Semantic Interoperability Between Clinical Research and Healthcare: The PONTE Approach

Anastasios Tagaris[1], Efthymios Chondrogiannis[1], Vassiliki Andronikou[1], George Tsatsaronis[2(✉)], Konstantinos Mourtzoukos[1], Joseph Roumier[3], Nikolaos Matskanis[3], Michael Schroeder[2], Philippe Massonet[3], Dimitrios Koutsouris[1], and Theodora Varvarigou[1]

[1] National Technical University of Athens, 9, Heroon Polytechniou Str., 15773 Athens, Greece
{tassos, dkoutsou}@biomed.ntua.gr, {chondrog, vandro, kmour, doravarv}@mail.ntua.gr
[2] Biotechnology Center (BIOTEC), Technische Universität Dresden (TUD), Tatzberg 47–49, 01307 Dresden, Germany
{george.tsatsaronis, ms}@biotec.tu-dresden.de
[3] Centre d' Excellence en Technologies de l' Information et de la Communication (CETIC), Rue des Frères Wright 29/3, 6041 Chareroi, Belgium
{joseph.roumier, nikolaos.matskanis, philippe.massonet}@cetic.be

Abstract. The adoption of ICT technologies in healthcare for recording patients' health events and progression in Electronic Health Records (EHRs) and Clinical Information Systems (CLIS) has led to a rapidly increasing volume of data which is, in general, distributed in autonomous heterogeneous databases. The secondary use of such data (commonly anonymised for privacy reasons) for purposes other than healthcare (such as patient selection for clinical trials) comprises an emerging trend. However, this trend encapsulates a great challenge; semantic interlinking of two different, yet highly related, domains (in terms of semantics) i.e., clinical research and healthcare. This paper aims at presenting an analysis of the heterogeneity issues met in this effort and describing the semantically-enabled multi-step process followed within the PONTE project for achieving the interlinking of these two domains for the provision of the size of the eligible patients for participation in a trial at the cooperating sites.

Keywords: Semantic interoperability · Electronic health records · Ontology mapping

1 Introduction

Clinical research aims, among others, at revealing the therapeutic potential of substances, methodologies and devices in order for them to be developed into real-world therapies. A critical step in this process lies in the design and conduction of clinical

E. Simperl et al. (Eds.): ESWC 2012 Satellite Events, LNCS 7540, pp. 191–203, 2015.
DOI: 10.1007/978-3-662-46641-4_14

trials, which comprise the investigation of the efficacy and safety of these candidate treatments. Over the years great debate has been taking place concerning the therapy development timeline, the devoted resources as well as the weakened R&D productivity; i.e. the number of therapies which reach the real-world vs the number of investigational therapies researched upon. The reported figures in drug development demonstrate that of every 5000 molecules which are pre-clinically tested, only 1 will in the end be approved and will enter the market [1]. In the meanwhile, the number of new active ingredients entering the market has been significantly reduced over the years [2], while the estimated average cost per drug candidate reaches € 900 million [3], with recently reported figures indicating that this cost may even reach € 9 billion [4] per drug approved. And what is more, the therapy investigation and development comprises a heavily prolonged process with the new drug development timeline being 11.3 years on average [3]. The latter poses significant limitations in the advancement of the domain, while thousands of patients anticipating for a new treatment remain untreated or are following treatments of low/medium efficacy, of not necessarily minor health risks and/or of reduced quality of life.

Key steps in the process of Clinical Trial Design and Implementation are (i) the definition of the inclusion and exclusion criteria (aka eligibility criteria) that describe the target population of the study and eventually the criteria which, patients should meet for participating in the trial and (ii) the recruitment of these patients. The eligibility criteria specification is among the most important steps of study design as it determines at a significant degree the feasibility and the applicability of the study conduction, as well as the value of the study outcome. Patient recruitment comprises another critical step in the clinical trial lifecycle. In fact, prolonged recruitment periods are reflected on study costs, while poor recruitment in trials (i.e. inability to reach the required sample size) is a significant bottleneck for the evaluation of new therapies and quite often leads to study findings of low external validity or even trial termination. One of the key reasons for the series of issues related to recruitment is the limited ability to reach patients due to the insufficient means used such as trial advertising and oral communication of the intent.

