Exploitation of ontologies for the management of clinical archetypes in ArchMS

María del Carmen Legaz-García[1], Catalina Martínez-Costa[12], Marcos Menárguez-Tortosa[1], Jesualdo Tomás Fernández-Breis[1*]

[1]Departmento de Informática y Sistemas, Universidad de Murcia, CP 30100, Spain
[2]Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria

ABSTRACT

An archetype is a formal model of a domain concept, expressed as a set of constraints based on a reference model and used in a particular domain. In clinical domains, an archetype is defined as a formal model of a clinical concept. Archetypes are considered by many researchers a fundamental tool for the achievement of semantic interoperability of electronic healthcare records. To date, different libraries of archetypes and a series of tools have been developed for archetype-based standards like ISO 13606 and openEHR. However, despite the importance of the semantics of the clinical content and the demonstrated usefulness of ontologies and semantic web technologies, the available platforms for managing archetypes do not exploit them. In this paper we present our archetype management system, which enables the semantic management of archetypes, performing a series of activities based on their OWL representation. The activities include the transformation of archetypes between standards, their validation or the recommendation of appropriate learning contents.

1 INTRODUCTION

The semantic interoperability of clinical information remains being a major goal to achieve in the medical informatics field. Among the Electronic Healthcare Record (EHR) standards envisaged for making it possible, the European ISO 13606 standard (2012), defined by the Technical Committee 251 (CEN/TC 251), the openEHR specification (2012), proposed by the openEHR Foundation, or Health Level Seven (2012) can be pointed out. All of them define an EHR architecture to enable clinical information sharing among information systems, and they are considered a promising solution for the semantic interoperability (see, for instance, Stroetmann *et al.* (2009)). That architecture allows separating the information model from the domain or knowledge model. In this way the information model can be used as a stable model for representing the clinical information in software systems while different domain models or archetypes can be built by clinical experts. ISO 13606 and openEHR follow this architecture in which archetypes, according to Kalra and Archana (2008), are considered the minimal information unit that clinical information systems can exchange and thus the basic semantic interoperability unit. For example, a blood pressure test, a discharge report or a body weight measurement can be defined by means of archetypes.

Archetypes include bindings to terminologies, which provide the clinical meaning of the terms of the archetype. Each clinical institution might develop their own archetypes, thus mechanisms for managing and assuring their quality are needed. In Kalra *et al.* (2008), the need for quality assurance of archetypes is identified, since they will direct how clinical data is captured, processed and communicated. Archetype concepts should be bound in a consistent and appropiate way to clinical terminologies. EuroRec (2012) and the openEHR Foundation are developing governance practices for archetype development and the quality criteria and editorial policies by which certified libraries of EHR Archetypes can be recognised. They will be in charge of evaluating the quality of archetypes.

Despite the growing importance of archetypes, there are not many systems for their management. In Garde *et al.* (2007), the openEHR Foundation described the Clinical Knowledge Manager (CKM), which facilitates basic classification and searching tasks. CKM intends to provide users with the opportunity and means of participating in the creation and/or enhancement of an international set of archetypes. This system works with archetypes defined in ADL, which is a formal language that represents clinical models in a generic way independently of a specific information model. ADL does not allow performing semantic activities like checking the correctness of archetypes, finding similar ones, suggesting appropriate annotations and so on. The use of semantics in tools like CKM is limited to a domain governance ontology that helps to classify archetypes according to the health area, and only the taxonomy is exploited.

In recent years, the bio-ontologies community has developed more than 200 biomedical ontologies, terminologies and controlled vocabularies, most of them being available in the Bioportal website (`http://bioportal.bioontology.org/`). Such knowledge might play an important role in the semantic activities associated with archetypes and whose aim is the achievement of semantic interoperability. However, to date, such set of ontologies has not been used for such tasks, and its application would also contribute to more effective translational research processes.

