Knowledge-Based Systems 105 (2016) 175–189

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A semantic web based framework for the interoperability and exploitation of clinical models and EHR data

María del Carmen Legaz-García[a], Catalina Martínez-Costa[b], Marcos Menárguez-Tortosa[a], Jesualdo Tomás Fernández-Breis[a,*]

[a] Departamento de Informática y Sistemas, Universidad de Murcia, IMIB-Arrixaca, Spain
[b] Institute of Medical Informatics, Statistics, and Documentation, Medical University of Graz, Austria

ARTICLE INFO

Article history:
Received 14 June 2015
Revised 8 May 2016
Accepted 9 May 2016
Available online 12 May 2016

Keywords:
Semantic web
Medical informatics
Electronic health records
Ontology
Semantic interoperability

ABSTRACT

The advent of electronic healthcare records (EHR) systems has triggered the need for their semantic interoperability, which is reinforced by the opportunities for the 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. In this paper we describe an OWL-based framework that leverages EHR and Semantic Web technologies for the interoperability and exploitation of archetypes, EHR data and ontologies. It also enables the secondary use of clinical data. This framework has been implemented in the Archetype Management System (ArchMS). We also describe how ArchMS has been used in a real study in the colorectal cancer domain.

1. Introduction

The increasing use of electronic health records (EHRs) in our globalised world leads to a situation in which patients' health data are spread across different health systems. This situation demands the semantic interoperability of clinical information, that is, its meaningful communication across EHR systems. The lack of such interoperability 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 [1].

In [2], the SemanticHEALTH project identified that EHR standards, ontologies and terminologies are key players to achieve the desired semantic interoperability. In the last decades, many efforts have addressed the development of EHR standards and specifications, including openEHR[1] or ISO 13606[2]. They 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, which are a promising way of sharing clinical data in a formal and scalable way [3]. The interest in archetypes is reinforced by the commitment of the Clinical Information Modeling Initiative (CIMI) to use them.[3] HL7 specifications[4] have also evolved to include artifacts similar to clinical models with the aim of facilitating sharing and interoperability. An example is the recent Fast Healthcare Interoperability Resources (FHIR) specification.[5]

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. Therefore, the development of tools that permit to exploit the archetypes and the archetype-based data is needed [2].

Besides, the advent of EHR systems has also created new opportunities for the secondary use of data such as rapid cohort identification, quality of care assessment, comparative effectiveness research, data privacy and de-/re-identification research, phenotyping methodology and predictive modelling [4]. Some secondary

* Corresponding author. Fax.: +34868884151.
E-mail address: jfernand@um.es (J.T. Fernández-Breis).

[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.

http://dx.doi.org/10.1016/j.knosys.2016.05.016
0950-7051/

uses require combining data from different systems, which requires semantic interoperability between such systems, and in works like [5,6] the corresponding solutions are based on standards.

The achievement of the semantic interoperability will depend on the effective development and application of technologies able of supporting tasks such as detecting semantically equivalent archetypes and EHR data, or the joint exploitation of clinical models, clinical data, terminologies and ontologies for both primary and secondary uses.

The Semantic Web [7] is described as a new form of Web content meaningful to computers, and [8] proposed the Semantic Web as a natural space for the integration and exploitation of biomedical data. Semantic Web technologies are meant to enable the joint exploitation of heterogeneous, distributed content because machines are able to understand the meaning, which is provided in a precise way by means of ontologies. An ontology is defined in [9] as a common, shareable and reusable view of a particular application domain. Besides, Semantic Web technologies permit to infer new information by using automated reasoning, which can be very useful when working in semantic interoperability settings, in which discovering relations between content generated by different systems will be needed.

Our hypothesis is that Semantic Web technologies provide an appropriate support for performing the previously described tasks in the area of semantic interoperability. By Semantic Web technologies we mean the formalisms and languages that permit the semantic representation, query and exploitation of information and knowledge. Hence, in this paper we propose a Semantic Web-based framework for the joint exploitation of clinical data, archetypes, ontologies and terminologies for semantic interoperability environments. This framework has been implemented in the Archetype Management System (ArchMS), whose technological infrastructure permits to manage archetypes and EHR data from different standards using Semantic Web technologies.

The contributions of this work are (1) the exploitation of patient data, archetypes and classification rules using Semantic Web formalisms; (2) the reuse of content from existing archetypes and ontologies for the management and exploitation of clinical models and EHR data; and (3) enabling reuse of the ArchMS content by third parties because of the application of Semantic Web representation principles. As a prototypical tool, we think that ArchMS represents a good example of how Semantic Web technologies can contribute to semantic interoperability environments.

The structure of the rest of the paper is described next. In Section 2, some background and description of the state of the art in archetypes and Semantic Web technologies are presented. Our Semantic Web framework is described in Section 3. The validation of the platform in a real scenario is described in Section 4. Finally, some discussion and conclusions are put forward in Sections 5 and 6.

2. Background

2.1. Archetypes technologies

Archetypes are used to specify clinical recording scenarios such as a laboratory test, a blood pressure measurement, a medication order, etc. An archetype 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. Archetypes are expressed in the Archetype Definition Language[6] (ADL), which structures the content in 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 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 the most recent version of ADL. For example, the openEHR blood pressure archetype 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.

In the last years, a series of tools have been developed by the archetype community. LinkEHR[7] and the tools developed by the openEHR community, like the Archetype Editor[8] (AE), ADL Workbench[9] (AW) and the Clinical Knowledge Manager[10] (CKM) are likely to be the most widely used ones. LinkEHR permits the edition of archetypes, the representation of legacy data using archetypes as described in [10], and the view of EHR extracts. 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. The state of the art on tooling in this community shows the following limitations from a semantic interoperability perspective: (1) data, archetypes and terminologies are not represented and exploited in the same formalism, what limits the effectiveness of these approaches; (2) they are based on ADL technologies, whose limitations to perform activities as detecting equivalent archetypes have been shown in works like [11], since it is not easy to perform or support automated reasoning on ADL-based content.

2.2. Semantic web technologies

The Web Ontology Language (OWL)[11] is the *de facto* standard for ontology implementation, and it enables the precise description of data meaning. The subset of OWL based on Description Logics (DL), namely, OWL DL, permits the use of DL reasoning, which in this context enables performing inference tasks over the clinical models and the clinical data. In recent years, different works based on Semantic Web technologies have provided preliminary results of the feasibility of our research hypothesis:

1. OWL representations of clinical information and clinical models from different EHR standards such as ISO 13606, openEHR, HL7 or Clinical Element Models (CEM)[12] have been proposed [11–13].
2. OWL representations of clinical information and clinical models have supported the transformation of clinical models and clinical data between different EHR standards in [14,15].
3. OWL reasoning has been used for validating and checking the consistency of clinical models in works like [16] (for checking the correctness of terminological bindings) or [17] (for checking the correctness of specialization relations between archetypes).
4. OWL reasoning has been used to support the transformation of clinical models between specifications [18].
5. OWL reasoning has been used for the detection of isosemantic content in heterogeneous EHR systems, that is content with the same meaning but structurally different [19].

[6] http://www.openehr.org/releases/trunk/architecture/am/adl2.pdf.

[7] http://www.linkehr.com/.

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

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

[10] http://www.openehr.org/ckm.

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

[12] http://www.clinicalelement.com.

6. Semantic Web technologies have been used to support secondary use of EHR data [13].

The analysis of the state of the art reveals that more efforts have been devoted to clinical models than to clinical data. On the models side, the proposals mainly use representations based on OWL classes. So far, reasoning has been used to check the correctness of clinical models to support their transformation and to detect equivalent data instances. However, these uses have been carried out by independent efforts. There are no experiences of the use of OWL for the representation and exploitation of clinical models, data and biomedical domain knowledge (i.e., ontologies and terminologies). Besides, hundreds of biomedical ontologies are available in OWL format in repositories as Bioportal, including many medical terminologies. This makes the joint exploitation of clinical data, models, ontologies and terminologies natural, in contrast to ADL technologies. In this work, we use OWL DL ontologies for representing clinical models, clinical data and biomedical domain knowledge, which will permit to make domain knowledge explicit and therefore exploitable by means of automated reasoning. The use of automated reasoning is another major advantage of Semantic Web technologies against ADL ones.

On the data side, 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 are limited in terms of semantic processing of information and do not provide support for automatic reasoning. In the last years, different methods for representing relational and XML content using Semantic Web formats like RDF and OWL have been proposed.

Methods for transforming relational databases propose to represent tables as ontology classes, records as individuals, and columns as properties. The W3C RDB2RDF[13] specification proposes a canonical transformation/mapping for relational databases to RDF. Such a transformation can be considered a change of format, because the real meaning of the entities represented is not used in such a process. Triplify [20] is an example of this type of tools, whose main limitation to support our work is that they perform a generic transformation of the data, that is, they do not take into account their underlying model of meaning. Our experience reveals that the semantic representation of the data sometimes needs additional content that is not made explicit in the XML schema or in the corresponding table, so the additional meaning cannot be provided by the canonical transformation. There are also methods and tools in the area of ontology-based data access that permit to use relational data with queries in Semantic Web query languages like SPARQL.[14] Examples of such tools are presented in [21,22]. However, these approaches would not permit the representation of the data in the same formalism as the clinical models and the biomedical domain knowledge, which would limit our exploitation tasks and would not permit achieving the objective of the representation in a common formalism.