A new trend towards improving the process for the selection of eligible patients (based on these inclusion and exclusion criteria) involves taking advantage of ICT technologies and more specifically existing Clinical Information Systems (CLIS) or Electronic Health Record (EHR) systems currently in operation at hospitals and healthcare centres. Although the secondary use (i.e., other than healthcare provision purposes) of healthcare patient data sets has been investigated as of great importance and potential impact, their isolated and specific-focus development has led to significant variations in the organisation and representation of information, the technologies exploited and the implementation of these systems among others. These variations increase the complexity of such efforts, especially in the cases that interoperability needs to be achieved between these systems but also with external ones, such as clinical research information systems. In fact, an additional level of complexity is introduced due to the semantic distance between the different domains which are required to be interlinked.

Focus of this paper is at the challenges posed in the second step which stem from the native interoperability open issues in the world of CLIS and EHRs which becomes an

even harder challenge to address, due to the different nature and way of view between clinical research and clinical practice. Moreover, in this paper we present the approach followed within the PONTE project [12] for overcoming the variety of heterogeneity issues between clinical research and clinical care domains. In fact, among the key objectives in PONTE is the semantic interlinking of clinical research, with particular focus on the clinical trial eligibility criteria, and the clinical patient data at healthcare for providing the researchers with an instant view of the size of the eligible population available at the cooperating hospitals.

More specifically, in Sect. 2, we analyse the great challenges faced in order to achieve interoperability among clinical research systems and healthcare clinical data sources which stem from the variety of heterogeneity issues at structural, syntactic, semantic and interfacing levels. In Sect. 3, we present the methodology developed and applied within the PONTE project for allowing communication with healthcare patient data sources for two purposes: (a) retrieving the size of the available population satisfying the eligibility criteria of a clinical trial during its design and (b) selecting the eligible patients for screening purposes after approval of the study protocol. Section 4 presents the key open issues while Sect. 5 summarises the main findings of the presented approach.

2 The Challenges

Given a set of eligibility criteria for patients to be selected and participate in a clinical trial, there is a need for a systematic communication with CLIS or EHRs of hospitals or healthcare centres in order to identify the patients that satisfy this set of criteria. To our knowledge there is no automated and systematic procedure to perform this process due to the heterogeneity of the multiple and different datasources. A methodological approach to overcome the challenges posed by high heterogeneity (at system, syntax, structure, semantics and interface/messaging level [5]) of datasources is presented in this paper.

According to [6, 7], the heterogeneity between data sources is classified into the following four categories:

i. *System Heterogeneity*: encapsulates the differences at the level of different hardware used and operating systems.
ii. *Syntactic Heterogeneity*: is related to the data models (relational, object oriented, hierarchical model, etc.) used to organize our knowledge. Based on the data model selected, a different language (SQL, OQL etc.) may be used to access the data. Moreover, many systems do not provide direct access to the data, but enable data retrieval using user/system specific queries. In such cases additional heterogeneity issues arise at interface/messaging level, concerning the means of accessing the data as well as the structure of the queries posed and the results retrieved.
iii. *Structural Heterogeneity*: is based on the way we organize our information. We can represent the same information in many different ways even if we use the same data model. Different schemas may be used for representing the same

information. For example, a many-to-many relationship can be implemented by using one-to-many and many-to-one relationships, but there is no standard way to find this implemented in the existing CLIS or EHRs. Different implementations may follow different approaches depending on the needs of the specific hospital department and applications.

iv. *Semantic Heterogeneity*: The semantic heterogeneity is related to the meaning of the elements of the schema. It represents the way people understand a specific domain of knowledge. They can use different terms to refer to the same concepts, while they may use the same concept to refer to different things. It can be found both at schema level and data/instance level. Thus, there might be cases of different terms referring to the same concepts (synonyms), the same term referring to different things (homonyms), missing data across EHRs, concepts used that have a broader, narrower or overlapping meaning. Also, differences could exist in the unit or scale of the measurements.

All the above have been depicted in Fig. 1.