In addition to this, the research carried out by our group on the use of ontologies for representing archetypes has produced a series of methods for importing, exporting, validating, annotating and searching archetypes using OWL technologies. In this paper, we will try to provide an integrated solution that combines such methods and the Bioportal for providing semantic archetype management methods, which are available in our archetype management system (ArchMS). ArchMS adopts a semantic approach in which most of the activities are performed over the OWL representation of archetypes and it is able to work with both ISO 13606 and openEHR standards. In order to perform the semantic tasks, the system will combine its internal semantic representation of archetypes with the ontologies available in the Bioportal by using the NCBO web services. The availability of

*To whom correspondence should be addressed: jfernand@um.es

such flexible ontological infrastructure makes possible to represent and exploit all the aspects related to archetype in a common technological, semantic environment. This will make easier the adaptation of the platform to the needs of particular groups of users, since each group might annotate and classify the archetypes according to their own clinical needs. The semantic methods would then exploit such semantic dimensions of interest for providing them with the appropriate archetypes for performing such taks. Given the referred common technological space, those methods would be able to combine governance and terminological annotations of the archetypes to get relevant and consistent results.

ArchMS also addresses the recommendation of personalized learning contents based on archetypes. According to Stroetmann *et al.* (2009), the electronic health record of citizens should play a fundamental role for developing mechanisms of information and training for both citizens and professionals. Given that archetypes are considered the knowledge level in dual-model architectures, developing intelligent methods for recommending such learning contents based on archetypes seems sensible.

2 AN OVERVIEW OF THE ARCHETYPE MANAGEMENT SYSTEM

ArchMS (`http://sele.inf.um.es/archms`) can be used by clinical institutions to develop its own archetypes repository or as a central common one. Its purpose is to facilitate the local management of archetypes and to import new ones from other repositories. The architecture of the system is depicted in Figure 1. It provides functions for finding similar archetypes, validating them, and defining search conditions. For this purpose, the system will group different services and tools which will support the execution of those semantic activities.

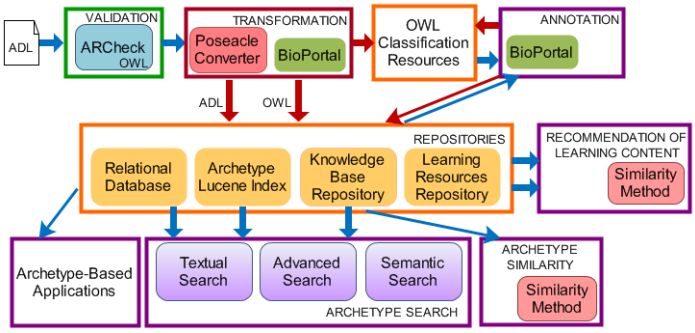

Fig. 1. The architecture of the ArchMS System

We described now such activities:

- Archetype Representation: ArchMS exploits OWL-based archetype representations as both individual and classes. On the one hand, the individual-based representation is the result of the development of a series of OWL ontologies for representing the structural semantics of archetype standards. As a result of this interpretation process, ontologies for the reference and archetype models of both ISO 13606 and openEHR were developed (see Martínez-Costa *et al.* (2010)). The current implementation includes ISO 13606 and openEHR and a common ontology that will allow, in the future, dealing with other EHR standards. Such common ontology has been defined to allow representing archetypes from both standards, in this way it includes concepts from both reference models. Since archetypes are usually represented in ADL, a methodology for automatically transforming them into OWL was designed (see Martínez-Costa *et al.* (2009)). Such representation is used in all the tasks performed in ArchMS except for the validation of the archetypes, which uses a class-based representation. The OWL class-based representation of the EHR standard reference model is generated by following the rules proposed in the Ontology Definition Metamodel specification (ODM) of the Object Management Group (2012). Each concept is defined in this representation by means of an OWL class, and its constraints are defined using OWL-DL axioms. ADL archetypes are then represented by creating constrained subclasses of the OWL reference model classes.
- Archetype Validation: An archetype is consistent if its set of constraints defined over both the reference model and the parent archetype are satisfiable. Methods for checking the consistency of OWL archetypes based on OWL reasoning have been implemented by our group in the ARCheck tool (see Menárguez-Tortosa and Fernández-Breis (2011)), trying to reduce the effort required for implementing quality assurance and validation methods. The validation process consists of identifying whether the definitions of an specialized archetype are consistent with the parent ones.
- Archetype Transformation: The import function allows for including a new ADL archetype into the system. The archetype is then stored in both ADL and OWL format. ArchMS allows to view the ADL definition of an archetype, to download it in ADL or OWL and also to download the result of transforming it from ISO 13606 to openEHR or vice versa (see Martínez-Costa *et al.* (2010)). This transformation is carried out by the Poseacle Converter, which is publicly available online at `http://miuras.inf.um.es/PoseacleConverter/`.
- Archetype Annotation: Archetypes can be classified according to different criteria such as standard, language and other description metadata. Moreover they usually contain terminological information that can be used as annotations. However, such metadata might not be enough when a healthcare institution needs classifications based on particular criteria like application domains or related diseases. For this purpose, ArchMS allows adding new classification resources for annotating archetypes and then using them in the search and comparison tasks.
- Archetype Search: The definition of the same archetype according to different standards or defining more than one archetype for the same purpose may make the process of finding the right one tedious. In order to make it easier ArchMS includes different ways of searching. The *Textual Search* uses the textual content of an archetype to retrieve those with text matches. It uses the Apache Lucene API (`http://lucene.apache.org/`) to execute the query over the Lucene index directory created in the import action, and returns the list of archetypes matching the query. For each archetype, the textual content is extracted from both

