URL: https://ieeexplore.ieee.org/abstract/document/8679239
DOI: 10.1109/BIGCOMP.2019.8679239

A Novel Approach for Clinical Data Harmonization

Efthymios Chondrogiannis
National Technical University of Athens
9 Heroon Politechniou Str.
Athens, Greece
chondrog@mail.ntua.gr

Vassiliki Andronikou
National Technical University of Athens
9 Heroon Politechniou Str.
Athens, Greece
vandro@mail.ntua.gr

Efstathios Karanastasis
National Technical University of Athens
9 Heroon Politechniou Str.
Athens, Greece
ekaranas@mail.ntua.gr

Theodora Varvarigou
National Technical University of Athens
9 Heroon Politechniou Str.
Athens, Greece
dora@telecom.ntua.gr

***Abstract*— A city transforms into a "smart" one when it utilizes the data from smart devices, cloud infrastructures, applications and repositories in order to develop and provide new services, products and insights with the goal being to offer a safer, more efficient and robust environment for government, citizens and businesses, and accelerate sustainable economic growth. Although the variability across cities in terms of cultural background, demographics, current infrastructures, topology, among others, might drive significant variations in the individual goals of a smart city, the role of health in achieving the aforementioned goals is of paramount importance in any city. One of the starting points of introducing a health-related orientation into a smart city developments and operations is the intelligent processing of the overwhelming amount of clinical data being continuously generated during healthcare provision, daily activities and clinical research. Great challenges that need to be met in order to offer big data analytics over clinical data for the purposes of a smart city lie in their high heterogeneity at system, syntactic, structural and semantic level as well as their sensitive nature from the legal and ethical perspective, which may prevent and/or limit access to and analysis of the data. In this work we will present a novel approach for expressing clinical data using a common formalism with explicit semantics of the terms being used. Based on this established semantic ground, questions on health-related data can be placed for disease prevention, treatment monitoring, post-marketing surveillance, policy making, among others.**

Keywords*— *Data Harmonization, Heterogeneity Issues, Semantic Web, Clinical Studies

I. Introduction

In an effort to improve management of assets and resources and, consequently, improve the lives of citizens, boost business innovation and offer more opportunities for sustainable economic growth for society, businesses and government, cities worldwide are struggling to transform into smart cities. Meanwhile, the generation of a wealth of data at a dizzying pace further pushes the introduction of a data-driven approach in smart cities development and operations in an effort for the latter to leverage data to realize the full notion of a smart city.

Data generated in the broad clinical domain, including healthcare, daily activities, literature and clinical research, are of too high volume, too fast growing, highly heterogeneous and too complex for interested parties, including healthcare providers, policy makers, regional medical unions, government and citizens, to process and interpret with existing operations. Meanwhile, their great value, not only for well-targeted and effective healthcare provision but also disease prevention, resource planning, policy making and citizens' self-empowerment regarding their health and well-being, pushes the introduction and advancement of clinical big data analytics in the context of smart cities, following an extremely challenging multi-disciplinary task; the clinical data harmonization across the various data providers.

Achieving harmonized datasets of clinical orientation, although highly complicated in its very core, will allow for important advancements in the context of smart cities with irreplaceable benefits for all stakeholders, from citizens to healthcare providers to governments and business, and towards the innate goals of smart cities themselves, including quality of life, economic competitiveness, growth and sustainability, improved government operations. In fact, building advanced big data analytics on top of harmonized clinical-related datasets would set the basis for the realization of highly impactful services and processes. Apart of primarily health-related examples of the latter, such as deeper understanding of disorders and genes mechanisms, new opportunities for disease prevention and more effective treatment development, the expected impact is much broader. Hence, more robust postmarketing surveillance [1] based on detailed and accurate patient records across healthcare providers with direct benefit to the citizens (treatment efficacy, early treatment withdrawal in cases of safety issues), pharma (faster and less costly response to treatment safety alerts) and government (reduced healthcare costs due to treatment complications) could be achieved.