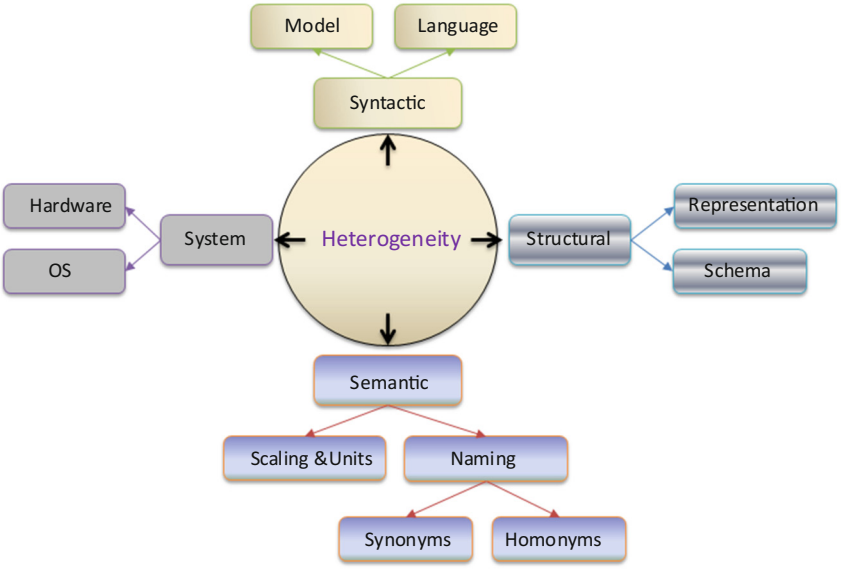

Fig. 1. CLIS and EHR datasources heterogeneity

A typical example of synonyms at the *schema level* of the different datasources relates to the name of the field, under which the "disease" of a patient is recorded, with terms such as "disorder", "diagnosis", "disease", "condition", being used across different EHRs. An example of *semantic heterogeneity* across EHRs at the *data/instance level* lies in the vocabularies used for recording the disorders for patients. For example, "myocardial infarction" may also be documented within a datasource as myocardial infarct or MI – an acronym of the latter ones - (synonym), acute myocardial infarction (narrower meaning) or heart disorder (broader meaning) with different codes being used in different coding systems, vocabularies and classifications.

The best case scenario, in which international terminology standards (such as the ICD-10 [8] or SNOMED-CT [9]) have been adopted for the documentation of a patient's disease, still imposes important challenges. Although certain mapping efforts across those terminologies can be used for at least overcoming this type of heterogeneity (a typical example of which is the NCI Metathesaurus [10]), handling terms of narrower, broader or overlapping meaning, even in this case, still remains a challenging task. The challenge gets even greater in cases that *local coding schemas and/or vocabularies* are used within the hospitals, which in turn requires mapping to an international standard vocabulary or coding system in order for any external application to query upon this data across hospitals in a transparent and efficient manner.

In order to achieve successful communication with the different data sources, we need to handle all the aforementioned heterogeneity issues and provide a way to the end user (who in most of the cases is expected to have limited IT background being a clinical researcher) to pose queries without needing to know the internal characteristics of these datasources. A safe way to overcome the challenges in the different interoperability levels is needed and the presented work aims at realising this through the use of semantics and appropriate ontologies to converge between the clinical research and clinical care domains.

Still several challenges exist, as clinical care systems are based mainly on RDBMS implementations with no semantic or ontological elements. A first challenge is to convert a relational-based model to an ontological model. Assuming that the clinical research domain (and in our case the eligibility criteria) is also represented using an ontological model, the next challenge is the *alignment* of the different ontologies. Hence, the eligibility criteria ontology and the CLIS or EHR ontology (derived from the corresponding CLIS or EHR system respectively) need to be aligned, while *mapping services* together with *transformation rules and mechanisms* (to address terminological and structural issues) have to come in place. This is depicted in Fig. 2.

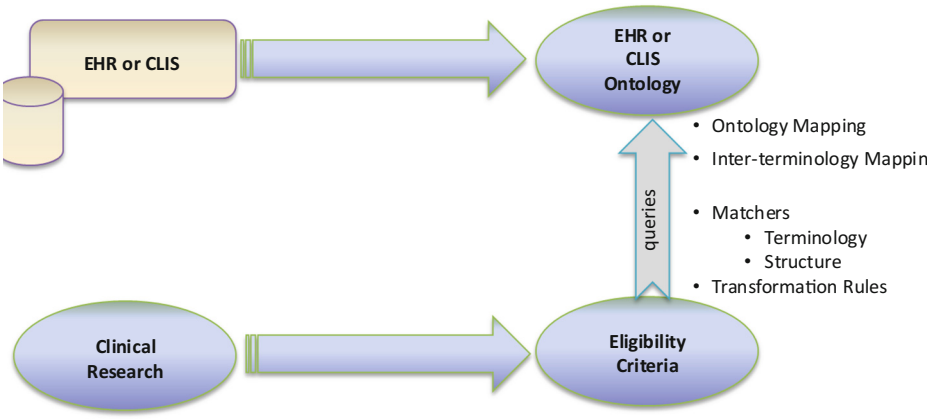

Fig. 2. Convergence of Clinical Research and Clinical Care using Ontologies

According to [11], combining and relating ontologies is by no means a straightforward procedure. On the contrary, many challenges are required to be overcome depending on the "mismatches" between those ontologies. These mismatches arise when two or more ontologies describe (partly) overlapping domains (apart from language mismatches which may also exist) (Fig. 3):