3. The ArchMS framework

ArchMS[15] is a framework for the management and reuse of clinical archetypes and data by applying Semantic Web technologies. ArchMS activities can be classified by the content managed (i.e., archetypes or EHR data) and by the EHR data use (i.e., primary or secondary). The framework has been designed having in mind three types of users: (1) administrators, who are able to perform all kinds of activities described in this section; (2) patients, who have access to their clinical data and to additional knowledge based on their EHR profiles; and (3) physicians, who have access to the patients' medical histories and can get additional knowledge from them (e.g. their classification by colorectal cancer risk).

Fig. 1 shows an overview of the ArchMS architecture. The figure contains three layers which correspond to different types of entities and activities included in the framework. The bottom of the figure shows the inputs to the system. ArchMS works with ADL archetypes and with XML EHR data extracts. In particular, those archetypes and extracts must be compliant with the openEHR and ISO 13606 specifications. The mapping between archetypes and ontologies that will be used in data transformation and the associations between archetypes and ontologies that will be exploited in this framework for data classification are also inputs for ArchMS.

The layer *acquisition* shows the first activities that are performed when the archetypes and data are input into ArchMS. Archetypes are validated and converted into OWL (see Section 3.1.1) and data extracts can be directly input into ArchMS (see Section 3.2.1) and transformed into OWL (see Section 3.2.2). Once processed, the archetypes and data are stored.

The items in the *repositories* layer store the primary data about the clinical archetypes and the clinical extracts:

- A relational database (ADL Archetypes) that stores the non-semantic archetype content, such as archetype metadata properties (e.g. name, language, purpose, etc.) as well as the ADL file.
- A semantic repository (OWL Archetypes) that stores the OWL representation of the archetypes, allowing for the semantic exploitation of the content.
- An XML repository of clinical data extracts (XML EHR data).
- A semantic repository (OWL EHR data) that stores the data in OWL.
- The repository of associations between archetypes and ontologies, which are used for data classification. These associations are stored in an in-house defined format.

Other repositories in this layer store semantic resources:

- The set of ontology repositories contains (1) EHR ontologies and local ontologies uploaded by users; and (2) external resources, in this work, Bioportal ontologies. Such semantic resources are ontologies, terminologies and controlled vocabularies in OWL format.
- A semantic repository of annotations (Semantic Annotations) for both archetypes and clinical data.

The layer *exploitation* includes the services for the exploitation of the archetypes and the EHR data: archetype annotation (Section 3.1.2), archetype search and similarity (Section 3.1.4), data visualization (Section 3.2.1), semantic profiling (Sections 3.1.3 and 3.2.3) and data classification (Section 3.2.4). Data transformation (Section 3.2.2) is shown only on the acquisition layer for readability, but can be invoked at any moment. The generation of web applications shown in the figure is used with the purpose of data entry, so it will be described in Section 3.2.1. For legibility, Fig. 1 shows how some services enrich the semantic repository of annotations but does not show which repositories are used by the different services in the exploitation layer, this will be explained in the next subsections in which we provide further information about the methods included in the ArchMS framework.

3.1. Archetypes in ArchMS

The main activities that can be performed with archetypes are: conversion, validation, annotation, and searching for similar archetypes or for archetypes with specific properties. They are described next.

[13] http://www.w3.org/2001/sw/rdb2rdf/.
[14] http://www.w3.org/TR/rdf-sparql-query/.
[15] http://sele.inf.um.es/archms.

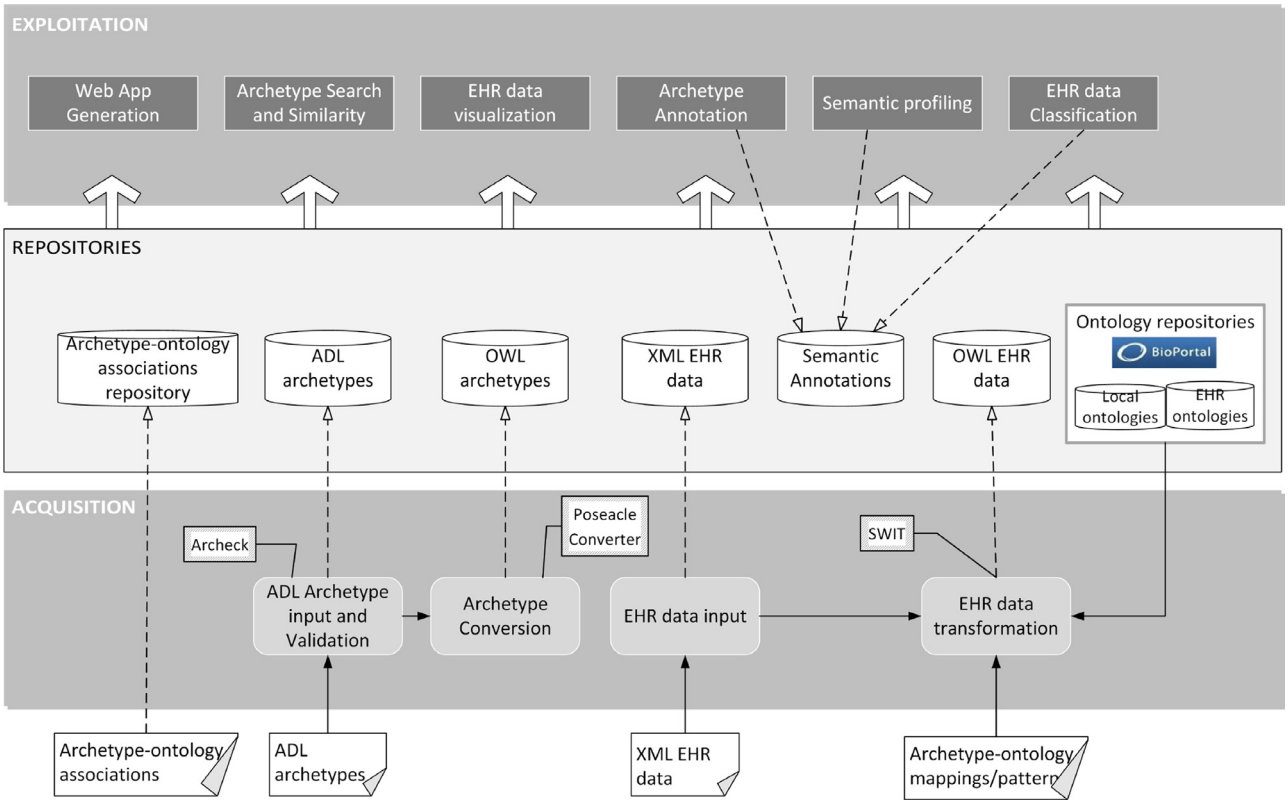

Fig. 1. ArchMS architecture overview, including three layers (from bottom to top): acquisition (light grey rounded-corner rectangles), repositories (white cylinders) and exploitation (grey rectangles).

3.1.1. Representation and management of archetypes

When an ADL archetype is imported into the system, the following activities are made: (1) validation, (2) storage in ADL, and (3) storage in OWL. The validation of the archetype is performed using our Archeck[16] method. This method, described in [17], represents archetypes as OWL classes and allows for checking the consistency of a specialised archetype with respect to its parent ones. ADL archetypes are transformed into OWL by applying the Ontology Definition Metamodel (ODM)[17] specification for expressing UML models in OWL. The import process fails if errors are found in the archetype. If the archetype is successfully validated, it is stored in the ADL Archetypes repository. The OWL version of the archetype is also stored in the OWL Archetypes repository.

Besides, the OWL individuals-based representation of the archetypes described in [11] is used in ArchMS for: (1) transforming archetypes and EHR data from openEHR into ISO 13606 and viceversa, and (2) calculating the semantic similarity between archetypes. This conversion is obtained by applying our PoseacleConverter.[18] It should be noted that OWL2 punning enables the use of the same URI for referring to an entity as individual or class depending on the context, which permits to use the different representations internally in a transparent way for the users. Hence, we store two OWL representations of the imported archetype.

3.1.2. Semantic annotation of archetypes

Archetypes are linked to terminologies by terminological bindings but additional semantic content may be needed for specific uses of the archetype. Given a specific repository of ontologies, controlled vocabularies and terminologies in OWL, 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. We are currently using Bioportal as main annotations resource, because it contains at the time of writing more than five hundred biomedical ontologies, terminologies and controlled vocabularies. This corpus of semantic resources is exploited through the Bioportal Web Services, presented in [23], which provide candidate terms for a given text. Besides, ArchMS uses the Apache Lucene API[19] to create an index using the content of the local semantic resources. This index is exploited using a search method based on the Apache Lucene API for finding exact and partial matches between the text and the content of the local semantic resources.

We represent the annotations in OWL format, which facilitates its jointly exploitation with the OWL representations of the archetype. These annotations and the terminology bindings are exploited in our approach in tasks such as archetype comparison and search. The annotations of the archetypes 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. Besides, EHR data are indirectly annotated through the annotations of the archetypes used to capture such data.