the description and ontology sections. The *Advanced Search* allows for using several filtering criteria like text (using the textual search explained above), the information model root concept used in its definition (e.g., cluster, observation, etc.); language or terminological annotations. Finally, the *Semantic Search* allows to incorporate into the queries all the constraints modeled in the archetype ontology. For this purpose, ArchMS integrates our Semantic Flexible Query System (see Martínez-Costa *et al.* (2010)). The queries are created by selecting the concepts involved and the constraints that apply to such concepts. The query is then automatically translated into SPARQL and issued over the knowledge base repository.

- Archetype-based Applications: The development of applications based on EHR standards requires their deep knowledge. In order to facilitate this task our research group has developed the ArchForms system for generating web applications from archetypes (see Menárguez-Tortosa *et al.* (2012)).
- Archetype Similarity: Identifying similarities between archetypes is another step forward towards the achievement of semantic interoperability. ArchMS implements methods based on semantic similarity to detect how similar two archetypes are. Currently, it uses such similarity for suggesting annotations for the imported archetypes. The way this semantic similarity is calculated will be explained in Section 3.
- Recommendation of learning contents: Given a specific archetype, ArchMS provides the learning contents that are closely related to it. Such contents may help patients or clinicians to improve their knowledge about a specific health issue. In order to provide these recommended learning contents the system calculates the semantic similarity between the semantic profile of an archetype and the annotations of the learning contents.

The persistence layer includes four different repositories:

- The knowledge base stores the OWL archetypes which will allow the system to use the semantic representation for different purposes. This repository was developed using Jena (`http://jena.sourceforge.net/`).
- The Archetype Lucene Index refers to the index directory created by the Apache Lucene API, a software library which provides full text indexing and searching capabilities. This repository is mainly used by the search modules.
- The relational database contains the archetype information which is continuously accessed by the system, such as its archetype identifier, language, lifecycle state, etc.
- The learning contents repository is a semantic repository of SCORM learning contents. This knowledge repository was developed using also the Jena Semantic Web Framework.

3 ONTOLOGIES IN ARCHMS

In this section we will provide a more detailed explanation of how ontologies are used in ArchMS. They are mainly used with the following purposes:

- Modelling framework for the domain knowledge;
- Background knowledge for supporting classification processes;
- Semantic context for calculating similarity. The ontologies used in ArchMS are written in OWL and are either ontologies developed by our group or reused from the Bioportal.

3.1 Ontologies as modelling framework

As it has been described in Section 2, ontologies for representing archetypes and EHR standards (EHR ontologies) have been produced. Such ontologies are the semantic backbone of most ArchMS functionality. Despite ADL being the input format for archetypes in our system we will use their ontological representation in order to perform the semantic activities implemented in ArchMS. More specifically, we will use two different ontological representations, individual-based and class-based representations, depending on the tasks that will be performed. In this way, when the archetype is imported into the system it will be transformed into both representations.

On the one hand, the individual-based representation of the archetype will be used in order to recommend annotations for it by using similarity based searching methods, for generating web applications and for making semantic queries. On the other hand, the class-based representation will be used for validating the correctness of the archetype during the import process.

Apart from managing archetypes, ArchMS also manages learning contents, in the form of learning objects. According to Rehak and Mason (2003), a learning object is a digitized entity which can be used, reused or referenced during technology supported learning. SCORM is the current standard for exchanging learning contents so it is used in ArchMS. The semantic representation of SCORM objects was enabled by a semantic extension done to such standard in Esteban-Gil *et al.* (2009).