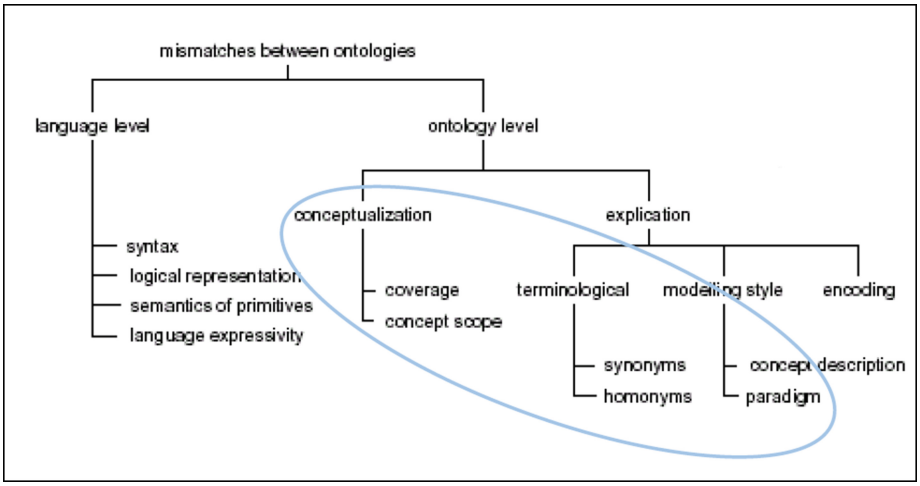

Fig. 3. Mismatches between ontologies [11]

i. *Language Level Mismatches*: This category of mismatches comes from the different languages that can be used to describe ontologies. Each one has its own syntax, representation, semantics and expressivity.
ii. *Conceptualization Mismatch*: These mismatches are met due to the different ways a domain of knowledge can be interpreted (conceptualized). They are further classified into "scope" and "model coverage and granularity" mismatches. In the first case, two classes which represent the same concepts may not have the same instances. In the second case, the part of the domain that is covered is different and some models are more detailed than others
iii. *Explication Mismatch*: These mismatches are driven by the way a domain of knowledge is specified. They are classified into the following subcategories:

   – *Paradigm*: Different paradigms can be used to represent concepts. Paradigm mismatch is the problem when two different approaches are to be used at the same time. The most common case of this idiom is in between ObjectModeling and RelationalModeling. The use of different "top-level" ontology is also an example of this kind of mismatch.
   – *Concept Description*: These mismatches arise from the choices made while conceptualizing a domain. For instance, the full name of a person can be specified using one property ("full_name") or using two properties ("first_name" and "last_name").

– *Terminological Mismatches*: The same concept may be represented using different names (*Synonym*) or the same term may have different meanings (*Homonym*).
– *Encoding*: These mismatches occur due to the variety of ways in which a value may be represented. For instance, a date may be represented as "dd/mm/yyyy" or "dd/mm/yy" or "dd-mm-yyyy".

Another challenge that comes into place has to do with the alignment of datasources heterogeneity with the ontologies heterogeneity.