3.1.3. Semantic profile of an archetype

A semantic profile is defined in [24] as the semantic description of a dataset, that is, its semantic interpretation. Semantic profiles permit an efficient, effective processing without needing the use of the whole information about a particular information entity but such semantic interpretation. The use of ontologies for such pur-

[16] http://miuras.inf.um.es/archeck/.
[17] http://www.omg.org/spec/ODM/1.0/.
[18] http://miuras.inf.um.es/PoseacleConverter/.
[19] https://lucene.apache.org/core/.

pose permits to make decisions and recommendations based on formal specifications of domain knowledge. In our context, the annotations of an archetype are a representative generalization of its semantic content, and constitute its semantic profile.

The semantic profile of an archetype can be defined as the union set of all the terminological bindings of the archetype and its external annotations. The semantic profile of an archetype is generated when the archetype is imported into ArchMS, and it is updated when new annotations are added to the archetype.

3.1.4. Semantic similarity between archetypes

Studying the similarity between 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 a more complex process, and finding similarities may help to bridge the gap between different representations.

We use a method that performs the similarity of two given archetypes by comparing both the semantic and the structural content associated with them. Eq. 1 defines the function for calculating our semantic similarity score between two archetypes.

$$\begin{aligned}&\text{semantic\_similarity}(A_1, A_2)\\&= z_1 * \text{profile\_similarity}(A_1, A_2)\\&\quad + z_2 * \text{structural\_similarity}(A_1, A_2)\\&\quad \textit{where } \leq z_k \leq 1, \sum z_k = 1\end{aligned} \tag{1}$$

The factor *profile_similarity* depends on the similarity of the semantic profiles of the archetypes. The method for obtaining the similarity of the profiles (see Algorithm 1) uses the ontologies of the information and archetype models, and the ones associated with the annotations of the two archetypes as semantic context for the analysis. The result of the comparison is a number in the range [0,1].

Algorithm 1: Pseudo code of the semantic similarity function of the semantic profiles of two archetypes

```
Data:
A1, A2: Archetypes to compare
O: Set ofArchetype and Information Model Ontologies and Biomedical Ontologies
w1:weight given to taxonomic similarity metric
w2: weight given to propertiessimilarity metric w3: weight given to linguistic similarity metric
tr: threshold
1  PS1 ⟵ getSemanticProfile(A1);
2  PS2 ⟵ getSemanticProfile(A2);
3  for a_i ∈ PS1 do
4      for a_j ∈ PS2 do
5          pairwise_similarity[i][j] ⟵ 0;
6  for a_i ∈ PS1 do
7      for a_j ∈ PS2 do
8          d ⟵ taxonomicSimilarity(a_i, a_j, O);
9          ps ⟵ propertiesSimilarity(a_i, a_j, O);
10         ls ⟵ linguisticSimilarity(a_i, a_j, O);
11         score ⟵ w1 * d + w2 * ps + w3 * ls;
12         if score ≥ tr then
13             pairwise_similarity[i][j] ⟵ score;
14 for all i ∈ pairwise_similarity do
15     x ⟵ 0;
16     for all j ∈ pairwise_similarity[i] do
17         if x < pairwise_similarity[i][j] then
18             x ⟵ pairwise_similarity[i][j];
19     best_score[i] ⟵ x;
20 return mean(best_score);
```

This function compares all the pairs of elements in the semantic profiles of the archetypes, obtaining a similarity score for each pair (lines 8 to 11 in Algorithm 1). The process returns the similarity score as a result of executing the following steps:

- Comparing all the pairs and selecting those ones whose score is higher than a given threshold, namely, pairwise_similarity (lines 12 to 13 in Algorithm 1). If the value of the threshold is 1, only exact or equivalent matches are obtained. Threshold values below 1 permit to find elements which are similar enough but not equivalent.
- Finding the set of best pairs, namely, best_score, that maximises the mean of the similarity scores that include only one pair per element of the semantic profile of each archetype (lines 14 to 19 in Algorithm 1).
- Returning the mean of the similarity scores of those best pairs (line 20 in Algorithm 1).

The function that obtains the similarity between two elements of the semantic profile is based on the one described in [25], which uses the following factors:

- Taxonomic similarity (d): It measures the hierarchical distance between the classes associated with the two elements C_i and C_j. This function uses the union set of taxonomic ancestors of both classes (the taxonomic ancestors) and the intersection set of ancestors of both classes (common taxonomic ancestors). The difference in size of the set of all the ancestors and the common ones represents the length of the path between those two classes that goes through the common ancestor, that is, the shortest taxonomic path between the two classes. Certain classes might present multiple taxonomic ancestors in an ontology. In such cases, all the distances are calculated, and the shortest one is used. Thus, this score is calculated as shown in Eq. 2:

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

- Properties similarity (ps): It measures the similarity between the set of properties associated with the classes C_i and C_j. It takes into account how many properties are associated with both classes, namely, common(C_i, C_j), and how many properties are associated only with one class, namely, different(C_i, C_j). Thus, this score is calculated as shown in Eq. 3. Note that different(C_i, C_j) and different(C_j, C_i) may have different values because both operations can be seen as the difference between the sets of properties associated with each class. Let us suppose that C_i has 7 properties, that C_j has 4 and that three of them are common to both classes, then different(C_i, C_j) would have value 4 and different(C_j, C_i)) would have value 1.

$$\begin{aligned}&ps(C_i, C_j)\\&= \frac{|\text{common}(C_i, C_j)|}{|\text{common}(C_i, C_j)| + 1/2 * |\text{different}(C_i, C_j)| + 1/2 * |\text{different}(C_j, C_i)|}\end{aligned} \tag{3}$$

- Linguistic similarity (ls): A string-based calculation of the terms associated with the ontological elements compared. If we are

comparing two classes from the OWL representation of two archetypes, this calculation uses their term definitions. When comparing two concepts from a terminology, the labels or the local names are used. Our current implementation uses the distance defined in [26].

These three factors are combined to obtain the similarity between two elements of the semantic profile of the archetypes as shown in lines 11-13 in Algorithm 1, where $0 \leq w_k \leq 1, \sum w_k = 1$.

On the other hand, the factor *structural_similarity* depends on the types of the archetypes compared. The semantic context for this comparison is the information model ontology of the corresponding EHR standard. For instance, the structural similarity method assumes that two archetypes of the same type COMPOSITION are more similar than two archetypes of types COMPOSITION and SECTION. This score is obtained by applying the taxonomic similarity function to the types of both archetypes.

Our similarity functions include a similarity threshold and a series of weights. These parameters can be determined by the users of ArchMS, so enabling them to decide which notion of semantic similarity to apply. It should be noted that once an archetype is successfully imported, ArchMS looks for similar archetypes using the similarity threshold. Such search enables (1) checking if an equivalent archetype already exists in the repository; and (2) recommending annotations associated with similar archetypes.

3.1.5. Searching archetypes

The ArchMS framework includes the following methods for searching archetypes from the different repositories:

- The textual search interface uses a Lucene index for finding archetypes that contain the search text in their textual content, including metadata properties like name, language, purpose, etc. The query "histopathology English" returns the archetypes that contain "histopathology" or "English" in any text field.
- The faceted search 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" returns the archetypes available in the language "English" that contain "histopathology" as an annotation.
- 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 exploits the hierarchical structure of the ontologies, terminologies and controlled vocabularies, and it applies the semantic similarity function 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 for such concept, archetypes with annotations or bindings similar enough to "Histopathology" could be suggested. For this purpose we provide a specific semantic query language for archetypes based on SPARQL.

3.2. EHR data in ArchMS

Fig. 2 shows the ArchMS user interface for working with EHR data extracts. The upper part of the figure shows the options for an EHR extract: view the archetypes associated with the extract (*Archetypes*), view the XML extract (*Extract*), download the XML file (*Download*). The lower part of the figure lists the archetypes associated with the extract. Next, we describe the methods associated with these options. Besides, the patient data can also be classified (not shown in the figure), what requires obtaining the semantic profile of the patient.

3.2.1. Data import

ArchMS processes XML EHR extracts from both ISO 13606 and openEHR specifications. Such extracts may have two major sources:

1. XML extracts generated from third-party ISO 13606 or openEHR systems.
2. XML extracts generated using the ArchMS infrastructure. ArchMS permits to generate web forms from archetypes for data entry by applying the ArchForms method [27].

Once the extract is imported into ArchMS, the user can visualise both the content of the extract (Extract option in Fig. 2) or the archetypes used to capture the data (Archetypes option in Fig. 2). 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. In this case, the data are stored as one extract associated with multiple archetypes. This is the case of the example shown in Fig. 2, since the lower part of the figure lists the two archetypes associated with the extract "Histopathology report".

3.2.2. Data transformation

The ArchMS framework provides methods for the exploitation of EHR data using Semantic Web technologies. Given that the extracts are imported in XML, ArchMS uses a generic approach driven by domain knowledge to transform XML data into OWL.

The transformation process is driven by the mapping between the ADL archetypes used to capture the EHR data and domain ontologies. The mapping rules define how archetyped data are transformed into ontology individuals, and such rules must be defined by an expert. 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. Those mappings can be defined using our Semantic Web Integration Tool (SWIT)[20] and then uploaded into ArchMS. SWIT services are actually invoked from ArchMS to automatically execute the data transformation. The corresponding OWL content is stored in the OWL EHR data repository.

Fig. 3 provides a more detailed explanation of this process. The mapping rules are defined among the archetypes associated with the extracts and domain ontologies. We can also see that the transformation process uses ontology patterns [28]. Such patterns define the templates for the creation of OWL individuals with specific types of axioms and enable the definition of complex mapping rules. A pattern is created from formal descriptions of types of entities and their relations. Ontology patterns are expressed using the entities of the target domain ontology and may contain variables. The variables are mapped onto the archetype data elements.

