A Repository of Semantic Open EHR Archetypes

Fernando Sánchez, Samuel Benavides, Fernando Moreno, Guillermo Garzón, Maria del Mar Roldan-Garcia, Ismael Navas-Delgado, Jose F. Aldana-Montes

University of Málaga, Malaga, Spain

Abstract — **This paper describes a repository of openEHR archetypes that have been translated to OWL. In the work presented here, five different CKMs (Clinical Knowledge Managers) have been downloaded and the archetypes have been translated to OWL. This translation is based on an existing translator that has been improved to solve programming problems with certain structures. As part of the repository a tool has been developed to keep it always up-to-date. So, any change in one of the CKMs (addition, elimination or even change of an archetype) will involve translating the changed archetypes once more. The repository is accessible through a Web interface (http://www.openehr.es/).**

Keywords — **openEHR, Semantic Web, CKM**

I. INTRODUCTION

AS stated in [1] EHRs (Electronic Health Records) and ePrescribing have a real impact in the healthcare at a service level and also at the economical level. However the economic impact is reflected in the net benefits only in an average period of 7 years. The use of standards for the establishment of EHRs in healthcare systems would reduce this latency period.

The development of health information systems has been guided by the need for health systems to manage the huge amounts of information that make the use of physical methods unfeasible. However, these systems are not usually constrained to standards. Thus, different hospitals working together or even different services within the same hospital cannot share information about their patients.

Most advanced EHR architectures and standards are based on the dual model-based architecture, which defines two conceptual levels [2].

OpenEHR has at its core the aim of providing the necessary elements for managing electronic health records, providing ways of modelling all the agents implied in a health environment. The openEHR Foundation provides specifications which define a health information reference model together with a language for developing archetypes (clinical models). This language is not part of any software or query language by default. This architecture, based on archetypes, enables the use of external health terminologies (SNOMED CT, LOINC and ICD). OpenEHR uses the dual-model architecture, which has also influenced HL7 CDA. In dual model approaches, archetypes constitute a tool for building clinical consensus and this enables interoperability between different health information systems.

In this approach we are working towards extending how the models are published by providing new perspectives in the use of OWL as a language to provide semantically rich clinical models. Using a translator, we have built a repository of OWL models derived from public ADL models. Ongoing work is helping this proposal to provide ways of improving this semantics by aligning archetypes and health records with ICD-10 and SNOMED-CT. However, because the structure of the EHR is annotated with such terminologies, the information contained in an EHR is mostly composed of text descriptions without terminology annotations on the patient data.

Section 2 presents some related work. Section 3 describes the archetype translation process. Section 4 presents the current version of the repository and its user interface, to conclude with Section 5 explaining the main conclusions and ongoing work.

II. RELATED WORK

Archetypes are considered an important element in the achievement of the semantic interoperability between EHR systems. So, the design of methods to manage them is fundamental [3]. The translation of openEHR archetypes to OWL is not a novel proposal. [4] presents the first proposal of an ontology for representing archetypes in OWL. This ontology is divided into seven integrated ontologies:

- EHR EXTRACT Reference Model. It defines the semantics shared by all kinds of Extract requests and Extracts from openEHR data.
- EHR Reference Model. It contains a representation in OWL of the information model of the openEHR EHR.
- Data Structures Reference Model. It represents the shared data structures used in openEHR reference model, including lists, tables, trees, and history, together with one possible data representation (hierarchical).

- Support Reference Model. It defines identifiers, assumed types, and terminology interface specification used by openEHR reference model.

- Common Reference Model. It contains shared concepts, including the archetype-enabling LOCATABLE class, party references, audits and attestations, change control, and authored resources.
- Demographic Reference Model. It describes the architecture of the openEHR Demographic Information
- Data Types Reference Model. It represents data types, including quantities, date/times, plain and coded text, time specification, multimedia and URIs.