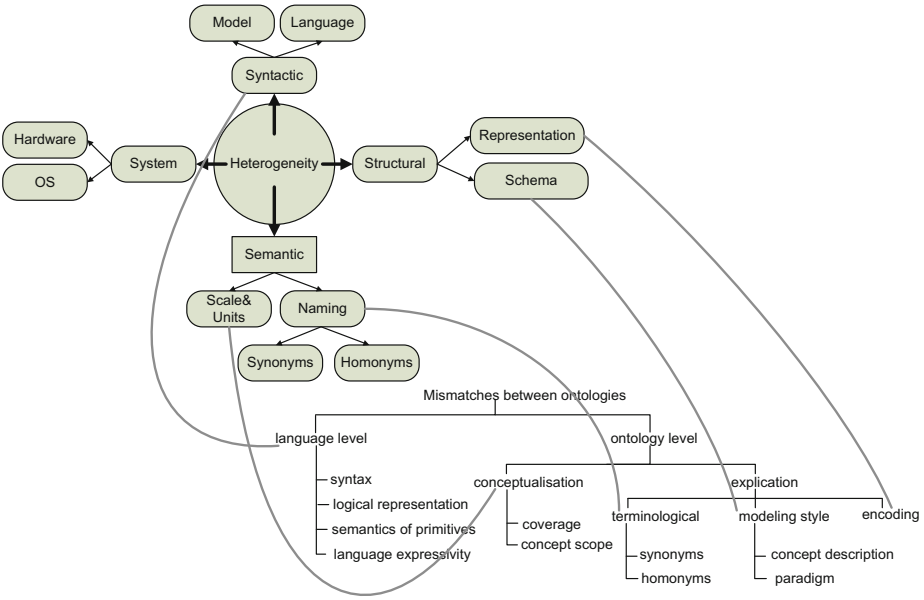

Fig. 4. Ontologies Heterogeneity and Data sources Heterogeneity

As we can see in Fig. 4, the Language Level mismatches are related to the Syntactic Heterogeneity. From the *Ontology Level* mismatches, the Conceptualization mismatches are related to the Scaling & Unit Heterogeneity (a subcategory of the Semantic Heterogeneity) whereas the *Explication mismatches* are associated with both Structural and Semantic Heterogeneity. More precisely, the Modelling Style and Encoding mismatches are linked with the Schema and Representation Heterogeneity accordingly (Structural Heterogeneity) while the Terminological mismatches are related to the Naming Heterogeneity (which comprises a subcategory of the Semantic Heterogeneity).

In the following section we present the methodological approach that we have followed in PONTE in order to address the aforementioned challenges.

3 Semantic Interoperability in PONTE

3.1 The PONTE Approach

Within PONTE we follow an ontological approach covering the whole representation of information requested and retrieved; from the clinical research-oriented information (i.e., the eligibility criteria) to the healthcare patient data on which the queries are posed. For this purpose two ontologies have been developed using the Web Ontology Language (OWL) namely: (i) the *Eligibility Criteria ontology* aiming at representing the eligibility criteria parameters (such as gender, life expectancy, contraindications to a treatment etc.) as well as the relationships among them and (ii) the *Global EHR ontology* which comprises the PONTE-side representation of the healthcare patient data. The latter include demographics, general characteristics and health-related parameters (such as administered medication, diagnosed conditions, operations scheduled or performed, etc.).

The semantic distance of the two types of information represented by the two aforementioned ontologies can be better understood through the following example: a potential inclusion criterion for a trial could be "*patients with adequate liver function*". Such information is not expected, in most cases, to be found within a CLIS or an EHR system. In order to be able to apply such a criterion in those systems, it needs to be converted into a query applicable to such systems based on their semantics; i.e., with parameters – at least partially - semantically correlated with this criterion. In the aforementioned example this corresponds to patients who have *never* been diagnosed with "*liver cirrhosis*" or are not *currently* suffering from "*Hepatitis*", etc. Hence, the ontological approach has been chosen in order to exploit the semantic relationships among the parameters for alignment purposes across the two domains (Fig. 5).

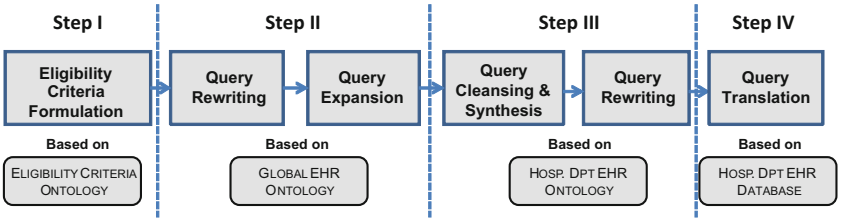

Fig. 5. PONTE approach for Interlinking with healthcare patient data

In the figure above, the main steps followed in the methodology developed and adopted within the PONTE platform are presented. Based on this approach, the initial set of eligibility criteria is expressed as SPARQL queries. Overall, a series of transformations from each initial SPARQL query representing the eligibility criteria and expressed over the Eligibility Criteria Ontology to one final SQL query applicable to the relational EHR database take place. This approach is driven by the effort to break down the heterogeneity issues so that they are almost individually addressed at a series of steps and thus reduce its complexity.

In Step I, the eligibility criteria are formulated as SPARQL queries expressed over the terms of the Eligibility Criteria ontology. Rather than generating a single SPARQL query for all the criteria, each criterion is formulated as a single query, so that it can be processed individually. One of the main reasons behind this decision is the fact that not all criteria are expected to be applicable at the different patient datasets (for example, information indicating that a patient will be willing to provide informed consent (a typical inclusion criterion) is not expected to be found in any EHR).