Our current process uses the type of ontology patterns defined in the context of the European project SemanticHealthNet (SHN)[21], which pursues to ease the mapping process and to create semantically interoperable datasets in OWL. ArchMS patterns are currently implemented using the Ontology Pre-Processing Language version 2 (OPPL2).[22] 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.

[20] http://sele.inf.um.es/swit.
[21] http://www.semantichealthnet.eu/.
[22] http://oppl2.sourceforge.net/.

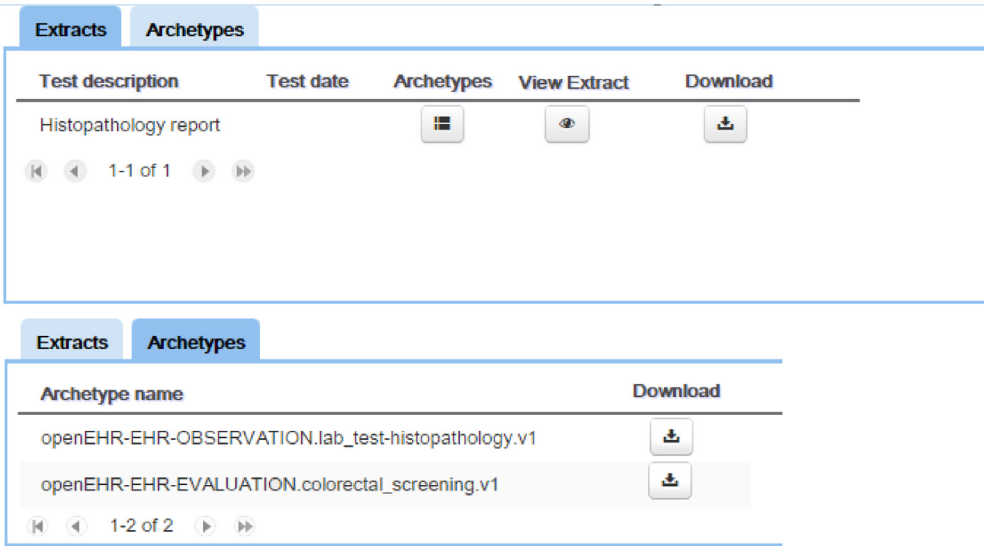

Fig. 2. Screenshot of the options available for data extracts and information of the archetypes used to capture the extract.

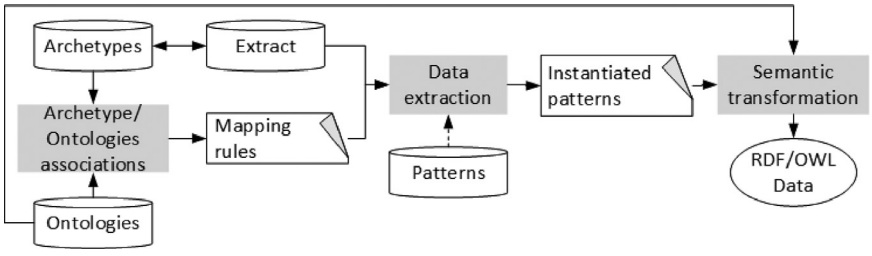

Fig. 3. Overview of the method for the transformation of archetype-based EHR data into RDF/OWL.

An excerpt of the SHN Medication Administration pattern in OPPL2 is shown next. This pattern defines that a medical record is a medication administration process whose patient is the pharmaceutical product. 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 pattern completes the rest by itself.

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

3.2.3. Semantic profiles of EHR data

An initial semantic profile of the EHR data is derived from its associated archetype (e.g. the semantic profile of an extract about blood pressure will be derived from the semantic profile of the blood pressure archetype). This profile can be enriched by analysing the EHR data. 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 the classification rules (see Section 3.2.4), which would permit to know when a patient has to be associated with "hypotension".

The semantic profile of an EHR data extract is generated when the extract is transformed into OWL, and it is updated when: (1) the profiles of its associated archetypes are updated; (2) new classifications of the patients become available (see Section 3.2.4).

3.2.4. Patient data classification

Patient classification means grouping patients in different categories conforming to certain clinical criteria. For example, in our case study in colorectal cancer, the patients can be classified by level of risk according to the rules based on the properties of the adenomas, the histopathology reports, etc.

The ArchMS framework uses OWL-DL reasoning to classify EHR data according to rules that define the clinical status of the patients. Two types of axioms associated with OWL-DL classes 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.

Ideally, the classification rules should be 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. An example of classification rule may define that a patient has hypotension if the systolic blood

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	Not sessile	Low degree	Tubular adenoma	5	Normal adenoma

pressure is lower than 90 mmHg or the diastolic ones is lower than 60 mmHg. This rule could be expressed into OWL-DL as follows:

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

Hence, patients with properties systolic value 80 or diastolic value 50 would be classified as *Hypotensive*. Given that the EHR data are associated with archetypes, ArchMS permits to associate classification ontologies with archetypes. This association means that such ontologies will be used every time EHR data captured with such archetype are classified. Once defined the associations, the classification method uses two inputs: (1) the OWL ontologies with the classification rules, and (2) the OWL representation of the patient EHR data. The classification rules are applied to the data by using DL reasoning. ArchMS stores the classifications as annotations of the EHR data, and they are included in the semantic profile of the patients. This approach is generic, and the same patient data can be automatically analysed and classified by applying many different classification ontologies.

4. Validation: ArchMS in a colorectal cancer study

In this section we describe how ArchMS has been used in a colorectal cancer study that included data from more than 20,000 anonymised patients' records from the 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 [29]) and American[23] colorectal cancer guidelines.

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 the colorectal cancer screening tests. For example, both guidelines define that patients with less than 3 normal adenomas with size lower 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, the third one being a Normal Adenoma with 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 [30]. Here, we focus on how the semantic activities offered by ArchMS were useful to carry out this study.

4.1. Selecting the archetypes

The source data for the study was a relational database of colorectal cancer screening tests. Such data were transformed into XML EHR extracts according to openEHR to be imported into ArchMS. This was done using LinkEHR, and required to use a series of archetypes for recording histopathology and colorectal screening information.

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 that match best (see Fig. 4). The query keywords were extracted from the original records: "histopathology", "finding", "size", "dysplasia" or "sessile". We found an archetype suitable for the histopathology report (i.e. openEHR-EHR-OBSERVATION.lab_test-histopathology.v1), but the repository did not contain anyone appropriate for colorectal screening.

Since our study required recording more information than the one provided by the histopathology archetype (e.g. information about the findings such as type, maximum size, dysplasia degree, sessile, etc.), we specialised the existing archetype (see left side of Fig. 5). Besides, we created an archetype for colorectal screening (see right side of Fig. 5) which allows for 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.

4.2. Importing and annotating the archetypes

The two mentioned archetypes are then imported into ArchMS. This implies checking the consistency of the imported archetypes regarding their 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 existing annotations. Two similar archetypes are found for openEHR-EHR-OBSERVATION.lab_test-histopathology-colorectal_screening (see Fig. 6), both of them related to it through specialization. The results are sorted by similarity score, which permits the user to know how similar the archetypes are. No similar archetypes are found for the archetype openEHR-EHR-EVALUATION.colorectal_screening. This is not an unexpected result since no archetype related to such domain was available in the repository.

In addition to the annotations retrieved from similar archetypes (see Fig. 6), we can add new annotations from SNOMED CT and MeSH. For that, ArchMS suggests related terms based on the archetype textual content in their keywords and ontology sections. Fig. 7 shows part of the recommendations for the openEHR-EHR-OBSERVATION.lab_test-histopathology-colorectal_screening.v1 archetype. The archetype contains terms such as "accession", "adenoma", "colorectal", etc. as keywords and in the ontology section. The system looks for matches for those terms in the selected annotation resource (MesH in this case) and shows the terms suggested. 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.

4.3. 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

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

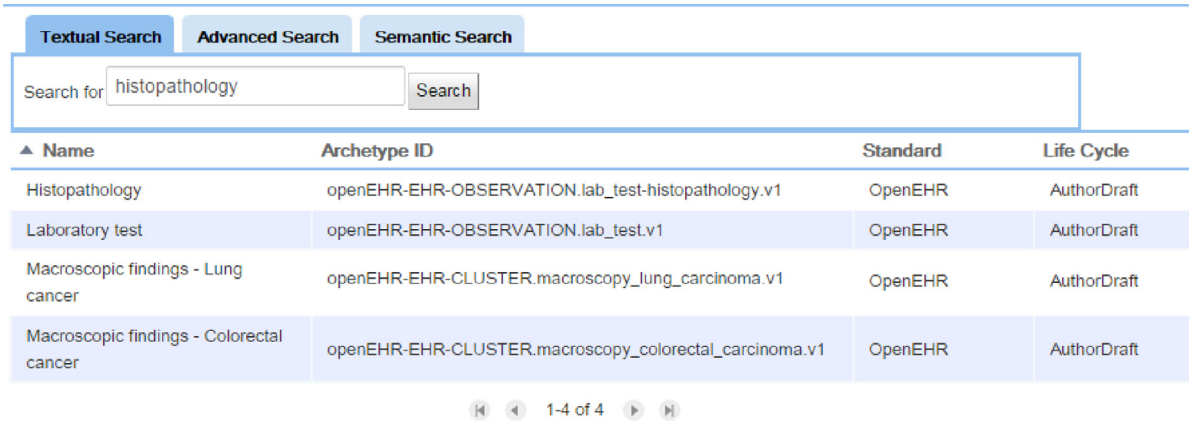

Fig. 4. Use of textual search with keyword histopathology.