Policy (being health-related in the broader sense, from resource planning across healthcare providers to school and social initiatives to urban planning and infrastructures) design and development [2] would be more well-targeted, effective, influential and impactful. Meanwhile, policy impact assessment, being based on massive clinical data rather than solely on surveys and statistics, would allow for more effective policy monitoring, evaluation and revision, if necessary. Given that public expenditure on health and long-terms care in OECD countries has been rising over the past years and is expected to increase from around 6% of GDP today to almost 9% of GDP in 2030 and as much as 14% by 2060 [3], ICT solutions, and primarily big data and IoT, are expected to play a key role into finding a balance between reducing healthcare costs and improving citizens' health and quality of life. Moreover, clinical research priorities would be set on a more effective and efficient basis covering real patients' needs, facing new challenges promptly and being of higher impact for patients, society, government and economy.

Within this context, we have developed a Reference Model covering both the structure and the vocabularies of parameters related to demographics, diseases, treatments, laboratory tests, assessments based on questionnaires, among others, which aims at serving the backbone for the harmonization of clinical datasets which may be generated during healthcare provision as well as clinical research. The Reference Model and the accompanying semantic interlinking mechanisms have been developed within the context of the HarmonicSS project [4], which aims at bringing together and bridging the semantic gap among 23 different cohorts across Europe and the USA. Although the focus of this project lies primarily in patient stratification, disease progression, treatment monitoring and policy making regarding pSS (primary Sjögren Syndrome), the models and mechanisms have been developed in order to be applicable to other clinical domains, are extendable and adaptable.

In particular, and given also the strong legal and ethical restrictions posed due to the innate sensitive nature of this data, in this work we will present a semi-automatic approach for data harmonization purposes that enables the expression of patient data using a common formalism in a data-blind manner, i.e. the software engineers require no access to the actual patient data. In the core of this process is a Reference Ontology, which provides a meaningful, machine-processable description of the clinical parameters, as well as tools and mechanisms developed for specifying the correspondence among the terms used within each specific clinical dataset with the terms of the Reference ontology. Moreover, mechanisms which use these mappings for expressing the patient data based on the Reference ontology terms are presented.

The document is structured as follows. In section 2 we present related work in this field as well as existing algorithms and tools for correspondence detection and data transformation. In section 3 the approach followed along with the ontology developed is being presented. In section 4 we describe the tools and mechanisms developed for data harmonization purposes. In Section 5 we present how we have used these tools for expressing the data from a specific cohort based on the terms of the Reference ontology. A discussion follows in section 6 along with our next steps. Finally, in the last section we summarize the main points of this work.

II. Related Work

A. *Heterogeneity Issues*

There are many different ways to express concepts, observations and facts even in the same computer-based language (e.g., XML [5], JSON [6], Relational Database, RDF [7], etc.). In our case, the latter provide the syntax for organizing the data recorded about each patient. However, they do not provide any constraint about the meaning of terms being used (i.e., sequence of characters) in the name of the fields or their values, how these terms should be linked and/or what they actually represent. As a result, the representations of patient data across entities often show significant differences in the structure and meaning of the elements specified. For example, some entities record all the blood test results that belong to the same person in one table. Meanwhile, in other entities the same information may be distributed in several tables (e.g., one table per lab test). Also, for each laboratory examination the unit of measurement may be provided along with the value in the same field, or in two separate fields. Moreover, an interpretation of values in terms of being normal / abnormal may be given along with the range of normal values in many different ways. Furthermore, the terminology being used for each lab test (but also disease and drug) is not the same in all entities, given the variability of the possible terms for the same test (e.g., Hemoglobin and Haemoglobin both of which abbreviated to HG or HGB).

For mitigating the heterogeneity issues international organizations developing standards such as HL7 [8], CDISC [9], WHO [10] and FDA [11] have published several documents regarding the representation and exchange of patient data, including Reference Models, such as the CDISC Biomedical Research Integrated Domain Group (BRIDG) Model and HL7 Reference Information Model (RIM), and Controlled set of terms, such as the International Classification of Diseases (ICD) and Logical Observation Identifiers Names and Codes (LOINC) [12]. However, the poor adoption of standards by the different data providers as well as the differences existing among these standards perplex the representation and access to patient data, which continue to pose serious discrepancies.