In Step II a query rewriting process takes place during which the initial SPARQL queries generated in Step I are transformed into SPARQL queries expressed over the terms of the Global EHR ontology. In order for this transformation to take place, an initialization phase has been set up during which the Eligibility Criteria ontology has been aligned with the Global EHR ontology and transformation rules have been built to be used at run-time. The heterogeneity issues addressed at this step are purely semantic and structural. For example, an expected eligibility criterion would be age range (e.g., patients from 20 to 50 years old), whereas the date of birth is expected to be found within an EHR. Applying such a criterion would require at this step alignment of age with date of birth and a transformation rule indicating the calculation of age from the date of birth. At this point it should be noted that the vocabularies and classifications (for disorders, active substances, laboratory examinations, race, etc.) used in these two ontologies are kept the same for semantic consistency reasons.

In this step, *query expansion* comprises an accompanying process, which might be required in some cases. Let's take for example the case that an exclusion criterion describes that "Patients suffering from any heart disorder" should not participate in the trial. This criterion, in fact, encapsulates a wide range of disorders, while in the EHRs (where the *diagnosed* diseases of patients are recorded) a possible record (which is semantically related to this criterion but not equivalent) would be "ST elevation (STEMI) myocardial infarction of anterior wall". Thus, in this case the query expansion process aims at collecting the terms of narrower meaning and expanding the SPARQL query (for any concept included in a criterion) so that all possible values, for the EHR parameter included in the query, are covered. After this step, as described above, the eligibility criteria are *semantically* much closer to the EHRs than when initially generated.

At this point it should be noted that, in most – if not all – of the cases, CLIS and EHRs are implemented using relational or object oriented database servers. In order to extract such database schemas to a semantic representation (ontological format) we have used the D2R server [13]. The D2R server takes as input a mapping file expressed in the D2RQ Language [14]. The D2R server provides also a tool that can be used to generate this mapping file based on the schema of the database. The reason for performing this process and extracting the ontology of the EHR database instead of directly transforming the Global EHR ontology-based SPARQL queries into the final SQL query for the EHR database is that such a transformation would require addressing a variety of heterogeneity issues, namely *semantic, structural, syntactic and interface heterogeneity*, *at the same time*.

Hence, in Step III the queries produced by Step II are synthesized into one SPARQL query which encapsulates all the eligibility criteria, each of which previously comprised a separate SPARQL query. The reason behind this is that in order for a

patient to be eligible at the clinical trial s/he should meet all the inclusion and exclusion criteria specified for this study. Before the synthesis of the SPARQL query, a cleansing mechanism is applied. During this process the SPARQL queries representing eligibility criteria which cannot be applied on the healthcare patient data sets are removed. The need for this mechanism is driven by two main reasons. The first one lies in missing information at structural level of the patient data sets. For example, an eligibility criterion might exclude "Patients who have been smoking for the past 5 years", whereas no information about the tobacco use is being recorded for patients at the EHR or CLIS. The second reason involves missing information at data level. In this case, the healthcare department (e.g. cardiology) might be recording specific health-related events for patients, such as a specific subset of diseases (for example only heart and metabolism diseases). If an eligibility criterion aims at including "patients who suffer from liver cirrhosis", then the querying of such a data set will provide no results. *The analysis of this response, however, shows that it is of low confidence since it is based on the wrong assumption that the patient data sets provide a full description of the patients' health and thus lack of information means no such health event for the patient ever occurred.*

The resulting SPARQL query (which is still expressed over the terms of the Global EHR ontology) is then transformed into a SPARQL query expressed over the Hospital Department EHR *ontology* (aka Local EHR ontology) which comprises the ontological representation of the healthcare patient data sets. As in Step II, a mapping process between the Global EHR ontology and each one of the Local EHR ontologies has been held and a set of transformation rules has been developed during the initialization phase. The latter set is being used during the query rewriting process.

Taking into consideration that the vocabularies and coding systems used at the side of the EHRs are not expected to be the same as within the PONTE platform, the query rewriting process at this step also encapsulates *semantic mapping of the terms* used. For example, there might be the case that the diagnosis in an EHR are coded based on ICD9. In this case, transformation of the disorder values in the eligibility criteria from ICD10-CM (which is used within PONTE) to ICD9 would be required. For this purpose, the EVS Vocabulary Servers[1] are being used. This step focuses on overcoming semantic and structural heterogeneity issues. In Step IV, the SPARQL query synthesized during the previous step is translated into an SQL query based on the mapping file generated by the D2R server as described above.