Figure 1 shows a part of this ontology. As can be observed, the design of this ontology is directly driven by the syntactic structure of the archetypes, including their main types, without taking into account compressibility or reusability. From a semantics point of view, this is an inconsistent ontology (tested using the Pellet reasoner in Protégé 4.3), so it cannot be used for reasoning purposes. However, the positive aspect of this ontology is that it is complemented by translation software [5] for obtaining OWL versions of ADL archetypes. This translator is based on the ADL API and the Archetype Object Model (AOM). The OWL model is built using Jena to construct the ontology model in memory while the ADL archetype is simultaneously parsed. A negative aspect of this translator is that it only includes the translation of 2 of the 4 archetype types, and many of the archetypes in these two types cannot be translated due to programming errors.

In this paper we present the roadmap from this approach to reach some goals:

- A comprehensible and reusable consistent OWL ontology.
- A complete translator for any ADL archetypes to consistent OWL ontologies.
- A repository of archetypes and translations able to trace the evolution of the archetypes.
- Software able to automatically align clinical records with external vocabularies.

III. Archetype Translation

An archetype constrains the entities of the reference model. The constraints are applied to the attributes defined for each entity: range, cardinality, etc. In this way, each constrained entity is defined by means of an OWL class in which the corresponding constraints are defined [6].Using the existing translator we have taken several steps to improve it.

A. Error detection.

In this step we have tested the translator using public archetypes in the openEHR CKM (http://www.openehr.org/ckm/). The automatic execution of these archetypes showed the following errors that were solved on the translator provided in our portal (http://www.openehr.es/):

- Non-existing nodes. Some ADL nodes were not expected at certain parsing steps, and this lead the software to an error, stopping the translation process. These nodes were analysed and the translator extended to deal with them properly.
- Repeated class names. The names of the classes in the translation directly rep-resent ADL nodes. ADL does not prevent us from using the same name for different nodes, but OWL does not allow the use of the same name in different classes. In order to solve this problem, the names for these classes were automatically detected and changed to a new name using the parent class name as a prefix.
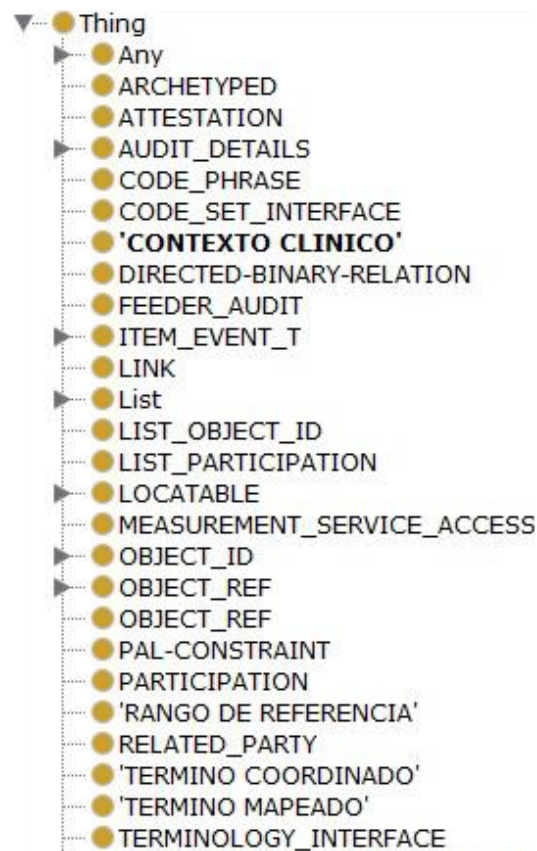

Fig. 1. Main structure of the archetype ontology

B. Incompleteness.

The translator has been shown to be incomplete and so unable to deal with over half of the public archetypes we wished to translate automatically.