3.2 Ontologies for supporting classification processes

The semantic annotations are defined over the set of ontologies managed by ArchMS. Therefore, both archetypes and learning contents are classified using the same knowledge areas, which permits developing effective recommendation methods. Such supporting ontologies may have different origins in ArchMS. First, users can upload their own OWL annotation resources by using the corresponding option in ArchMS. Second, when an archetype is imported, its term bindings are processed. Term bindings are annotations of the archetype content using clinical terminologies and ontologies. In case such terminologies or ontologies are not found in ArchMS, they are searched in the Bioportal by using the NCBO Web Services. If such resource is available in the Bioportal in OWL format, then it is imported into ArchMS.

We have mentioned that ArchMS suggests annotation terms based on the similarity between archetypes. However, it also suggests annotation terms from Bioportal resources. For this purpose, NCBO Web Services are used with text pieces of the archetype. By proceeding in this way and leading towards Bioportal resources-driven annotation processes, we pursue to generate sets of interoperable archetypes annotations.

3.3 Calculating similarities

ArchMS also uses ontologies for calculating two different types of semantic similarity. On the one hand, the similarity between archetypes is calculated by finding how similar two individuals of the same ontology are. The semantic similarity functions are based on the ones described in state of the art literature (see, for

instance, Resnik (1999); Castellanos-Nieves *et al.* (2011)), and return a similarity score in the range [0,1]. The compared entities are considered similar if such score is greater than a user-defined similarity threshold. The similarity function combines the following three factors:

- Linguistic similarity: The similarity between the terms associated with the archetype concepts using a string-based calculation.
- Taxonomic similarity: The distance between the terminological annotations of the archetype concepts in a taxonomic structure. In case one or both archetypes are not annotated, the root concept of the archetypes are used. This similarity uses the EHR and the annotation resources ontologies.
- Properties similarity: The similarity between the set of properties associated with the archetype elements. This means that not only the taxonomic structure of the ontology is used but also the axioms defined in the EHR ontologies.

On the other hand, the sets of annotations of both archetypes and learning resources are also compared in order to identify the relevant resources for a given archetype. For each archetype and learning resource, their semantic profiles are obtained by processing its annotations. In this case, the similarity score is based on the taxonomic similarity of the pairs of annotations and only the annotation resource ontologies are used.

4 PRACTICAL USAGE OF ARCHMS

In this section, part of the functionality provided by ArchMS will be shown by using the openEHR archetype for non-TNM staging scores for the colorectal cancer.

4.1 Importing the archetype

Once the ADL archetype is submitted to ArchMS its correctness is validated with ARCheck. In case this process is successful, then it is transformed into the individual-based OWL representation and inserted into the repositories. Then, an initial annotation process looks for the most similar archetypes. This would allow reusing annotations from those archetypes. Figure 2 shows the result of such tasks for our running example. In addition to this, the terminological bindings of the archetype are processed to generate new annotations. In this case, since the archetype has an annotation from SNOMED-CT, this annotation is also stored. Besides, the archetype is also processed with the Lucene API to generate information useful for the search facilities.

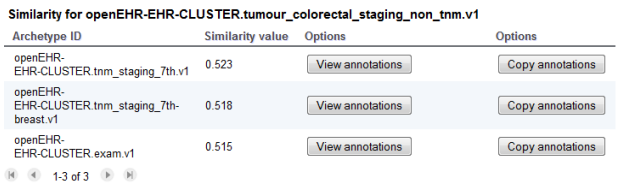

Fig. 2. The most similar archetypes

4.2 Searching for archetypes

ArchMS offers query facilities for searching existing archetypes. For example, using the semantic search interface the user could query for archetypes stored in ArchMS which contain the keyword *cancer*. Such query is built by adding constraints to the different properties defined for the concept *ARCHETYPE* in the EHR ontologies. The definition of this query would be retrieve all the archetypes that:

- *has_description* is an *ARCHETYPE_DESCRIPTION*
- The *details* of such *ARCHETYPE_DESCRIPTION* is an *ARCHETYPE_DESCRIPTION_ITEM*
- Such *ARCHETYPE_DESCRIPTION_ITEM* has a *keywords* whose value is *cancer*

As a result the tool returns the archetypes *ARCHETYPE_openEHR-EHR-CLUSTER.tnm_staging_7th.v1*, *ARCHETYPE_openEHR-EHR-CLUSTER.tnm_staging_7th-breast.v1* and *ARCHETYPE_openEHR-EHR-CLUSTER.tumour_colorectal_staging_non_tnm.v1*.