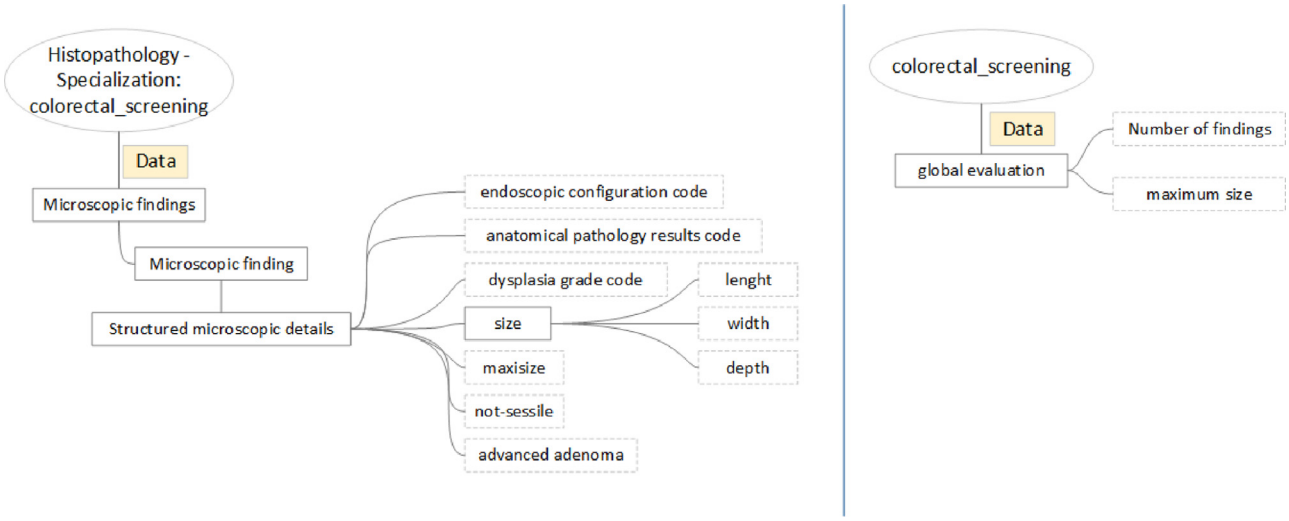

Fig. 5. (Left) Excerpt of Histopathology - Specialization colorectal screening (openEHR-EHR-OBSERVATION.lab_test-histopathology-colorectal_screening.v1); (Right) Excerpt of Colorectal Screening archetype (openEHR-EHR-EVALUATION.colorectal_screening.v1).

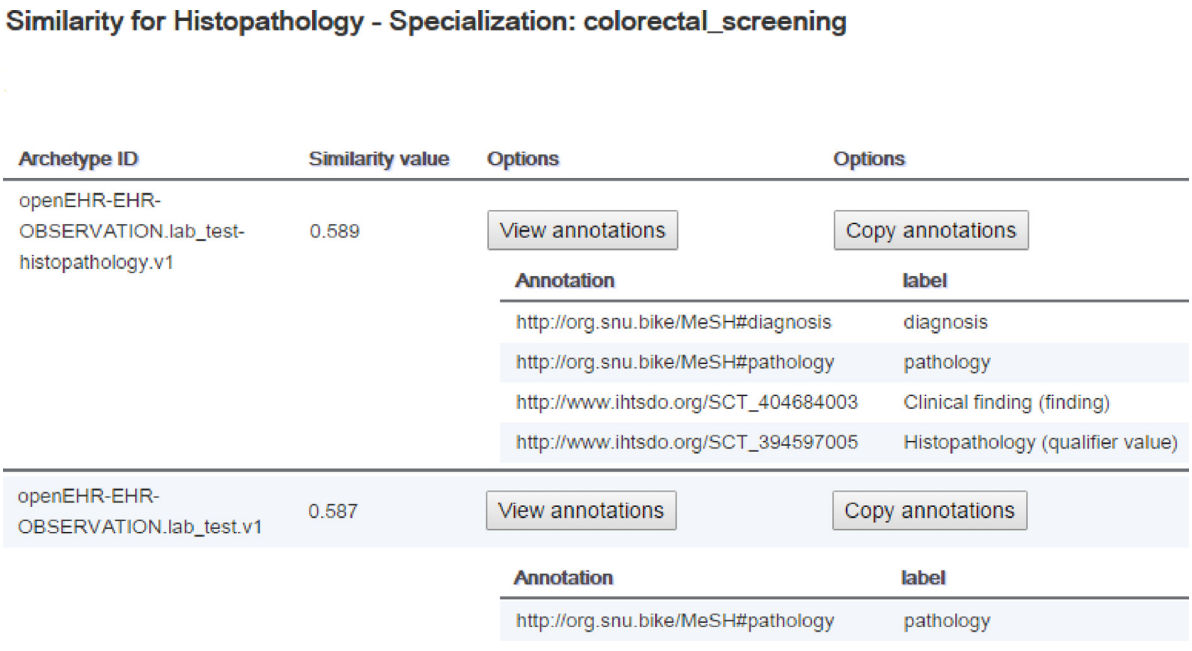

Fig. 6. Similar archetypes to openEHR-EHR-OBSERVATION.lab_test-histopathology-colorectal_screening.v1 and their annotations.

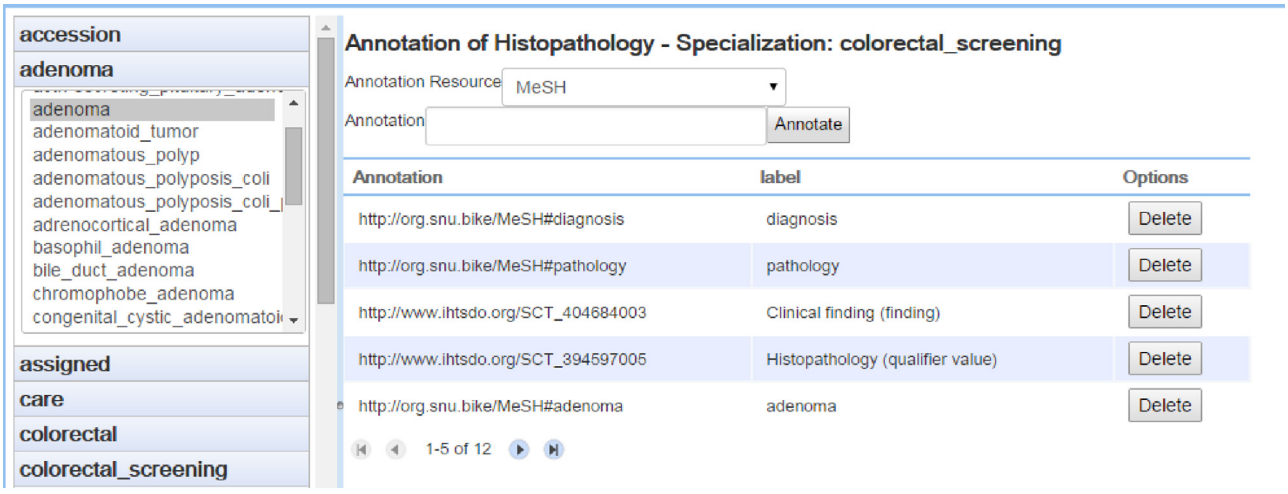

Fig. 7. Suggested annotations for openEHR-EHR-OBSERVATION.lab_test-histopathology-colorectal_screening.v1. The left part shows suggested annotations from the selected annotation resource (MeSH in the figure). The right part shows a table with the terms already used to annotate the archetype.

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

by using LinkEHR, and we used the SWIT services in ArchMS to define and execute the mappings between the archetypes and the ontology, and also to transform the data into OWL.