- Archetype types that are not translated. The original translator does not trans-late ACTION AND INSTRUCTION archetypes. INSTRUCTION and ACTION have been added to the reference ontology and now the archetypes in this cate-gory are properly translated.
- External vocabulary annotations (SNOMED, ICD, etc.) are not translated. The first step in solving this issue has been to add a new concept to the resulting ontology:

ONTOLOGY_CONCEPT. Thus, the translator has been extended with a component for detecting and dealing with external vocabulary annotations. When the ADL parser detects these annotations this component is activated to add a new instance of the new ONTOLOGY_CONCEPT indicating the external vocabulary used (SNOMED, ICD, etc.) and the term is referenced in the ADL annotation. These annotations will be of help when trying to align clinical data with external vocabularies as this will provide a context to be used by the text mining process.

C. Improve the resulting OWL ontology.

The translator is being modified to eliminate the generation of unnecessary nodes. Some of the concepts added to the OWL ontology were direct translations from the ADL language and are not needed to rep-resent the information of the archetypes. This part of the translator is being modified to use a different structure of the OWL ontology without using these intermediate class names, reducing the complexity of the resulting OWL ontology. This modification which will lead to a totally different translator is still ongoing work which will describe in the following sections.

D. Test case generation.

The translation of archetypes to OWL enables the possibility of using RDF Database Management Systems to deal with clinical data represented as instances (individuals) of these OWL ontologies. However, there are no examples of how clinical data should be represented in these ontologies. Thus, we have developed an instance generator to provide test cases for the data management. Our instance generator asserts individuals in a given ontology in two different ways: inserting individuals according to certain data or inserting individuals randomly generated in a given range.

In order to insert individuals by given data we should follow these steps:

- Instance the reference ontology using "columnX" where "X" is the number of the column from which the program should take the data. We should keep in mind that the first column is "column0".
- The name of each instance in the reference ontology has to be given, with its version at the end, e.g. "example.1", "example.2", "example.1.1", "example.2.3.4", [...].
- This algorithm can be configured to take input files, and decide where to write the results. The separator of data by default is tabulator.
- The program will insert as many individuals as there are lines in the input file.

In order to insert randomly generated individuals in a given range we should follow these steps:

- The input file must have a first line with the type of value that we would like to use separated by spaces (being I=Integer, D=Double and S=String.
- The input file must include one line for each of the types we put in the first line.
- The reference ontology is instanced in the same way as in the previous case, but the data is not collected from the input file, rather it is randomly generated, taking the data types indicated.
- The output is a file that can be used as input for the previous case. Thus, it is possible to create a workflow that uses both cases together, although their maintenance is independent. So, the changes in the reference ontology will only affect the first case, but not the second.

E. Translation examples

Current translation implies that the result is an ontology with a similar structure as an ADL file. Thus, a simple ADL file (Figure 2) will produce a complex structure based on subsumption and object properties.
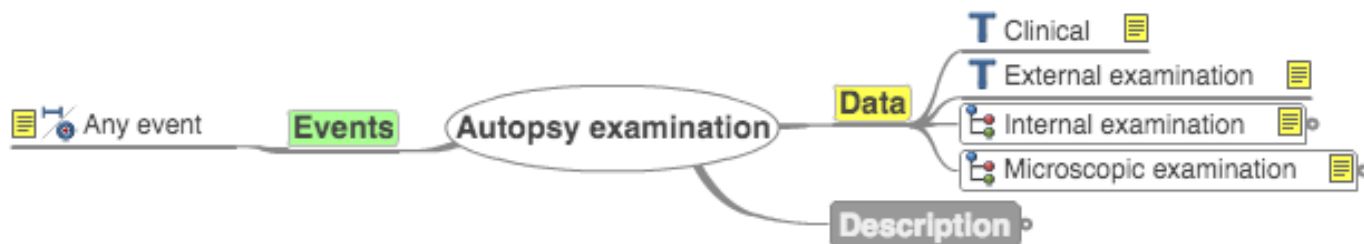