3.2 The Mapping Issues

As mentioned above, during the initialisation phase of the presented methodology a set of mapping steps is required among the ontologies involved in each step. The mapping process itself poses a series of requirements on the language which will be used for representing it. In [15] the functionality/expressivity that should be provided from an ontology mapping language is investigated. According to this paper, the mapping is defined as a list of assertions (mapping rules) each of which defines the relation

[1] https://cabig.nci.nih.gov/community/concepts/EVS/.

between a set of ontological entities (concepts and relations). For each one we should determine whether it is bidirectional or not. Also we should keep information about the nature of the mapping rules (generated from an ontology alignment algorithm or defined by the user). The mapping language should be expressive enough to allow the definition of both simple and complex correspondences in order to handle the *syntactic and structural* differences (or conceptualization and explication mismatches). It should also contain the necessary operators which would allow for specifying that two concepts are equal or that one concept is more general than the other one or that they are partially overlapping. Additionally it should give us the functionality needed to define more complex relations such as that a concept is equal with the intersection of two or more concepts from the other ontology.

As an example let's take the full name of a person. This may be represented with the property "full_name" in one ontology, while the same information may be represented using the properties "first_name" and "last_name" in another ontology. The mapping language should allow us to specify this correspondence. In this case, simply defining that the full name is equal with the intersection of the first and last name is not enough. We should also specify that the first name is the first token of the full name whereas the last name is the second token of the full name.

Let's see another example. Suppose that we have in the first ontology the class "Person" in which the age is specified using the property "has Age". In the second ontology we have only the concept of the Adult. These ontologies don't have exactly the same information. In the first ontology the information for a person is the actual age, whereas in the second ontology the corresponding information describes whether s/he is an "adult" or not. As we know, an adult is a person whose age is greater than 18 years old. So, it is obvious that there is a correspondence between these ontologies. In this case, we should specify that the Person which has Age greater than 18 is considered an Adult person and vice versa.

Due to the variety of the correspondences between the two ontologies, their specification requires much more than what a declarative language can offer. Hence, we need a procedural language which enables us to specify every possible relation identified. These issues, which comprise real problems that we should overcome in order to achieve successful communication with existing CLIS or EHR systems, need to be considered when deciding upon the mapping language to be used. Concerning the specification of the correspondence between Global and Local EHR ontologies and in order to overcome the semantic and structural interoperability issues, we take advantage of the "Expressive and Declarative Ontology Alignment Language" EDOAL [16], which allows for representing correspondences between the entities of different ontologies. Its key strength and the main reason for its selection is the fact that it extends the ontology alignment format and it enables the representation of complex correspondences.

4 Challenges and Open Issues

As mentioned in Sect. 2, interoperability between clinical research systems and healthcare datasources comprises a great challenge due to the variety of heterogeneity issues primarily at the semantic level. The proposed approach followed within the PONTE

project breaks down the process of communication with EHRs into different steps, each of which focuses on specific aspects of heterogeneity and aims at reducing the semantic distance between the eligibility criteria specified in clinical research and the healthcare patient data.

Among the greatest challenges met in this process rises from the fact that, although international vocabularies and classifications are being developed the past years, their adoption in the healthcare domain still remains limited. Mapping local vocabularies to the corresponding international ones, is a process which requires significant manual effort and involves experts both from the medical and the technical domains. A similar challenge is related to the structure of the EHRs. Despite efforts being made in the agreement on and provision of guidelines and specifications describing the information and data elements which should be recorded in an EHR, still great variations are found across EHRs in terms of parameters, their naming, structure, etc. In our approach, in order to reduce the effort required to perform the mapping between the Global EHR ontology and the ontology extracted from each EHR data and structure, one of the next steps in our work will involve the development of semi-automatic mechanisms allowing for the specification of the ontology mappings.

Moreover, given the wealth of eligibility criteria which are met across clinical trials and, consequently, a researcher would potentially specify, the enrichment of the Eligibility Criteria ontology and the Global EHR ontology comprises part of our ongoing work. This effort includes deep analysis of patterns and categories across eligibility criteria specified in trials registered in clinicaltrials.gov[2] and EU Clinical Trials Register[3] together with close interaction with the clinical experts of the PONTE project.