Fig. 8 shows a part of the mapping between the archetypes and the domain ontology for colorectal cancer.[24] In this case this pattern 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 has Finding ?finding,
ADD ?histopathologyReport number ?number,
ADD ?histopathologyReport maxsize ?size
END;
```

The pattern variables (preceded by ?) represent the parts that are bound 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. Each extract captured using openEHR-EHR-OBSERVATION.lab_test-histopathologycolorectal_screening.v1 archetype corresponds to a histopathology report, so we bind the root of the archetype to the *?histopathologyReport* variable. The "Microscopic finding" element represents the findings reported, so we bind that element to the variable *?finding*, whereas the values for the variables *?size* and *?number* are obtained from openEHR-EHR-EVALUATION.colorectal_screening.v1 (see Fig. 8).

4.4. 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 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 10 mm. 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 classification rules and which import the domain ontology.[25] The application of the mapping to the patient data in Table 1 results in the data shown in Fig. 9. This figure corresponds to the OWL representation and classification of a histopathology report as viewed in Protégé.[26] The upper-left side shows the histopathology report, which has three findings, but only one is an adenoma (*finding_3*),

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

[25] http://miuras.inf.um.es/ontologies/colorectalscreening-rules.owl.

[26] http://protege.stanford.edu/.

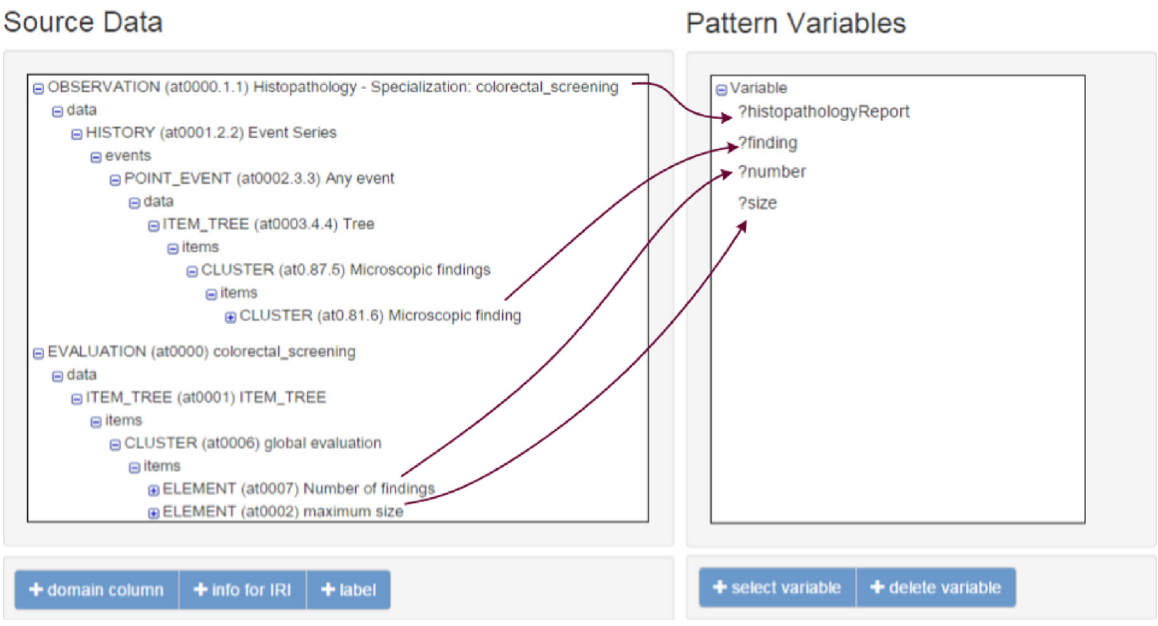

Fig. 8. Mapping between the two archetypes (left) and the variables of the pattern (right). The mappings will be applied to transform the EHR data extracts into RDF/OWL.

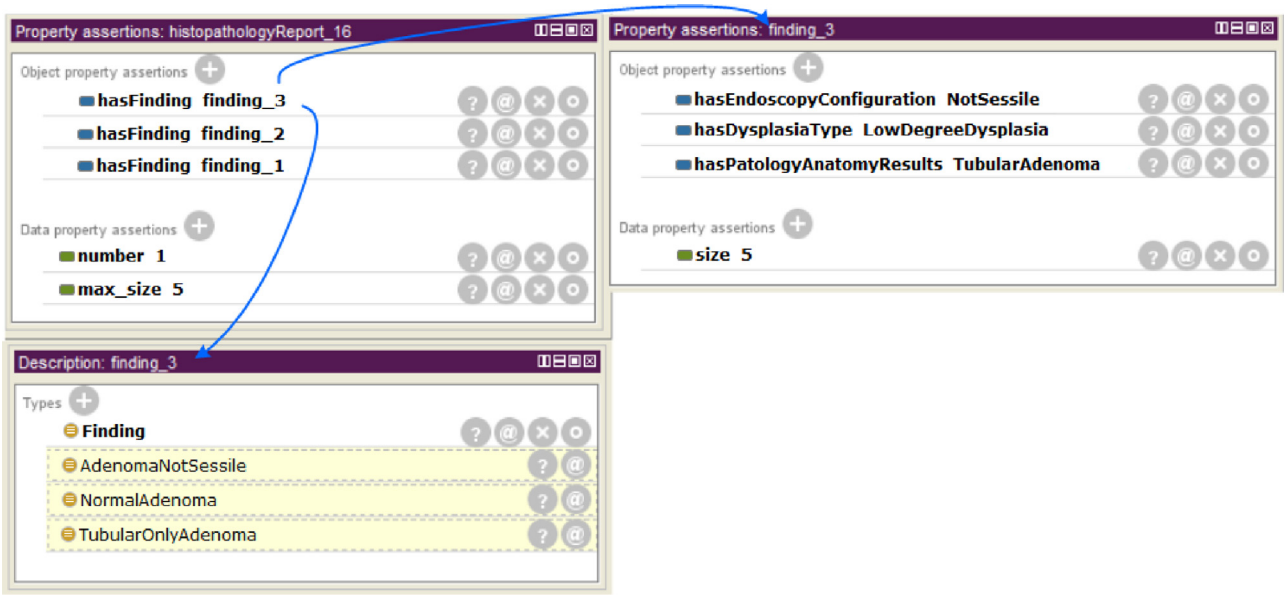

Fig. 9. OWL representation of the findings for patient in Table 1. The histopathology report of this patient contains three findings (upper left part) and the right part shows the properties of the finding_3. The bottom-left part shows the types of the adenoma finding_3.

and the largest adenoma has size 5 (*max_size*). The right side expands the properties of *finding_3*. The classification of a finding depends on 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 a low degree of dysplasia, it is not sessile and it is a tubular adenoma. Therefore, the patient is classified as low risk according to the American and European guidelines.

Once generated the complete dataset in OWL we applied automated reasoning to assign a level of risk to each patient and we evaluated the performance of the classification using a sample of 503 patients. The evaluation consisted in comparing the OWL classification with the level of risk recorded by the physicians in the database. The results showed 63% of agreement, which does not mean precision as it will be justified next. In 79% of the discrepancies the physicians had assigned a level of risk higher than the suggested by the OWL classification. The discrepancies were analysed with the physicians, who confirmed that the OWL classification was correct in all the cases, and that they were sometimes recording a level of risk different from the one suggested by the protocols due to restrictions related to the assignation of treatments of the patients, meaning that they are not always applying the protocol, which is also an interesting result of this effort.

5. Discussion

The achievement of the 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 ArchMS, which is a Semantic Web framework

for managing archetypes and EHR extracts. ArchMS is a prototypical tool that offers functionality for semantic enrichment, standardization and interoperability of clinical data and archetypes, whose application in a real study has been described in this paper. The evaluation of some individual modules included in ArchMS has been reported in previous papers. Performing a complete evaluation would require to 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.

Our case study has shown some advantages 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 OWL 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 [31]. Annotation is one major use of ontologies in biomedical domains since the development and success of the Gene Ontology [32]. Ontologies are exploited in annotation scenarios as controlled vocabularies, since the ontology classes are the annotation entities.

Archetype terminology bindings should not be confused with the annotations provided by ArchMS. Those are usually added to the archetype data elements or terms during its development, but ArchMS does not intend to support the design and development of archetypes. The ArchMS annotations should be understood as archetype metadata since they are associated with the archetype as a whole and not with their individual terms or elements. 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 methods like the ones presented in [33] and [34] whose integration into ArchMS should be studied.

The ontological infrastructure used for representing and annotating archetypes supports the execution of consistent semantic searches, since such ontologies share both the knowledge schema and the semantic context. The archetype search methods exploit the semantic representation of both archetypes and their annotations for retrieving the archetypes that meet the query constraints.

The annotations suggested by ArchMS are based on the semantic similarity of archetypes, which is calculated by applying state of the art semantic similarity functions. In this case, ontologies are used 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.

The semantic similarity function can be customised by specifying the values of the threshold and the weights. We consider the decision on the values of the weights a way of modeling the semantic similarity strategy in an ArchsMS deployment. The default values used in ArchMS are reused from previous projects of our research group, because we found the results appropriate for a series of tests we performed since we are using similar OWL resources. Major changes in the weights and in the threshold will impact on the results of the semantic similarity function, but that should not be an issue if they have been defined according to the desired policy.

The weights should depend on the type of OWL resources used in ArchMS. So far, most biomedical ontologies are limited in the number of properties defined for each class but their taxonomic structure is rich. Consequently, assigning a higher weight to the taxonomic distance makes sense. In case the set of resources is richer in terms of properties, increasing the weight of the properties similarity factor might be of interest. For us, the linguistic similarity weight is the smallest one, because it does not really provide information about the particular structure or meaning of the knowledge entity. Nevertheless, it should be noted that the linguistic level might discover relations that are not made explicit in the formal content of the resource.

Additional research should be made to learn optimal sets of parameters depending on the properties of the archetypes compared and the ontologies used in the annotations of the archetypes. To this end, we are currently analyzing the properties of the ontologies contained in Bioportal, which is also of interest for other research activities in our research group such as ontology evaluation and ontology enrichment.

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. On the one hand, the tasks related to the transformation of archetypes and EHR data between ISO 13606 and openEHR exploit the representation of archetypes as individuals. Those methods were developed before OWL2 was available and fully supported by tooling and reasoners. Such limitations made not possible to represent some types of archetype constraints as OWL axioms, so we decided to represent archetypes as individuals. OWL2 enabled a complete representation of archetypes as OWL classes, and this representation has been used for the most recent activities. Besides, OWL2 punning facilitates the use of the same URI for referring to an entity as individual or class depending on the context, which permits to use the different representations in a transparent way for the users. We are in the process of upgrading the methods for transforming archetypes and data extracts among standards to use the class-based representation of archetypes, which would simplify the maintenance of the platform. The transformation of data between standards should not be confused with the transformation of EHR data into RDF/OWL, since they are independent processes.

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, our work does not propose to replace ADL by OWL, but to use the most appropriate formalism for each task, trying to minimise 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.

The ArchMS infrastructure also permits to move EHR data from the archetype technological space to the Semantic Web one. The data transformation method is driven by domain ontologies, which play the role of knowledge schema in such transformation and is enhanced with the use of semantic patterns. ArchMS accepts any semantic pattern that can be expressed using the OPPL2 grammar, although in this paper we have emphasised the use of the semantic content patterns proposed by SHN 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. 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.

5.2. Use of automated reasoning