Fig. 2. Autopsy observation archetype

The generation of instances will produce instances for the whole structure of the given archetype ontology. This means generating a lot of instances for intermediate concepts that will serve only as the connection between the archetype and the given data. For example, the following input file (for Blood Pressure archetype) will generate a complex structure of instances as shown in Figure 3:

```
13.10      0 mm[hg] 8.100 mm[hg]  9.50  0
           mm[hg] 75

13.20      0 mm[hg] 8.200 mm[hg]  9.60  0
           mm[hg] 85

13.30      0 mm[hg] 8.300 mm[hg]  9.70  0
           mm[hg] 95

13.40      0 mm[hg] 8.400 mm[hg]  9.80  0
           mm[hg] 105

13.50      0 mm[hg] 8.500 mm[hg]  9.90  0
           mm[hg] 115

13.60      0 mm[hg] 8.600 mm[hg]  9.00  0
           mm[hg] 125
```
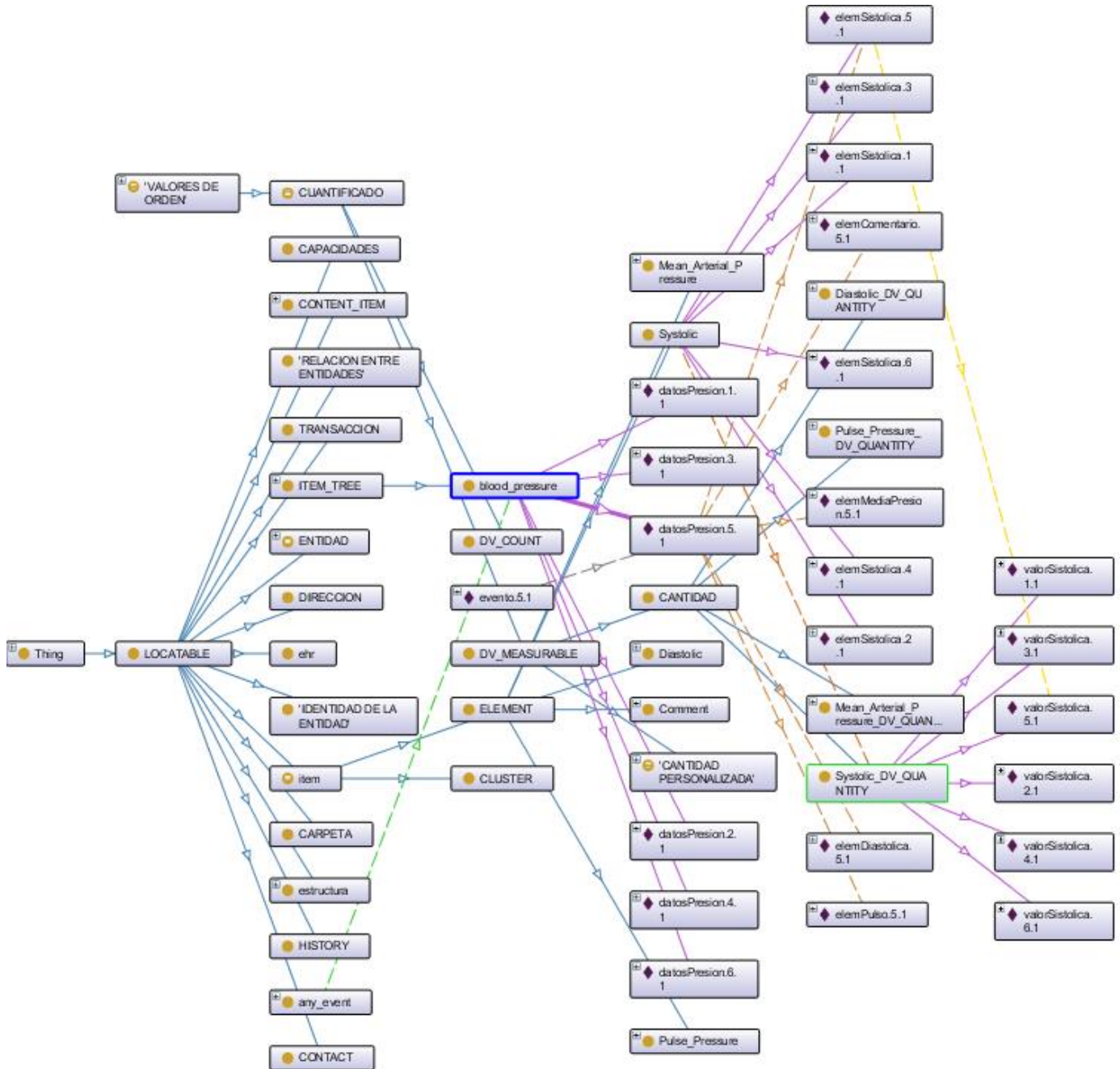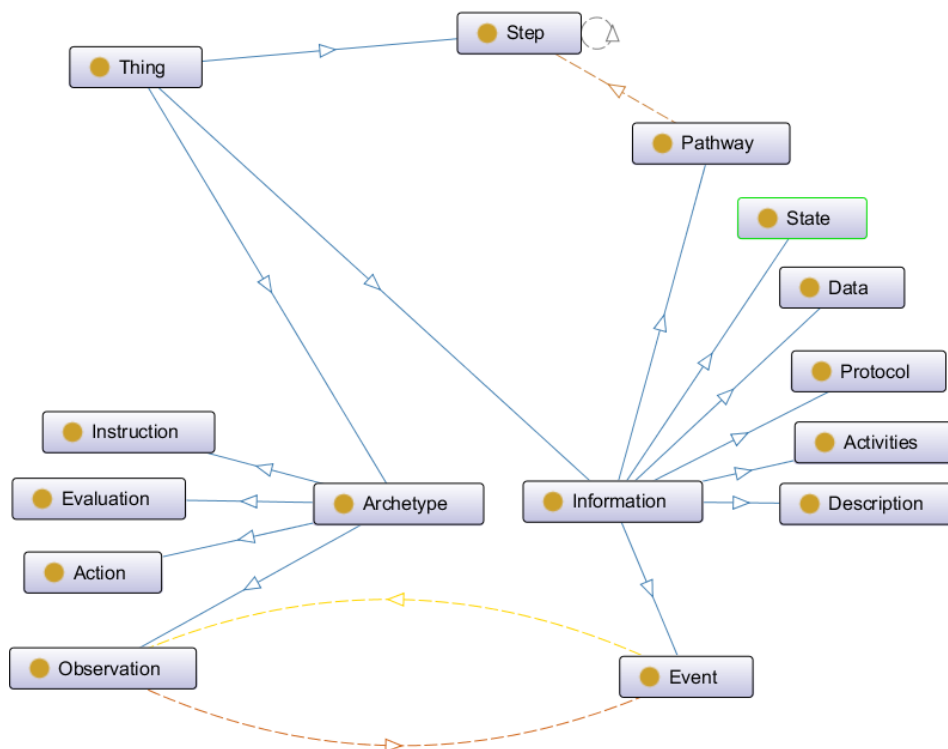