4.3 Annotating the archetype

Figure 3 shows the annotating interface, which helps the user in the annotation process by making suggestions. ArchMS makes suggestions from both its annotations resources and the ones available in the Bioportal by processing the textual information contained in the description and ontology sections of the archetype. In case such resources do not exist in ArchMS, they can be added by selecting the *install option*, which is not shown in the figure. If we select SNOMED-CT, the left panel would show the suggestions from such resource. Figure 3 displays suggestions for the keyword *metastases*. The lower part of the figure shows the existing annotations for the archetype. For instance, the third item, namely, *Multiple Malignancy*, from SNOMED-CT, was added to the archetype during the import process.

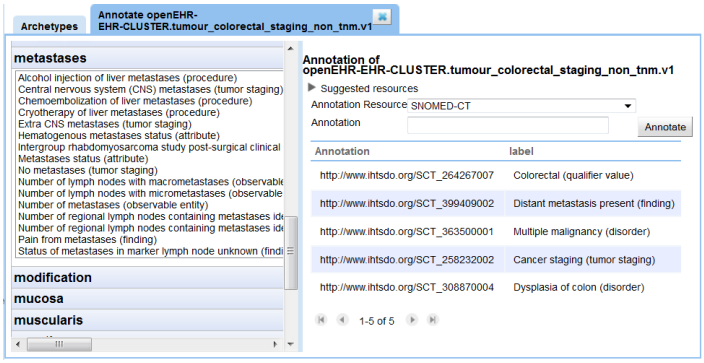

Fig. 3. Annotating an archetype

4.4 Searching associated learning resources

ArchMS allows retrieving learning contents related to an archetype. For this, the semantic profile of the archetype is obtained from its terminology bindings and annotations, and such profile is issued over the repository of learning contents. In our running example, the recommended content is the *clinical guideline for colorectal cancers*, which was manually annotated processing its appendix.

This guideline was recommended because the similarity between the archetype annotation *Multiple malignancy* and the guideline annotations *Adenosquamous carcinoma* and *Carcinoid tumor* is greater than the threshold, which for this example was 0.7. Figure 4 illustrates the taxonomic proximity of such concepts in SNOMED-CT.

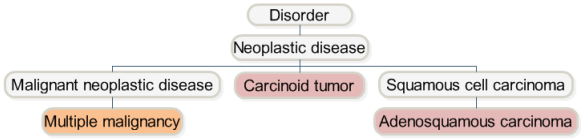

Fig. 4. Representation of the similar annotations in SNOMED-CT

5 CONCLUSIONS

Archetypes are considered an important element in the achievement of semantic interoperability among EHR systems. So, the design of methods to manage them is fundamental. In this work, we have proposed a system to manage archetypes which take advantages of their OWL representation in order to exploit their semantic relations. The system allows to classify the archetypes, to find similarities between different ones and to check their consistency. The current version of ArchMS integrates different tools developed by our research group in recent years. Some of them were developed for one of the EHR standards managed by ArchMS. However, given that we have methods for transforming archetypes between the standards, all the activities can be performed on both ISO 13606 and openEHR archetypes. ArchMS represents archetypes as both OWL individuals and OWL classes, having different methods for generating such representations. In technical terms, archetypes can be approached as a data object or as a knowledge object depending on the activity to be performed on it. For example, transforming an archetype between standards and checking the correctness of an archetype require a different manipulation and exploitation of the archetypes. In fact, our experience also reveals that given the different semantics of the ADL and OWL (e.g., specialization), the execution of different activities with OWL technologies might require slight changes in the OWL representation. This issue should not be compared with the dual class-individual representation provided by OWL punning. In our case, we aim at keeping such representations at the technical, *tooling* level, so users could benefit from functionality powered by semantics.

The annotation feature is a useful instrument to support the classification and identification of similarities in archetypes. For this purpose, ArchMS is able to use the Bioportal ontologies, which is in line with our objective of facilitating interoperability. Bioportal ontologies seem appropriate for supporting annotation processes in our system, because our current algorithm makes use of the taxonomic distance between the concepts. However, most of them have limitations in terms of the number of property axioms of the classes, which is a handicap for our property-based similarity factor. The future use and experiences with ArchMS might provide some hints about their appropriateness and their limitations for wide use with archetypes. We have described how archetypes can be used as an instrument for retrieving relevant learning contents. We plan to extend the approach to the data level by using archetyped EHR extracts. This would permit to obtain a more specific set of educational materials for a given citizen. Moreover, it would be interesting to classify resources according to the target audience, that is, professionals or citizens.