B. *Tools and Mechanisms*

For alleviating the heterogeneity issues there is a plethora of systems and tools that focus on the alignment of terms rather than how a transition from one representation to the other can be achieved. OPTIMA [13] is a general purpose ontology alignment tool. It is equipped with a graphical environment for the presentation and analysis of ontology elements. Possible correspondences are automatically detected based on the name of terms and ontology structure which can be accordingly stored in an XML document. AgreementMaker [14] is another ontology alignment tool that presents the given ontologies in the form of a tree along with the suggested correspondence among their terms using syntactic and lexical comparison algorithms as well as a lexicon. Many other ontology alignment tools also exist that focus one on 1-to-1 correspondences while they may either provided a simple graphical user interface (GUI) or not provide a GUI at all [15].

For bridging the gap among different set of terms used across data providers for demographics, diseases, drugs and laboratory examinations, these tools can prove to be very useful since they enable users to specify 1-to-1 correspondences among their terms. Nevertheless, in some cases much more complicated correspondences may be necessary (e.g., a term has the same meaning with a combination of a few other terms). Also, existing algorithms /

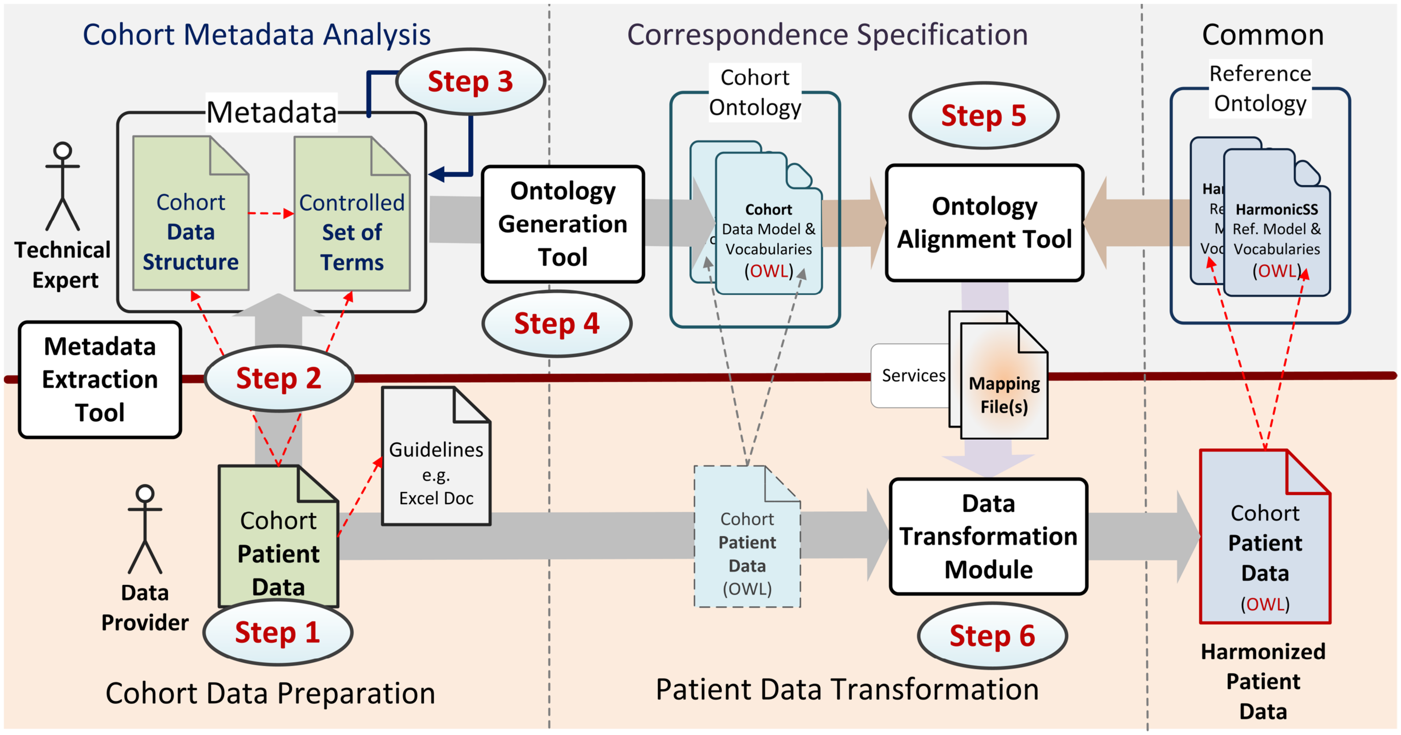


Fig. 1. Cohort Data Harmonization Process

tools / mechanisms for automatic correspondence detection can provide valuable results (mostly 1-to-1 correspondences) based on the sequence of characters being used in the description of each term (e.g., label, comments) as well as the axioms specified (e.g., hierarchy of terms, other constrains), which can be further improved with the use of external knowledge (e.g., a dictionary) [16]. These techniques can prove very useful when there is the need to specify the correspondence among thousands of terms and, hence, mapping them one-by-one would be impractical.

For harmonizing the data models of different data providers (even if an ontological representation of such models is generated so that these tools can be used) the above systems often cannot deal with the structural mismatches existing among them. For instance, the specification of a mapping rule that specifies an n-to-m relation where a data transformation is necessary is not covered. Some tools, like the Schema and Ontology Matching Tool, COMA++ [17], enable users to specify n-to-m correspondences among terms. However, when there is the need to specify the process that should be followed for moving from one representation to the other, such tools may be prove impractical or even fail, since several elements need to be introduced in each side of a mapping rule, while their relation needs to also be described in a procedural language. The above indicate that the correspondence specification so that moving from one representation to another is possible, but remains a challenging issue, especially when highly related yet isolated islands of knowledge have to be dealt with.

