasets using semantic formats. Berners-Lee[31] suggested a five-star deployment scheme for Open Data, and the upper levels can be reached through semantic web technologies. ArchMS follows the Open Data paradigm and, depending on how the datasets are used, it could use Linked Open Datasets. The methods provided by SWIT permit to achieve four stars data repositories that are exploited in ArchMS. The fifth star can be achieved by getting your dataset linked from an external one, which requires the repository used in ArchMS to be accessible by third parties through a SPARQL Endpoint, which would be a decision to be made by the system administrators.

5.2. Use of automated reasoning

Reasoning with ontologies is also exploited in ArchMS with both archetypes and data. On the archetypes side, ArchMS checks the correctness of archetypes including specialization by applying automatic reasoning over the OWL representation of archetypes. On the data side, automated reasoning is used for the classification of patients. Such activity is performed over the patient data imported into ArchMS as XML extracts and transformed into RDF/OWL using the SWIT methods. As it has been mentioned, such transformation is driven by domain ontologies, which play the role of knowledge schema in such transformation and enhanced with the use of semantic patterns. It should be noted that SWIT accepts any semantic pattern that can be expressed using OPPL2 grammar, although in the Section "Clinical Data in OWL" section we emphasized the use of the semantic content patterns proposed by SHN

[30] `http://linkeddata.org/`
[31] `http://www.w3.org/DesignIssues/LinkedData.html`

and its ontological framework, since they provided a formal representation of clinical data and they are intended as a solution for achieving the semantic interoperability of clinical information, since they allow the homogeneous query of isosemantic clinical information as we shown in Legaz-García et al. (2014). Once the EHR extracts are transformed into instances, inferencing is used for classifying such instances. For instance, in our colorectal cancer screening effort, the patient data were classified by level of risk according to the European and American protocols. This data transformation permits to move from the archetype technological space to the semantic web one. Currently, our transformed data do not keep information about the structure of the archetypes, since the transformation is purely driven by the domain ontology. In the future, we expect to also transform the structure of the archetypes to investigate which transformation approach can be more appropriate for different tasks, as we have done with the different OWL representations for archetypes. Nevertheless, OWL features like the open world assumption need to be taken into account to represent exclusion criteria. It should be noted that, if needed, some OWL-DL reasoners provide versions that work in a closed world.

5.3. Secondary use of EHR data

Another distinctive feature of ArchMS is that it includes secondary uses of EHR data. In this manuscript, we have described patient classification that exploits the complete EHR data by means of semantic web technologies. This use requires the availability of specific classification rules, which must be implemented in an OWL ontology by means of *equivalentClass* axioms. This is one of the possible ways for representing rules using semantic web technologies and has its limitations. Such rules can only contain constraints about the properties of the patient we are classifying. In case of needing to relate different patients, rule languages like SWRL[32] could be used, or the constraints could be expressed as SPARQL queries. Each alternative may have its own limitations. ArchMS has no support for SWRL, but SPARQL queries can be issued against the semantic repository in which the data are stored. Ideally, such classification rules should be extracted from clinical protocols or guidelines.

It may happen that the classification ontology had not been built by

[32] `http://www.w3.org/Submission/SWRL/`

reusing the domain ontology used for transforming the EHR data into OWL. In such case, mappings between the domain ontology used for the transformation and the domain ontology used for the classification ontology should be made explicit. For this purpose, *subClassOf* or *equivalentClass* axioms should be used, because they can be exploited by reasoners to infer the corresponding classifications. It should be noted that the design of the ontologies might make difficult, and even sometimes not possible, the definition of appropriate mappings between the domain and the classification ontologies. For such cases, ontology alignment tools and approaches might be helpful.

The integration of clinical guidelines and electronic healthcare records is in the agenda of major semantic interoperability initiatives like SemanticHealthNet[33]. Recently, the openEHR community has produced and applied the Guideline Definition Language (GDL), which exploits guidelines based on the openEHR specification (see Anani et al. (2014)). Further research on using GDL content in ArchMS will be carried out.

The semantic profile represents the semantic interpretation of the EHR data. Basically, it constitutes an abstraction from the EHR data to the semantic categories associated with such data. By semantic categories we mean concepts/classes in the terminologies and ontologies used for annotating the archetype and classes included in the classification ontologies applied to such data. All the semantic information available in ArchMS about the EHR data is used to build the profile. Consequently, connecting ArchMS with external semantic sources could permit to enhance the construction of such profile. By external source we mean in this context Linked Data, which has been previously mentioned in this discussion. Currently, Linked Open Data directories like datahub.io include more than 500 linked datasets related to health. The increasing awareness of the possibilities offered by such formats will certainly generate a higher number of datasets within the next years, which makes it a corpus worthy of study and exploitation.