ACKNOWLEDGEMENTS

This project has been possible thanks to the funding of the Spanish Ministry of Science and Innovation through grant TIN2010-21388-C02-02. M.C. Legaz-García is supported by the Fundación Séneca through grant 15555/FPI/2010.

REFERENCES

Castellanos-Nieves, D., Fernández-Breis, J. T., Valencia-García, R., Martínez-Béjar, R., and Iniesta-Moreno, M. (2011). Semantic web technologies for supporting learning assessment. *Information Sciences*, **181**(9), 1517 – 1537.

Esteban-Gil, A., Fernández-Breis, J., Castellanos-Nieves, D., Valencia-García, R., and García-Sánchez, F. (2009). Semantic enrichment of scorm metadata for efficient management of educative contents. *Procedia - Social and Behavioral Sciences*, **1**(1), 927–32.

EuroRec (Last accessed: March 2012). http://www.eurorec.org/.

Garde, S., Hovenga, E., Grnz, J., Foozonkhah, S., and Heard, S. (2007). Managing archetypes for sustainable and semantically interoperable electronic health records. *electronic Journal of Health Informatics*, **2(2):e3**, 1446–4381 (online).

Health Level Seven (Last accessed: March 2012). http://www.hl7.org.

ISO 13606 standard (Last accessed: March 2012). http://www.iso.org/iso/home.htm.

Kalra, D. and Archana, T. (2008). The EHR and Clinical Archetypes: time for clinical engagement! In *eHealth Planning and Management Symposium*, Copenhagen, Denmark.

Kalra, D., Freriks, G., Mennerat, F., Devlies, J., Tapuria, A., and Thienpont, G. (2008). Q-REC. Management and maintenance policies for EHR interoperability resources.

Martínez-Costa, C., Menárguez-Tortosa, M., Fernández-Breis, J. T., and Maldonado, J. A. (2009). A model-driven approach for representing clinical archetypes for Semantic Web environments. *Journal of Biomedical Informatics*, **42**(1), 150–164.

Martínez-Costa, C., Menárguez-Tortosa, M., and Fernández-Breis, J. T. (2010). An approach for the semantic interoperability of ISO EN 13606 and OpenEHR archetypes. *Journal of Biomedical Informatics*, **43**(5), 736 – 746.

Martínez-Costa, C., Miñarro Giménez, J. A., Menárguez-Tortosa, M., Valencia-García, R., and Fernández-Breis, J. T. (2010). Flexible semantic querying of clinical archetypes. In *Proceedings of the 14th international conference on Knowledge-based and intelligent information and engineering systems: Part I*, KES'10, pages 597–606, Berlin, Heidelberg. Springer-Verlag.

Menárguez-Tortosa, M. and Fernández-Breis, J. T. (2011). Validation of the openEHR Archetype Library by using OWL Reasoning. *Studies In Health Technology And Informatics*, **169**, 789–793.

Menárguez-Tortosa, M., Martínez-Costa, C., and Fernández-Breis, J. T. (2012). A generative tool for building health applications driven by iso 13606 archetypes. *Journal of Medical Systems*. 10.1007/s10916-011-9783-1.

Object Management Group (Last accessed: March 2012). Ontology Definition Metamodel specification. http://www.omg.org/spec/ODM/1.0/PDF.

openEHR specification (Last accessed: March 2012). http://www.openehr.org/home.html.

Rehak, D. and Mason, R. (2003). *Reusing Online Resources: A Sustainable Approach to E-Learning*, chapter Engaging with the Learning Object Economy, pages 22–30. Kogan Page.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res*, (95-130).

Stroetmann, V. N., Kalra, D., Lewalle, P., Rector, A., Rodrigues, J. M., Stroetmann, K. A., Surjan, G., Ustun, B., Virtanen, M., and Zanstra, P. E. (2009). Semantic Interoperability for Better Health and Safer Healthcare. Deployment and Research Roadmap for Europe. ISBN-13 : 978-92-79-11139-6.