III. Approach Followed

A. *Harmonization Process*

The expression of patient data recorded within each data provider using a common language is the first step for allowing their use by big data mining mechanisms for any purpose within the context of a smart city, including policy making, treatment monitoring, disease prevention, patient stratification for improved treatment provision (e.g., find the percentage of women under 25 y.o. with elevated Hemoglobin value). For this purpose, we have developed an OWL ontology [18] (aka Reference Ontology) that specifies the parameters of particular interest for patients diagnosed Sjögren Syndrome. Moreover, we have created a system consisting of several tools (presented in the following sections) that enables data providers to express their data using the terms specified in the OWL ontologies through a semi-automatic process (Figure 1).

Each data provider should prepare in advance a document with the actual patient data (Cohort Data), albeit pseudonymized, following some general guidelines concerning the data format (e.g., an MS Excel document [19] with specific rows and columns). Accordingly, the data providers can use the Metadata Extraction tool developed (section 4.1) that processes the pseudonymized patient data document and extracts the parameters (data fields) being used for the description of the patients along with the different values identified for each field among the existing patients' data (controlled set of terms), which are then stored in the Cohort Metadata document. The automatically generated version of the Cohort metadata document would normally contain a considerable amount of empty fields (Figure 2). For this purpose, the technical experts should examine this document, and especially the meaning of the fields and their values, in close collaboration with the clinical experts (data provider's side). The analysis of Cohort metadata is an iterative process during which the technical experts need to closely work with the data provider regarding the data structure and meaning, and the terminologies being used for capturing the patient data until there is a common understanding of the elements mentioned in the metadata document, so that they can be accordingly mapped with the terms of the Reference ontology. For this purpose, when the analysis of the metadata is completed, the Ontology Generation tool can be used (section 4.1) which automatically creates an ontological representation of the Cohort metadata.