Fig. 1. Example of instances for a given data file

Fig. 2. Reference OWL ontology

F. 3.2   Translation results

An archetype is consistent if its set of defined constraints over both the reference model and the parent archetype are satisfiable. It is necessary to analyse the results of the translation and to check the quality of the archetypes represented in OWL. Generated instances for the current archetypes have been manually evaluated to discover translation errors. This manual process is based on the comparison of the translated archetype as an OWL ontology with the original version in ADL. Nodes are compared by their name and relationships with the other nodes. This ensures that although at first glance the archetype represented in OWL seems to have been translated correctly, there are no hidden translation failures. The quality of the translation is an important part of the translation process in order to ensure a certain level of quality of the translations offered.

IV. REPOSITORY MANAGEMENT

The repository of archetypes is built and updated using a daily batch process connecting to a list of CKMs. This process checks all the archetypes contained in the external repositories, extracts them and compares the contents of the CKM with the local repository. If there are any differences, the process updates the archetypes in the local repository and translates the modified ones to OWL.

In order to connect to the CKM the system uses a web service that provides the CKM and returns a compressed file

with all archetypes structured in directories, classified by type. The following CKMs are currently being accessed:
- NEHTA = http: //dcm.nehta.org.au/ckm/
- openEHR = http: //www.openehr.org/ckm/
- uk = http: //clinicalmodels.org.uk/ckm/
- ezdrav = http: //ukz.ezdrav.si/ckm/
- russia = http: //simickm.ru/ckm/

An archetype can pass through several states (initial, draft, review team, etc.). If an archetype is "published", it cannot be modified. In this case, modifications should be done as an archetype with the same name and higher version number. This way of managing CKM prevents the modification of published archetype contents. The contribution of updating the repository is to keep all versions of archetypes to provide users with translations to the archetype version they are using in their Health Information System, even if a new version has been published. The synchronisation process is as follows:
- If a file has been modified internally, it is replaced in the local repository by the new one and the conversion to OWL is deleted.
- If a new archetype appears, then it is copied to the local repository, this occurs when a new archetype is created in the CKM or is versioned.
- Archetypes are not deleted from the CKM rather they are labeled as rejected or obsolete. Thus, it is not necessary to check whether an archetype is missing from the local repositories.
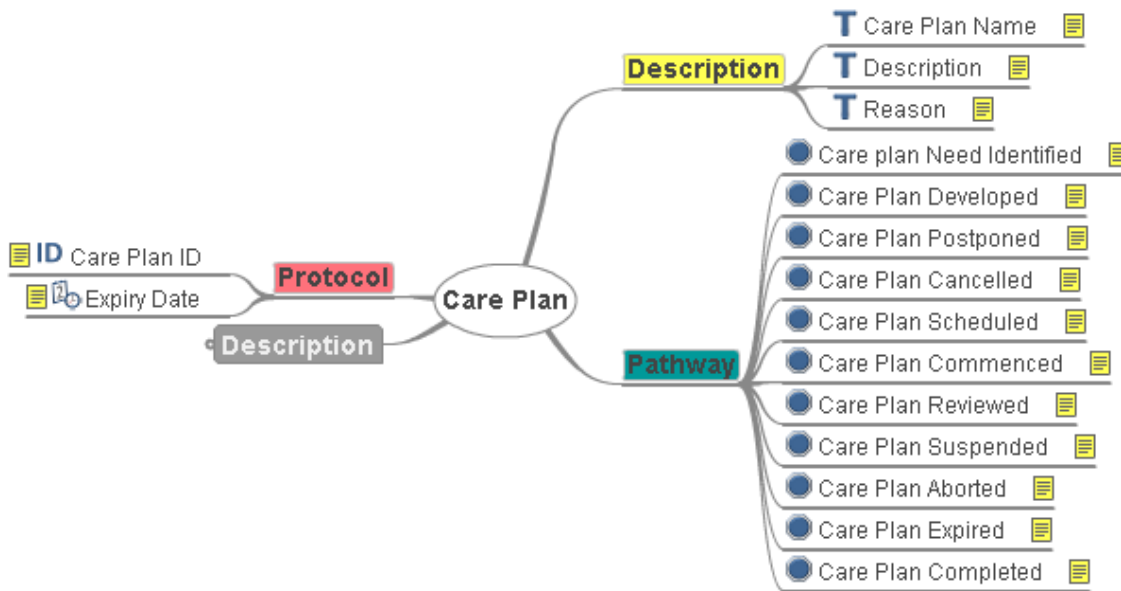
Fig. 3. Care Plan archetype

- Once the local and external repositories have been synchronised, the OWL translation process checks for each of the archetypes added or modified in the local repository.

The repository contains the automatically translated archetypes from public archetype repositories like, for example, CKM. However, ADL allows users to define their own archetypes. For this reason, the translation tool is included in the portal, so users can test its functionality. The system does not keep a copy of the archetype, or the translation, unless the user asks for them to be included in our repository.