5.4. Comparison with related tools

The use of a semantic web infrastructure is likely to be the major novelty of ArchMS over state of the art systems like the openEHR CKM, whose goals are different. CKM is based on ADL technology and is oriented to support to the construction and publication of existing archetypes. To the best of our

[33] `http://www.semantichealthnet.eu/`

knowledge, recent advances in CKM have concerned the improvement of the user visualization of archetypes and have not addressed a technological evolution towards the semantic web. One reason for this is that the specifications of the archetype model have not been designed having in mind the Semantic Web. An example is the fact that archetypes do not have URIs, which are the identifiers of resources in the Semantic Web. In ArchMS, we generate a URI for each archetype when represented and exploited in RDF/OWL. Another difference with CKM is that ArchMS stores both archetypes and data extracts. However, the key advantage of ArchMS against systems like CKM is the use of OWL technologies, which allow for the combination of information model, clinical models and terminologies. ArchMS does not provide, so far, functions for the edition of the archetypes, so the comparison with other openEHR tools like the Archetype Editor or the ADL Workbench are not relevant at this point. Despite terminology bindings can be defined in the ADL Workbench, these are different from the annotations created in ArchMS, which are not done at the ADL level, since our goal is not the authoring of archetypes.

5.5. Further standardization actions

ArchMS works only with ISO 13606 and openEHR content, since they use archetypes. CIMI recently decided to use ADL as representation formalism[34], and have started to create archetypes. This has also generated interest in the CEM community in transforming their models into archetypes. In fact, our PoseacleConverter includes the possibility of transforming CEM models into openEHR archetypes[35]. This permits an indirect semantic exploitation of CEM models using the ArchMS services. We plan as further work to be able to manage CIMI archetypes and CEM models in ArchMS.

As further work, we aim at adapting ArchMS to meet the ISO/IEC 11179 international standard for representing metadata for an organization in a metadata registry. We have already performed an initial mapping of the ArchMS entities with the ones of the standard. The availability of an OWL ontology for such standard will contribute to simplify the effort. In Sinaci and Laleci Erturkmen (2013) this standard in the Semantic Metadata Registry Repository (Semantic MDR) which is not focused on archetypes but on com-

[34] `http://informatics.mayo.edu/CIMI/index.php/London_2011`
[35] `http://miuras.inf.um.es/PoseacleConverter/`

mon data elements. For instance, the IMI EHR4CR project[36], which aims to improve healthcare research by making more efficient the access of academia and industry to EHR data and the participation of hospitals in clinical trials programs, proposes in its workpackage 4, a semantic interoperability framework for the correct share of clinical data between healthcare providers and clinical researchers, using a conceptual reference model (EHR4CR information model) implemented through the use of a meta data repository[37].

Our framework is able to deal with external resources in OWL format for the annotation of the clinical models and data. Investigating the integration content from terminology servers like LexEVS[38] or NCI CDE[39] would permit to use traditionally major biomedical terminological sources if not available in OWL.

5.6. Conclusion

We have presented ArchMS, which combines management and interoperability services previously developed by our group and new functions, among which the semantic transformation and exploitation of data can be pointed out. Our results show the potential of semantic web technologies for the management and exploitation of archetypes and EHR data, and we think that our approach could be applied to other dual model standards. Further work will focus on integrating new standards and improving the transformation and recommendation methods.

6. Acknowledgments

We thank the Programa de Prevención del Cáncer de Colon y Recto de la Región de Murcia for providing the data for performing the use case. This work has been funded by the Spanish Ministry of Science and Innovation and the FEDER programme through grants TIN2010-21388-C02-02, TIN2014-53749-C2-2-R, the Fundación Séneca through grants 15555/F-PI/2010 (MCLG) and 15295/PI/10.

[36]`http://www.ehr4cr.eu/`

[37]`http://www.ehr4cr.eu/files/ExecutiveSummary/EHR4CR-ExecutiveSummaryD4_1.pdf`

[38]`https://wiki.nci.nih.gov/display/LexEVS/LexEVS`

[39]`http://cdebrowser.nci.nih.gov/`

References

Abhyankar, S., Demner-Fushman, D., McDonald, C.J., 2012. Standardizing clinical laboratory data for secondary use. Journal of Biomedical Informatics 45, 642–50.

Anani, N., Chen, R., Prazeres Moreira, T., Koch, S., 2014. Retrospective checking of compliance with practice guidelines for acute stroke care: a novel experiment using openehr's guideline definition language. BMC Medical Informatics and Decision Making 14, 39.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al., 2000. Gene Ontology: tool for the unification of biology. Nature genetics 25, 25–29.