Representing archetypes and data in OWL permits the use of automated reasoning, which is used in the ArchMS framework for different tasks. On the archetypes side, ArchMS checks the correctness of archetypes, including specialisation relations, by applying automatic reasoning over the OWL class-based representation of archetypes.

On the data side, automated reasoning is used for the transformation of patients' data into OWL and for the classification of patients. The transformation of EHR extracts into OWL individuals uses reasoning to ensure that only logically consistent content is created. This means that if some OWL individuals would be not consistent with the domain ontology used for the transformation, those would not be generated. Hence, reasoning ensures the generation of consistent datasets, which can be used for the classification of patients. In this second application of reasoning, inferencing is used for classifying the data 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.

In addition to this, OWL reasoning may support the calculation of semantic similarity functions used for the recommendation of annotations and the methods for the extraction of semantic profiles. These activities are currently executed in ArchMS over the asserted model of the ontologies, because the obtention of the inferred model requires the use of the reasoner and it may take some time, and these tasks use the Bioportal ontologies, which permits to query the asserted models through the web services. It should be noted that ArchMS aims at storing as less information as possible, linking as much as possible to existing resources following the Semantic Web principles.

We are considering to use the recently established ontology repository AberOWL presented in [35] as source for inferred models, since it provides inferred versions of the ontologies, and most of its ontologies are crawled from Bioportal. The number and type of services offered by AberOWL must still augment, but it currently permits the execution of SPARQL queries over the inferred models.

In summary, ArchMS performs the most important types of OWL reasoning tasks:

1. Classification of the ontology to infer new classifications for the classes, specially for the validation of the archetypes and for the obtention of the inferred models of the ontologies.
2. Classification of the ontology to infer new classes for the individuals, specially used for the secondary use of the EHR data (classification of the patient data).
3. Checking the consistency of the content, which means checking the satisfiability of the classes, specially used for the validation of the archetypes and checking the logical consistency of the patient data when transformed from XML to RDF/OWL.

OWL reasoning uses the Open-World Assumption (OWA), which means that what is not known at a given moment and, therefore, it is not contained in the OWL knowledge base, may happen, so it cannot be assumed not to happen. The OWA influenced the OWL representation of the archetype constraints for being able to detect incorrect specialisations by using OWL reasoning. Such reasoning-oriented representation, highly focused on sufficient conditions (*equivalentTo* axioms), was not optimal for performing similarity calculations, which exploit necessary conditions (*subClassOf* axioms). The OWA also influences the representation of the classification rules for the patient data, specially the constraints associated with cardinality and exclusion criteria. For example, if maximum cardinality constraints are defined for a given property, the OWL reasoner must be sure that no individual can have, at most, such number of instances of such property. The effective application of exclusion criteria requires to pay special attention to cardinality and disjointness axioms to infer whether an individual meets the inclusion/exclusion criteria defined as OWL classes. For such situations, if the base OWL axiomatization is not optimal, specific axioms to simulate closeness might be added to the patient data. Nevertheless, the Semantic Web community has also developed versions of OWL DL reasoners which use the Closed-World Assumption, which can be an alternative for simulating such behaviour over OWL content.

5.3. Linked Open Data

The Semantic Web provides a natural space for the integration and exploitation of biomedical data. Linked Open Data[27] is the Semantic Web initiative that pursues the publication and sharing of datasets using semantic formats. Tim Berners-Lee[28] 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, which means that the ArchMS infrastructure is able to generate five stars, open datasets by means of the SWIT services. The data represented in RDF/OWL are given HTTP URI, so ArchsMS data are linkable. Given that the ArchMS RDF/OWL data are stored in a triple store, which is currently Virtuoso[29], a SPARQL Endpoint is automatically available for the external access to the RDF/OWL data. The fifth star is achieved by including links to data from external datasets. The data transformation methods included in ArchMS (SWIT services) enable this. For example, if the EHR data contain SNOMED CT codes, the RDF/OWL data can be linked to the corresponding classes in the OWL version of SNOMED CT. Including the links in the data transformation process requires the previous identification of the external resources to be linked. This is an open issue in the Semantic Web community and ArchMS does not provide support in the identification of the datasets yet. The consumption of Linked Open Data enables to use external content for the semantic analysis tasks performed in ArchMS.

Concerning the publication of ArchMS content, it stores the content about the clinical models and the EHR data in different graphs in Virtuoso. The content about the clinical models can be freely published, but the EHR data are subject to legal restrictions, so they cannot be freely published. This is why ArchMS does not enable by default a public SPARQL endpoint. State-of-the-art tools of the Semantic Web community provide flexible ways for managing and publishing the data in LOD format existing in a triple store, so we let the user organisations of ArchMS to decide how they want to enable the access. Hence, the current version of ArchMS generates the open datasets but does not impose a particular way of publishing. Third-parties can use the Linked Open Data generated for the models if the corresponding dataset is published.

5.4. Secondary use of EHR data

In this manuscript, we have described how Semantic Web technologies are used in ArchMS to classify patient data, which is a secondary use of EHR data. This use requires the availability of specific classification rules, which must be implemented in an OWL ontology by means of *equivalentClass* axioms. The quality of the OWL classification obtained will depend on the quality of the definition of the classification rules in the classification ontology. Ideally, such classification rules should be extracted from clinical

[27] http://linkeddata.org/.
[28] http://www.w3.org/DesignIssues/LinkedData.html.
[29] https://www.w3.org/2001/sw/wiki/OpenLink_Virtuoso.

protocols or guidelines. ArchMS has currently no support for other Semantic Web rule languages like SWRL.[30] OWL technologies have some limitations when the rules need to include operators such as count or negation. In our use case we solved this issue by performing all those types of operations at data level. The classification rules might have been implemented using other technologies such as Drools[31] or relational databases. However, these technologies would not be able to exploit automated reasoning, which is enabled by OWL technologies. For instance, in our use case, the type of an Adenoma is obtained by automated reasoning over the values of its properties. The OWL rules could have also been represented as SPARQL queries, but in such case the inferencing possibilities would be significantly reduced. Besides, coding the rules as *equivalentClass* axioms associated with OWL classes permit to reuse and exploit such knowledge, which would not be possible with other representation options.

It may happen that the classification ontology has not been built by 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. If the mappings between the ontologies are not easy to identify, 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.[32] Recently, the openEHR community has produced and applied the Guideline Definition Language (GDL), which exploits guidelines based on the openEHR specification (see [36]). 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 (Linked Open Data) could permit to enhance the construction of such profile. Currently, Linked Open Data directories like datahub.io include more than 500 linked biomedical datasets. The increasing awareness of the possibilities offered by such formats will certainly generate a higher number within the next years, which makes it a corpus worthy of study and exploitation.

5.5. Comparison with related tools

To the best of our knowledge, OWL is not currently being exploited by archetype tools. The use of a Semantic Web infrastructure is likely to be the major novelty of ArchMS over state of the art systems such the openEHR CKM, whose goals are different. CKM is based on ADL technology and is oriented to support the construction and publication of existing archetypes. To the best of our 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 is 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.