V. DISCUSSION AND CONCLUSIONS

The use of standards such as OpenEHR will reduce the time to return on the investment of putting an EHR system to work, with the corresponding economic impact. Additionally, the use of semantics opens new ways of interoperability even with other standards making worthy this initial economic effort. The automatic translation of openEHR archetypes to OWL has been approached in the past. However, in the cur-rent climate in which the interoperability of health information systems is a priority, this topic is of strong interest. For this reason, we have started with previous work and analysed the existing problems in these types of translations. Some of the problems detected have been solved, and an improved version of the translator has been used to provide a repository of OWL ontologies representing public archetypes.

However, there is still much work to be done in this approach. The main issue we are addressing is the design of a reference OWL ontology to lead the translation process towards consistent, comprehensible and reusable ontologies. The reference mod-el (Figure 4) we have designed in the first phase simplifies the representation of archetypes. For example, for the Care Plan Archetype (Figure 5), the translation would

be similar to the ontology in Figure 6.

Neither of the formal representations of ICD-10 presented in the literature has been classified nor their consistencies checked. Even more, some of then uses an OWL-Full component that prevents its use in a semantic classification system based on reasoning. Other approaches propose to model the ICD-10 exclusions using the owl:disjoint axiom, that could lead to a loss of important information and generate inconsistences in the model. There are no ontologies that combine SNOMED-CT and ICD-10-CM. SNOMED-CT and ICD-10 are broadly used in the field of medicine. In fact, SNOMED-CT is being used in most of the Health Information Systems. For this reason, our research group is working on modelling the ICD-10 (International Classification of Diseases, 10th version) [7] as an OWL ontology [8]. This medical classification standard, maintained and published by the WHO (World Health Organization) is used to classify diseases and health problems that have been recorded on death certificates and also in other records. Our ontology has also been aligned with SNOMED-CT [9]. SNOMED-CT terminology often referenced as an ontology, includes all those concepts that relate to each other logically within a specific domain [10]. As many openEHR archetypes are annotated with an ICD-10 code, this enables the possibility of aligning the OWL ontologies in our repository with our ICD-10 ontology. By means of this alignment, the reasoning capabilities of the OWL language can be exploited so as to obtain implicit information about the clinical concept described by the archetype, based on the information contained in ICD-10 and SNOMED-CT, such as its relationships with other clinical concepts, diseases and clinical procedures, to name a few.
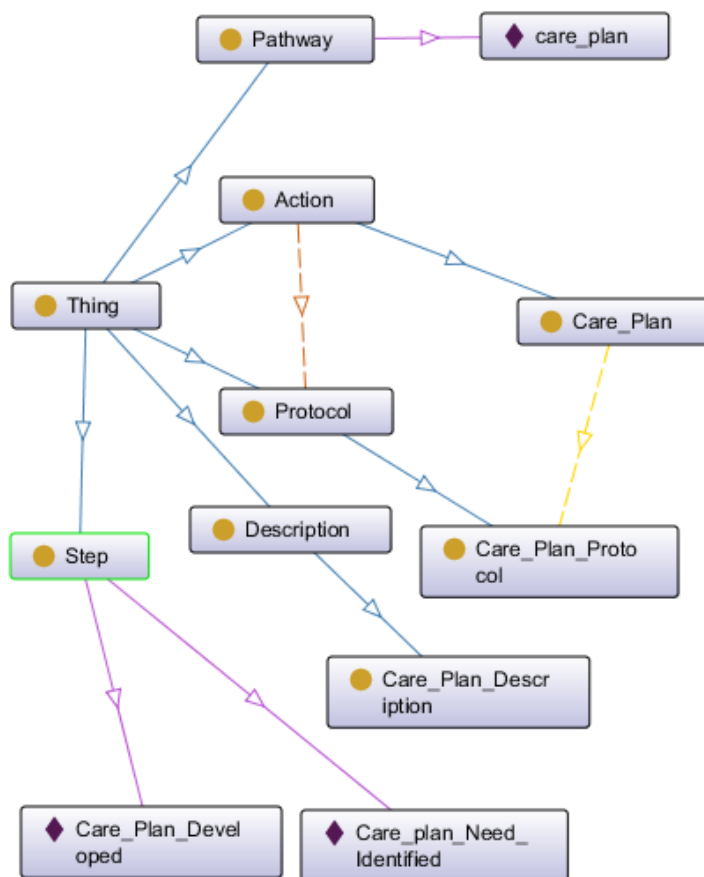
Fig. 4. Translation of Care Plan archetype