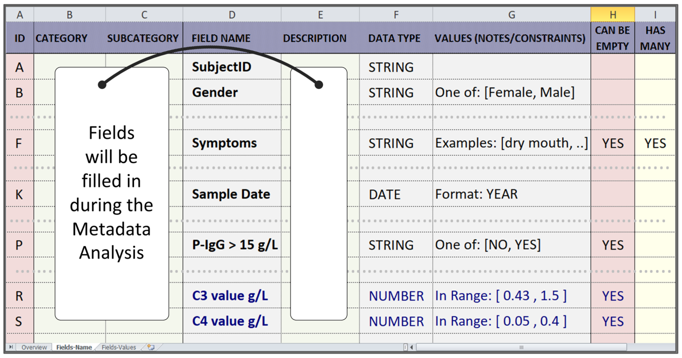
For specifying the correspondence among the terms of the Cohort and the ones existing in the Reference ontology, the Ontology Mapping Tool has been developed (section 4.2), which enables software engineers (in collaboration with clinical experts, if necessary) to bridge this gap through the specification of a few mapping rules (analogous to the number of the dataset fields). The mapping rules precisely determine the relation among the specified terms so that the initial data (so far being expressed based on the Cohort terms at the data provider's side) can be expressed through the Reference Ontology terms. For this purpose, a mapping rule specifies not only the relation among the entities of the two ontologies but also the process that should be followed when moving from one representation to the other. These mapping rules (when specified) are internally represented in a computer-based language (i.e., JSON) and can be, accordingly, used by another service for data transformation purposes (section 4.3).

It should be noted that the technical experts have only access to Cohort metadata which do not contain any personal data, while access to the latter have only the clinical experts, who can trigger the service for data transformation as soon as the mapping rules have been specified. The outcome of the latter process is an OWL document with the patient data expressed based on the Reference Ontology terms. This OWL document can then be imported in a relational database or even expressed in the form of an Excel document (if necessary) in a straightforward manner.

B. *Ontology Design*

The Reference Ontology (consisting of a Reference Model and Vocabularies) constitutes the core cohort representation with which any cohort data provided should be semantically aligned. It represents the cohort domain through a series of clinical-related parameters and their semantic relations (including classification and semantic linking) as well as their related vocabularies and value ranges.

The design of the Reference ontology was driven by clinical experts in HarmonicSS who described the mandatory and optional parameters of interest for each patient diagnosed with Sjögren syndrome. For each one of the parameters, relevant web sources were examined in order for its meaning to be precisely understood. Accordingly the parameters were organized in broader categories based on the domain they cover, such as Lab Test, Questionnaires, Symptoms, etc. Then for each of such categories the mandatory and optional properties were specified as well as the possible set of terms for each of the properties. For instance, regarding the smoking status of a person, the parameters of interest are the tobacco consumption status (i.e., current smoker, ex-smoker, never-smoker), the amount of cigarettes the person consumes per year (an optional parameter which is applicable in those cases when a person is an active smoker) and the date that these data were recorded. Eventually, the classes of data were organized in the following 7 categories [20]: Demographics, Lifestyle, Gender-specific conditions, Examinations, Medical Conditions, Interventions, Other.



A	B	C	D	E	F	G	H	I
ID	CATEGORY	SUBCATEGORY	FIELD NAME	DESCRIPTION	DATA TYPE	VALUES (NOTES/CONSTRAINTS)	CAN BE EMPTY	HAS MANY
A			SubjectID		STRING			
B			Gender		STRING	One of: [Female, Male]		
F			Symptoms		STRING	Examples: [dry mouth, ..]	YES	YES
K			Sample Date		DATE	Format: YEAR		
P			P-IgG > 15 g/L		STRING	One of: [NO, YES]	YES	
R			C3 value g/L		NUMBER	In Range: [0.43 , 1.5]	YES	
S			C4 value g/L		NUMBER	In Range: [0.05 , 0.4]	YES	

Fields will be filled in during the Metadata Analysis

Fig. 2 Cohort Metadata Document - Automatically extracted metadata part

According to the OWL ontology developed, a person is linked with zero or more classes of data, such Gender and Ethnicity (belong to Demographics), Smoking Status (belongs to the Lifestyle), Pregnancy Data (belong to Gender-specific Conditions), Blood or Urine Test and Questionnaires (belong to Examination), Symptoms (belong to Medical Condition) and Drugs Prescribed (belong to Interventions) as well as data about Family History and potential participation in another Clinical Study (belong to Other). Each of the categories has its own properties, the value of which is often derived from a controlled set of terms which have also been specified, taking into account the terms of particular interest to clinical experts as well as the ones specified in existing classification systems and codifications, including SNOMED CT [21], ATC [22], LOINC, etc.