Atkin, W., Valori, R., Kuipers, E., Hoff, G., Senore, C., Segnan, N., Jover, R., Schmiegel, W., Lambert, R., Pox, C., 2012. European guidelines for quality assurance in colorectal cancer screening and diagnosis. Endoscopy 10, 0032–1309821.

Auer, S., Dietzold, S., Lehmann, J., Hellmann, S., Aumueller, D., 2009. Triplify: light-weight linked data publication from relational databases, in: Proceedings of the 18th international conference on World Wide Web, ACM. pp. 621–630.

Berners-Lee, T., Hendler, J., Lassila, O., et al., 2001. The semantic web. Scientific American 284, 28–37.

Bhatt, M., Rahayu, W., Soni, S.P., Wouters, C., 2009. Ontology driven semantic profiling and retrieval in medical information systems. Web Semantics: Science, Services and Agents on the World Wide Web 7, 317–331.

Bizer, C., Cyganiak, R., 2006. D2r server-publishing relational databases on the semantic web, in: Poster at the 5th International Semantic Web Conference, pp. 294–309.

Bodenreider, O., 2008. Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearbook of Medical Informatics , 67–79.

Costa, C.M., Menárguez-Tortosa, M., Fernández-Breis, J.T., 2011. Clinical data interoperability based on archetype transformation. Journal of Biomedical Informatics 44, 869–80.

Danciu, I., Cowan, J.D., Basford, M., Wang, X., Saip, A., Osgood, S., Shirey-Rice, J., Kirby, J., Harris, P.A., 2014. Secondary use of clinical data: The Vanderbilt approach. Journal of Biomedical Informatics 52, 28–35.

d'Aquin, M., Motta, E., 2011. Watson, more than a semantic web search engine. Semantic Web 2, 55–63.

Falbo, R.A., Guizzardi, G., Gangemi, A., Presutti, V., 2013. Ontology Patterns: Clarifying Concepts and Terminology, in: Workshop on Ontology and Semantic Web Patterns, Sydney, Australia. pp. 1–13.

Fernández-Breis, J.T., Maldonado, J.A., Marcos, M., Legaz-García, M.d.C., Moner, D., Torres-Sospedra, J., Esteban-Gil, A., Martínez-Salvador, B., Robles, M., 2013. Leveraging electronic healthcare record standards and semantic web technologies for the identification of patient cohorts. Journal of the American Medical Informatics Association 20, e288–96.

Goble, C., Stevens, R., 2008. State of the nation in data integration for bioinformatics. Journal of Biomedical Informatics 41, 687–93.

Gruber, T.R., 1993. A translation approach to portable ontology specifications. Knowledge Acquisition 5, 199–220.

Iqbal, A.M., 2011. An OWL-DL ontology for the HL7 reference information model, in: Toward Useful Services for Elderly and People with Disabilities. Springer, pp. 168–175.

Kalra, D., Lewalle, P., Rector, A., Rodrigues, J.M., Stroetmann, K.A., Surjan, G., Ustun, B., Virtanen, M., Zanstra, P.E., 2009. Semantic interoperability for better health and safer healthcare. Research and Deployment Roadmap for Europe. SemanticHEALTH Project Report (January 2009), Published by the European Commission, http://ec. europa. eu/information_society/ehealth .

Kiryakov, A., Popov, B., Terziev, I., Manov, D., Ognyanoff, D., 2004. Semantic annotation, indexing, and retrieval. Web Semantics: Science, Services and Agents on the World Wide Web 2, 49–79.

Legaz-García, M.d.C., Martínez-Costa, C., Miñarro-Giménez, J.A., Fernández-Breis, J.T., Schulz, S., Menárguez-Tortosa, M., 2014. Ontology patterns-based transformation of clinical information. Studies in health technology and informatics 205, 1018–22.

Levenshtein, V.I., 1966. Binary codes capable of correcting deletions, insertions and reversals, in: Soviet physics doklady, p. 707.

Lord, P.W., Stevens, R.D., Brass, A., Goble, C.A., 2003. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 19, 1275–1283.

Maldonado, J.A., Costa, C.M., Moner, D., Menárguez-Tortosa, M., Boscá, D., Miñarro Giménez, J.A., Fernández-Breis, J.T., Robles, M., 2012. Using the ResearchEHR platform to facilitate the practical application of the EHR standards. Journal of Biomedical Informatics 45, 746–62.

Maree, M., Belkhatir, M., 2015. Addressing semantic heterogeneity through multiple knowledge base assisted merging of domain-specific ontologies. Knowledge-Based Systems 73, 199–211.

Martínez-Costa, C., Menárguez-Tortosa, M., Fernández-Breis, J.T., 2010a. An approach for the semantic interoperability of ISO EN 13606 and OpenEHR archetypes. Journal of Biomedical Informatics 43, 736–46.