REFERENCES

[1]   The socio-economic impact of interoperable electronic health record (EHR) and ePrescribing systems in Europe and beyond. http://www.ehr-impact.eu/downloads/documents/EHRI_final_report_2009.pdf

[2]   Catalina Martinez-Costa; Marcos Menárguez-Tortosa; Jesualdo Tomás Fernández-Breis. An approach for the semantic interoperability of ISO EN 13606 and OpenEHR archetypes. Journal of Biomedical Informatics. October 2010.

[3]   María del Carmen Legaz-García; Catalina Martinez-Costa; Marcos Menárguez-Tortosa; Jesualdo Tomás Fernández-Breis. Exploitation of ontologies for the management of clinical archetypes in ArchMS. University of Murcia (España), 2012.

[4]   Isabel Román Martínez. Aportaciones metodológicas a la integración de sistemas de informa-ción sanitarios basada en gestión semántica. Supervisors: Rafael Robles Arias, Germán Madina-beitia Luque. University of Sevilla (España), 2006.

[5]   Lezcano Matías, Leonardo. Combining ontologies and rules with clinical archetypes. Universi-dad de Alcalá. Departamento de Ciencias de la Computación. Supervisor: Sicilia Urbán, Miguel Ángel. 2012.

[6]   Marcos Menárquez-Tortosa; Jesualdo Tomás Fernández-Breis. OWL-based reasoning methos for validating archetypes. Journal of Biomedical Informatics. November 2012.

[7]   International Classication of Diseases (ICD) website, 775 http://www.who.int/classifications/icd/en/.

[8]   Pastor-Rubia, D.; Rold´an-Garca, M.M.; Navas-Delgado, I. Aldana-Montes, J. F. DTEKT: semantic classification for medical diagnosis Riccardo Bellazzi and Paolo Romano (eds.). Clinical bioinformatics. 11th international workshop. NETTAB 2011. Network tools and application in biology. October 12th-14th 2011, Collegio Ghislieri, Pavia, Italy, 2011, 86-87.

[9]   SNOMED-CT website, http://www.ihtsdo.org/snomed-ct/.

[10]   Jose Luis Allones; David Penas; María Taboada; Diego Martinez; Serafín Tellado. University of Santiago de Compostela (España). G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.

José F. Aldana-Montes is Full Professor at the University of Málaga, Spain. He is the main researcher of the Khaos Research Group. His research focuses on management, integration and analysis of data.

Ismael Navas-Delgado is PhD and Assistant Professor at the University of Málaga, Spain. His research focuses on the use of Semantics and Big Data technologies in Life Sciences.

Maria del Mar Roldan-Garcia had her PhD in Computer Science in 2012 by the University of Málaga, Spain. Currently she is a Lecturer at the same University. Her main interest are ontologies and Semantic Web, scalable semantic reasoning, smart data and biomedical ontologies development and applications.

Fernando Sánchez is Technical Engineer in Computer (Universidad de Málaga, 1997) and MSc in Software Engineering & Artificial intelligence (Universidad de Málaga, 2014). His research is focused on electronic health records and openEHR standard.

Samuel Benavides is student of Bioinformatics at University of Málaga. His major interests are text mining and semantic web. Currently, he is a grant holder working at Khaos, a research group by the University of Málaga.

Fernando Moreno is student of Bioinformatics at University of Málaga. His major interests are data mining, algorithmic and metaheuristics. Currently, he is a grant holder working at Khaos, a research group by the University of Málaga.

Guillermo Garzón Sánchez-Cruzado is studying the degree of Health Engineering the specialization of Bioinformatics Engineering by University of Málaga and University Sevilla. Currently, he is a grant holder working at Khaos, a research group by the University of Málaga.