IV. Tools and Mechanisms

A. *Metadata Extraction and Ontological Representation*

The role of the Metadata Extraction Tool is to allow for the automatic extraction of the dataset metadata, including structure (parameters' names) and vocabularies/value ranges. Apart from disengaging the Data Provider from the task of preparing the metadata files for their datasets, this tool also performs an early automatic check of the dataset in terms of structure and alignment with the cohort preparation guidelines.

For metadata extraction purposes, the user should initially select the cohort document (including all patient data) and accordingly press the metadata extraction button, which produces a new Cohort metadata document (Figure 2). The latter only includes the names of the cohort fields and recorded values for each of them, on condition that these values are not numeric or a date. The Metadata Extraction Tool collects the data recorded for each patient in the corresponding field, and then finds the different terms recorded for each one of them (especially in case of terms), the data type the data belong to (e.g., String, Integer, Date) along with the format of data (e.g., in case of a Date). For facilitating the analysis of parameters, the tool also calculates the range of the values (in case their value is a number) or widely used terms (in case their value comes from a controlled set of terms).

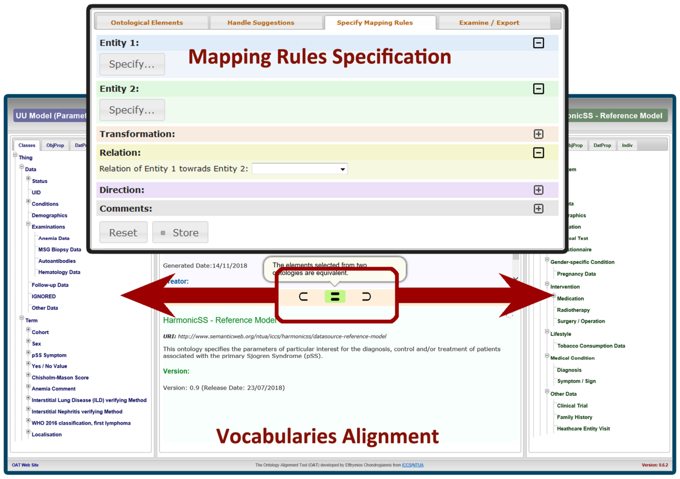
In the example presented in Figure 2, apart from the values in columns A and H, the values of all other fields can be modified, if necessary, as part of the metadata finalization process. In many cases a description of the meaning of the fields (Column E) is of great use for semantic analysis purposes. Also the fields may be organized in broader categories (Two levels supported – columns B and C). Further modifications can be applied. For example, the name of a field (column D) can be altered by providing the full name of a term rather than its abbreviation or vice versa. Also the Datatype of a property (Column F) can be changed and/or notes about the value of a parameter (Column G) can be added.

The main functionality of the Ontology Generation Tool is to automatically transform the metadata files into the respective ontological representation. In the generated ontology, information is organized in two separate categories: (Patient) Data and (Controlled set of) Terms. Parameters which are highly related are placed in the same category. All the information specified in the Cohort metadata document is included in the definition of each parameter. Also, in cases in which the value of a parameter comes from a controlled set of terms, the relevant class that can be used to provide the possible terms can be specified.

B. *Ontology Mapping Tool*

The main functionality of the Ontology Mapping Tool (which is an extension of the Ontology Alignment Tool [23]) is to allow for the user-friendly and efficient specification of the correspondence among the terms of two given OWL ontologies. It enables users to upload source and target ontologies and, accordingly, bridge the semantic gap among their terms. Users can easily specify 1-to-1 correspondences through the main panel, which simultaneously presents the definition of ontological elements from both source and target ontologies (Figure 3). Also, users can specify more complicated correspondences through the instantiation of the appropriate Ontology Patterns [24] via the highly interactive environment developed. The system further suggests possible correspondences among the terms of the given ontologies (mainly 1-to-1 suggestions) and enables users to examine (and potentially modify) the correspondences specified. Finally, users can export the correspondences in one of the supported formats, as appropriate, e.g., JSON for further processing by another module or HTML for presentation and validation purposes.