Martínez-Costa, C., Menárguez-Tortosa, M., Fernández-Breis, J.T., Maldonado, J.A., 2009. A model-driven approach for representing clinical archetypes for semantic web environments. Journal of Biomedical Informatics 42, 150–64.

Martínez-Costa, C., Miñarro-Giménez, J.A., Menárguez-Tortosa, M., Valencia-García, R., Fernández-Breis, J.T., 2010b. Flexible semantic querying of clinical archetypes, in: Knowledge-Based and Intelligent Information and Engineering Systems. Springer, pp. 597–606.

Martínez-Costa, C., Schulz, S., 2014. Ontology content patterns as bridge for the semantic representation of clinical information. Applied Clinical Informatics 5, 660–9.

Menárguez-Tortosa, M., Fernández-Breis, J.T., 2013. OWL-based reasoning methods for validating archetypes. Journal of Biomedical Informatics 46, 304–17.

Menárguez-Tortosa, M., Martínez-Costa, C., Fernández-Breis, J.T., 2012. A generative tool for building health applications driven by ISO 13606 archetypes. Journal of Medical Systems 36, 3063–75.

Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D.L., Storey, M.A., Chute, C.G., et al., 2009. Bioportal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Research 37, W170–W173.

Oren, E., Möller, K., Scerri, S., Handschuh, S., Sintek, M., 2006. What are semantic annotations. Technical report. DERI Galway .

Pesquita, C., Faria, D., Falcao, A.O., Lord, P., Couto, F.M., 2009. Semantic similarity in biomedical ontologies. PLOS Computational Biology 5, e1000443.

Qamar, R., Rector, A., 2006. Most: A system to semantically map clinical model data to snomed-ct, in: In the proceedings of Semantic Mining Conference on SNOMED-CT, pp. 38–43.

Rada, R., Mili, H., Bicknell, E., Blettner, M., 1989. Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics 19, 17–30.

Rea, S., Pathak, J., Savova, G., Oniki, T.A., Westberg, L., Beebe, C.E., Tao, C., Parker, C.G., Haug, P.J., Huff, S.M., Chute, C.G., 2012. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project. Journal of Biomedical Informatics 45, 763–71.

Rector, A., Brandt, S., Drummond, N., Horridge, M., Pulestin, C., Stevens, R., 2012. Engineering use cases for modular development of ontologies in OWL. Applied Ontology 7, 113–132.

Resnik, P., 2011. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. arXiv preprint arXiv:1105.5444 .

Rodrıguez-Muro, M., Hardi, J., Calvanese, D., 2012. Quest: efficient sparql-to-sql for rdf and owl, in: 11th International Semantic Web Conference ISWC 2012, Citeseer. p. 53.

Saleem, J.J., Flanagan, M.E., Wilck, N.R., Demetriades, J., Doebbeling, B.N., 2013. The next-generation electronic health record: perspectives of key leaders from the US Department of Veterans Affairs. Journal of the American Medical Informatics Association 20, e175–7.

Sánchez-Vera, M.d.M., Fernández-Breis, J.T., Castellanos-Nieves, D., Frutos-Morales, F., Prendes-Espinosa, M.P., 2012. Semantic web technologies for generating feedback in online assessment environments. Knowledge-Based Systems 33, 152–165.

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

Schulz, S., Martínez-Costa, C., 2013. How Ontologies Can Improve Semantic Interoperability in Health Care, in: Process Support and Knowledge Representation in Health Care. Springer, pp. 1–10.

Shearer, R., Motik, B., Horrocks, I., 2008. HermiT: A Highly-Efficient OWL Reasoner, in: OWLED.

Sinaci, A.A., Laleci Erturkmen, G.B., 2013. A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains. Journal of Biomedical Informatics 46, 784–94.

Stevens, R., Lord, P., 2009. Application of ontologies in bioinformatics, in: Handbook on Ontologies. Springer, pp. 735–756.

Tao, C., Jiang, G., Oniki, T.A., Freimuth, R.R., Zhu, Q., Sharma, D., Pathak, J., Huff, S.M., Chute, C.G., 2013. A semantic-web oriented representation of the clinical element model for secondary use of electronic health records data. Journal of the American Medical Informatics Association 20, 554–62.

Tapuria, A., Kalra, D., Kobayashi, S., 2013. Contribution of Clinical Archetypes, and the Challenges, towards Achieving Semantic Interoperability for EHRs. Healthc Inform Res 19, 286–92.

Whetzel, P.L., Noy, N.F., Shah, N.H., Alexander, P.R., Nyulas, C., Tudorache, T., Musen, M.A., 2011. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Research 39, W541–W545.

Yu, S., Berry, D., Bisbal, J., 2012. Clinical coverage of an archetype repository over SNOMED-CT. Journal of Biomedical Informatics 45, 408–18.