LinkEHR is also based on ADL technologies. It is an archetype editor but also a tool for defining mappings between data sources and archetypes and for executing such transformations. The nature of the mappings in LinkEHR and ArchMS are complementary. LinkEHR pursues to express data contained in legacy systems as instances of archetypes, so obtaining what they call normalised EHR data. Such transformation process provides some semantics to the EHR data, but this is limited by the use of ADL technologies. Nevertheless, the data normalised by LinkEHR is an interesting data source for our work. As shown in the case study, we can use LinkEHR to generate the EHR extracts used within ArchMS. We are currently studying the combination of the LinkEHR and ArchMS engines to standardise a process for getting semantic EHR data from legacy systems.

As we mentioned in Section 2, there are tools that permit a canonical transformation of XML and relational data into RDF/OWL, which is basically a change of format. In contrast, ArchMS provides a semantic representation of the input datasets by performing a transformation guided by domain knowledge using an OWL ontology. Fortunately, some of those tools and approaches are evolving and are beginning to consider that the domain knowledge is important to drive the transformation.

5.6. 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[33], and has started to create archetypes. This has also generated interest in the CEM[34] 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[36] 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. The Semantic Metadata Registry Repository [37] is not focused on archetypes but on common data elements. For instance, the IMI EHR4CR project[37], which aims to improve healthcare research by making more efficient the access of academia and industry to EHR data and the participation of

[30] http://www.w3.org/Submission/SWRL/.
[31] http://www.drools.org/.
[32] http://www.semantichealthnet.eu/.
[33] http://informatics.mayo.edu/CIMI/index.php/London_2011.
[34] http://www.clinicalelement.com.
[35] http://miuras.inf.um.es/PoseacleConverter/.
[36] http://metadata-standards.org/11179/.
[37] http://www.ehr4cr.eu/.

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 metadata repository[38].

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[39] or NCI CDE[40] would permit to use traditionally major biomedical terminological sources if not available in OWL.

6. Conclusions

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.

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/FPI/2010 (MCLG) and 15295/PI/10.

References

[1] J.J. Saleem, M.E. Flanagan, N.R. Wilck, J. Demetriades, B.N. Doebbeling, The next-generation electronic health record: perspectives of key leaders from the US Department of Veterans Affairs, J. Am. Med. Inf.Assoc. 20 (e1) (2013) e175–7.
[2] D. Kalra, P. Lewalle, A. Rector, J.M. Rodrigues, K.A. Stroetmann, G. Surjan, B. Ustun, M. Virtanen, P.E. Zanstra, Semantic interoperability for better health and safer healthcare, Research and Deployment Roadmap for Europe. SemanticHEALTH Project Report (January 2009), 2009.Published by the European Commission, http://ec.europa.eu-/information_society/ehealth.
[3] A. Tapuria, D. Kalra, S. Kobayashi, Contribution of Clinical Archetypes, and the Challenges, towards Achieving Semantic Interoperability for EHRs, Healthc Inf. Res. 19 (4) (2013) 286–292.
[4] I. Danciu, J.D. Cowan, M. Basford, X. Wang, A. Saip, S. Osgood, J. Shirey-Rice, J. Kirby, P.A. Harris, Secondary use of clinical data: The Vanderbilt approach, J. Biomed. Inf. 52 (2014) 28–35.
[5] S. Rea, J. Pathak, G. Savova, T.A. Oniki, L. Westberg, C.E. Beebe, C. Tao, C.G. Parker, P.J. Haug, S.M. Huff, C.G. Chute, Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project, J. Biomed. Inf. 45 (4) (2012) 763–771.
[6] S. Abhyankar, D. Demner-Fushman, C.J. McDonald, Standardizing clinical laboratory data for secondary use, J. Biomed. Inf. 45 (4) (2012) 642–650.
[7] T. Berners-Lee, J. Hendler, O. Lassila, et al., The semantic web, Sci. Am. 284 (5) (2001) 28–37.
[8] C. Goble, R. Stevens, State of the nation in data integration for bioinformatics, J. Biomed. Inf. 41 (5) (2008) 687–693.
[9] T.R. Gruber, A translation approach to portable ontology specifications, Knowl. Acquis. 5 (2) (1993) 199–220.
[10] M. Marcos, J.A. Maldonado, B. Martínez-Salvador, D. Boscá, M. Robles, Interoperability of clinical decision-support systems and electronic health records using archetypes: a case study in clinical trial eligibility, J. Biomed. Inf. 46 (4) (2013) 676–689.
[11] C. Martínez-Costa, M. Menárguez-Tortosa, J.T. Fernández-Breis, J.A. Maldonado, A model-driven approach for representing clinical archetypes for semantic web environments, J. Biomed. Inf. 42 (1) (2009) 150–164.
[12] A.M. Iqbal, An OWL-DL ontology for the HL7 reference information model, in: Toward Useful Services for Elderly and People with Disabilities, Springer, 2011, pp. 168–175.
[13] C. Tao, G. Jiang, T.A. Oniki, R.R. Freimuth, Q. Zhu, D. Sharma, J. Pathak, S.M. Huff, C.G. Chute, A semantic-web oriented representation of the clinical element model for secondary use of electronic health records data, J. Am. Med. Inf.Assoc. 20 (3) (2013) 554–562.
[14] C. Martínez-Costa, M. Menárguez-Tortosa, J.T. Fernández-Breis, An approach for the semantic interoperability of ISO EN 13606 and OpenEHR archetypes, J. Biomed. Inf. 43 (5) (2010) 736–746.
[15] C. Martínez-Costa, M. Menárguez-Tortosa, J.T. Fernández-Breis, Clinical data interoperability based on archetype transformation, J. Biomed. Inf. 44 (5) (2011) 869–880.
[16] S. Heymans, M. McKennirey, J. Phillips, Semantic validation of the use of SNOMED CT in HL7 clinical documents, J. Biomed. Inf. 2 (1) (2011) 2.
[17] M. Menárguez-Tortosa, J.T. Fernández-Breis, OWL-based reasoning methods for validating archetypes, J. Biomed. Inf. 46 (2) (2013) 304–317.
[18] M.C. Legaz-García, M. Menárguez-Tortosa, J.T. Fernández-Breis, C.G. Chute, C. Tao, Transformation of standardized clinical models based on OWL technologies: from CEM to OpenEHR archetypes, J. Am. Med. Inf. Assoc. 22 (3) (2015) 536–544.
[19] C. Martínez-Costa, S. Schulz, Ontology content patterns as bridge for the semantic representation of clinical information, Appl. Clin. Inf. 5 (3) (2014) 660–669.
[20] S. Auer, S. Dietzold, J. Lehmann, S. Hellmann, D. Aumueller, Triplify: light-weight linked data publication from relational databases, in: Proceedings of the 18th international conference on World Wide Web, ACM, 2009, pp. 621–630.
[21] C. Bizer, R. Cyganiak, D2R server-publishing relational databases on the semantic web, in: Poster at the 5th International Semantic Web Conference, 2006, pp. 294–309.
[22] M. Rodríguez-Muro, J. Hardi, D. Calvanese, Quest: efficient SPARQL-to-SQL for RDF and OWL, in: 11th International Semantic Web Conference ISWC 2012, Citeseer, 2012, p. 53.
[23] P.L. Whetzel, N.F. Noy, N.H. Shah, P.R. Alexander, C. Nyulas, T. Tudorache, M.A. Musen, BioPortal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications, Nucleic Acids Res. 39 (suppl 2) (2011) W541–W545.
[24] M. Bhatt, W. Rahayu, S.P. Soni, C. Wouters, Ontology driven semantic profiling and retrieval in medical information systems, Web Semantics 7 (4) (2009) 317–331.
[25] P. Resnik, Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language, arXiv preprint (2011) arXiv:1105.5444.
[26] V.I. Levenshtein, Binary codes capable of correcting deletions, insertions and reversals, in: Soviet Physics Doklady, volume 10, 1966, p. 707.
[27] M. Menárguez-Tortosa, C. Martínez-Costa, J.T. Fernández-Breis, A generative tool for building health applications driven by ISO 13606 archetypes, J. Med. Syst. 36 (5) (2012) 3063–3075.
[28] R.A. Falbo, G. Guizzardi, A. Gangemi, V. Presutti, Ontology patterns: clarifying concepts and terminology, in: Workshop on Ontology and Semantic Web Patterns, Sydney, Australia, 2013, pp. 1–13.
[29] W. Atkin, R. Valori, E. Kuipers, G. Hoff, C. Senore, N. Segnan, R. Jover, W. Schmiegel, R. Lambert, C. Pox, European guidelines for quality assurance in colorectal cancer screening and diagnosis, Endoscopy 10 (2012) 0032–1309821.
[30] J.T. Fernández-Breis, J.A. Maldonado, M. Marcos, M.C. Legaz-García, D. Moner, J. Torres-Sospedra, A. Esteban-Gil, B. Martínez-Salvador, M. Robles, Leveraging electronic healthcare record standards and semantic web technologies for the identification of patient cohorts, J. Am. Med. Inform.Assoc. 20 (e2) (2013) e288–96.
[31] R. Stevens, P. Lord, Application of ontologies in bioinformatics, in: Handbook on Ontologies, Springer, 2009, pp. 735–756.
[32] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al., Gene Ontology: tool for the unification of biology, Nat.Genet. 25 (1) (2000) 25–29.
[33] R. Qamar, A. Rector, Most: A system to semantically map clinical model data to SNOMED-CT, in: In the proceedings of Semantic Mining Conference on SNOMED-CT, 2006, pp. 38–43.
[34] S. Yu, D. Berry, J. Bisbal, Clinical coverage of an archetype repository over SNOMED-CT, J. Biomed. Inf. 45 (3) (2012) 408–418.
[35] R. Hoehndorf, L. Slater, P.N. Schofield, G.V. Gkoutos, Aber-OWL: a framework for ontology-based data access in biology, BMC Bioinf. 16 (1) (2015).
[36] N. Anani, R. Chen, T. Prazeres Moreira, S. Koch, Retrospective checking of compliance with practice guidelines for acute stroke care: a novel experiment using openehr's guideline definition language, BMC Med. Inf. Decis. Making 14 (2014) 39.
[37] A.A. Sinaci, G.B. Laleci Erturkmen, A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains, J. Biomed. Inf. 46 (5) (2013) 784–794.

[38] http://www.ehr4cr.eu/files/ExecutiveSummary/EHR4CR-ExecutiveSummaryD4_1.pdf.
[39] https://wiki.nci.nih.gov/display/LexEVS/LexEVS.
[40] http://cdebrowser.nci.nih.gov/.