In case of vocabularies alignment, the mapping rules are expected to be quite simple (mainly 1-to-1 correspondence) and can be easily specified by manually accepting/rejecting the automatically suggested mapping rules as well as manually specifying those missing. However, in the case of the data / reference models, much more complicated correspondences are expected to be met. In this case the appropriate ontology pattern can be used and then the entities as well as the appropriate data transformations (if necessary) can be specified



Relation of Entity 1 towrads Entity 2:

Generated Date:14/11/2018

The elements selected from two ontologies are equivalent.

URI: http://www.semanticweb.org/ntua/iccs/harmonicss/datasource-reference-model

This ontology specifies the parameters of particular interest for the diagnosis, control and/or treatment of patients associated with the primary Sjogren Syndrome (pSS).

Version: 0.9 (Release Date: 23/07/2018)

The Ontology Alignment Tool (OAT) developed by Efthymios Chondrogiannis from ICCS/NTUA.

Fig. 3. Ontology Mapping Tool

as described in our previous work [23]. This process can be particularly difficult and time consuming when there are significant differences among the terms of the source and target ontologies, as is the case in the mapping rules specified among the terms of the Cohort Ontology (which often consist of a list of parameters just grouped in a few categories) and the Reference Ontology (which provides a real conceptualization of the domain). For facilitating the correspondence specification process, lists of correspondence patterns have been specified that enable users to specify:

(i) The parameters participating in the left side of the mapping rule (i.e., one or more data type properties)

(ii) The mandatory and optional parameters of a specific type of data (i.e., OWL class) on the right side

(iii) The process that should take place when moving from one world of knowledge to the other one, through the use of built-in functions as well as custom-made services (in case, the existing functions cannot adequately cover the existing needs).

The correspondence specification through the use of predefined correspondence patterns would be clearer in the example presented in section 5.2. It should be noted that in many cases (especially in case of n:m correspondences), the implementation of the appropriate web service (if not already provided) that performs the required data transformation is essential. The Ontology Mapping Tool enables users to specify the data transformation that should take place in a semi-formal way (i.e., URI, input, output, parameters, and description). However, the actual implementation of the service should be in a procedural language and the functionality should be available in the form of a SOAR or REST service.

C. *Data Transformation Mechanism*

The main functionality of the Data Transformation Mechanism is to automatically express the Cohort data based on the terms of the Reference Ontology. In the background, the mapping rules specified through the Ontology Mapping Tool as well as the specific services (as specified within the mapping files) implementing the changes that should take place in the parameter values when moving from one side (cohort) to the other (common) are used. The outcome of the Data Transformation Mechanism is an OWL document with the ontological representation of the cohort data based on the terms specified in the Reference ontology.

The system/service initially reads the dataset and detects the parameters recorded for each patient. Accordingly, it examines the mapping rules specified as well as the parameters (i.e., fields of the initial patient data document) participating in the left side of the mapping rule for detecting the Mapping Rules that can be fired. Each mapping rule produces the appropriate class instance based on the data type specified in the right side of the mapping rule along with the properties being mentioned. For calculating the values of these properties, the data transformation mechanism uses the service included, which provides the values of these parameters using the actual data of each specific patient (i.e., the values of the properties mentioned in the left side of the mapping rule) as well as the attributes provided (e.g., unit of measurement, normal range of values) during the mapping rules specification. The above process is being repeated until no other mapping rule can be fired. Then system continues with the transformation of the Cohort data of the next patient. When this process completed all patient data are being expressed using the terms of the Reference ontology.