5 Conclusions

This paper presented an analysis of the different heterogeneity aspects met in the application of eligibility criteria in Electronic Health Records and Clinical Information Systems serving patient selection purposes for participation in clinical trials. The approach followed within the PONTE project for addressing the variety of the heterogeneity issues, including syntactic, structural, semantic ones, has been presented. In order to reduce the complexity of the process, a multi-step process has been adopted, with each step focusing on particular heterogeneity issues. At the same time each step aims at reducing the semantic distance between the eligibility criteria, expressed based on the semantics of the clinical research domain, and clinical care patient data, expressed based on the healthcare domain semantics. This approach requires an initial mapping process, which in cases that local vocabularies and health data structures are used, might become quite intensive. Hence, part of our future work will focus on building a tool allowing for the semi-automatic alignment of the Global EHR ontology and the produced local EHR ontologies. Moreover, in order to capture the richness of the eligibility criteria in clinical research, effort will be placed in further analysing the

[2] http://clinicaltrials.gov/.

[3] https://www.clinicaltrialsregister.eu/.

possible categories and parameters in inclusion and exclusion criteria and extending the Eligibility Criteria ontology as well as translating them into the respective EHR parameters.

Acknowledgements. This work is being supported by the PONTE project and has been partially funded by the European Commission's IST activity of the 7th Framework Programme under contract number 247945. This paper expresses the opinions of the authors and not necessarily those of the European Commission. The European Commission is not liable for any use that may be made of the information contained in this paper.

References

1. Kraljevic, S., Stambrook, P.J., Pavelic, K.: Accelerating drug discovery. EMBO Rep. **5**, 837–842 (2004). doi:10.1038/sj.embor.7400236
2. Van den Haak, M.A., Sculthorpe, P.D., McAuslane, J.: New Active Substance Activities: Submission, Authorisation and Marketing 2001. CMR International, Epsom (2002)
3. Di Masi, J., Hansen, R., Grabowski, H.: The price of innovation: new estimates of drug development costs. J. Health Econ. **22**, 151–185 (2003)
4. Harper, M.: The Truly Staggering Cost of Inventing New Drugs (2012). http://www.forbes.com/sites/matthewherper/2012/02/10/the-truly-staggering-cost-of-inventing-new-drugs/
5. Sheth, A.: Changing focus on interoperability in information systems: from system, syntax, structure to semantic. In: Goodchild, M.F., Egenhofer, M.J., Fegeas, R., Kotman, C.A. (eds.) Interoperating Geographic Information Systems, pp. 5–29. Kluwer Academic Publishers, Norwell (1999)
6. Ghawi., R.: Ontology-based Cooperation of Information Systems, 15 March 2010
7. Sheth, A.P.: Changing Focus on Interoperability in Information Systems:From System, Syntax, Structure to Semantics. In: Goodchild, M., Egenhofer, M., Fegeas, R., Kottman, C. (eds.) Interoperating Geographic Information Systems, vol. 495, pp. 5–29. Kluwer Academic Publishers, Dordrecht (1999)
8. International Classification of Diseases (ICD). http://www.who.int/classifications/icd/en/
9. Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT). http://www.ihtsdo.org/snomed-ct/
10. NCI Metathesaurus. https://cabig.nci.nih.gov/tools/NCI_Metathesaurus
11. Klein, M.: Combining and relating ontologies: an analysis of problems and solutions. In: A. Gomez-Perz, M. Gruninger, H. Stuckenschmidt, and M. Uschold. (eds.) Workshop on Ontologies and Information Sharing, IJCAI 2001, Seattle, USA, pp. 309–327 (2001)
12. PONTE project. http://www.ponte-project.eu
13. Bizer, C., Cyganiak, R.: D2R server – publishing relational databases on the semantic web. In: Poster at the 5th International Semantic Web Conference (ISWC 2006) (2006). http://richard.cyganiak.de/2008/papers/d2r-server-iswc2006.pdf
14. Bizer, C., Seaborne, A.: D2RQ - treating non-RDF databases as virtual RDF graphs. In: ISWC 2004 (2004). http://www4.wiwiss.fu-berlin.de/bizer/pub/bizer-d2rq-iswc2004.pdf
15. Scharffe, F., Bruijn, J.A.: Language to specify mappings between ontologies. In: Proceedings of the 1st International Conference on Signal-Image Technology and Internet-Based Systems (SITIS 2005), Yandoué, Cameroon. Dicolor Press, November 2005
16. Expressive and Declarative Ontology Alignment Language (EDOAL), available at:http://alignapi.gforge.inria.fr/edoal.html