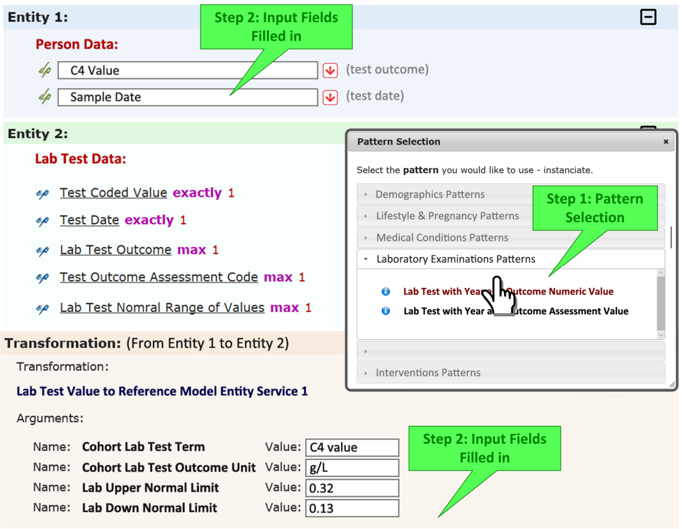
V. Example of Use

A. *Data Representation and Analysis*

For the Data Harmonization process, the Cohort data provider (Institute A) was initially asked to prepare a document (MS Excel in their case) with the data of 200 patients diagnosed with Sjögren Syndrome, based on predefined specifications given to them. For each patient, the spreadsheet contained information about 150 different parameters, including the patient ID (placed in the first column for each patient), symptoms, lab tests, biopsies, etc. Consequently, the excel document consisted of 201 rows (the first row contained the names of the parameters) and 150 columns.

Following that, the data providers used the Metadata Extraction tool for extracting the name of the fields and their possible values, which were stored in another Excel document sent to the technical experts (a small excerpt is presented in Figure 2). The data in this document were carefully examined and a definition of the terms was introduced along with the classification of the terms in broader categories. More precisely, for each term all the information of interest for the proper interpretation of its recorded values was collected. For example, in case of a laboratory examination taking numerical values, apart from the definition of the entity being measured and the technique being used (especially if being more than one), the unit of measurement and the normal range of values for the specific lab were collected and recorded. It should be noted that the description of the field was introduced by technical experts (and verified by clinical experts) but the values of the collected information of interest were provided by the corresponding data provider.

When the data analysis was completed, the Ontology Generation tool was used for transforming the document with the Metadata in the form of an OWL ontology that contained all of the recorded information. This step was essential so that the description of entities could be loaded in the Ontology Mapping Tool and then the relation of the terms with the ones specified in the Reference ontology could be specified. It should be noted that the automatic metadata extraction and their ontological representation (after their analysis) was feasible since both of them followed a predefined format.


Step 2: Input Fields Filled in
Step 1: Pattern Selection
Step 2: Input Fields Filled in
Pattern Selection
×
Demographics Patterns
Lifestyle & Pregnancy Patterns
Medical Conditions Patterns
Laboratory Examinations Patterns
Interventions Patterns

Entity 1:

Person Data:

C4 Value (test outcome)

Sample Date (test date)

Entity 2:

Lab Test Data:

Test Coded Value **exactly** 1

Test Date **exactly** 1

Lab Test Outcome **max** 1

Test Outcome Assessment Code **max** 1

Lab Test Nomral Range of Values **max** 1

Select the **pattern** you would like to use - instanciate.

Lab Test with Year and Outcome Numeric Value

Lab Test with Year and Outcome Assessment Value

Transformation: (From Entity 1 to Entity 2)

Transformation:

Lab Test Value to Reference Model Entity Service 1

Arguments:

Name: **Cohort Lab Test Term** Value: C4 value

Name: **Cohort Lab Test Outcome Unit** Value: g/L

Name: **Lab Upper Normal Limit** Value: 0.32

Name: **Lab Down Normal Limit** Value: 0.13

Fig. 4. *Mapping a Laboratory Examination with Reference Ontology Terms*

B. *Data Harmonization*

For bridging the gap among the Cohort Metadata and Reference Ontology terms we specified several mapping rules (almost 120) through the instantiation of the appropriate correspondence patterns. It should be mentioned in advance that more than one parameter (from the Metadata) is participating in the left side of the mapping rule. Also, it was not feasible for some parameters to be linked with the terms of the Reference Ontology since they had completely different meaning. In the rest of this section will be presented an example of how the mapping tool was used for specifying one Mapping Rule as well as how this mapping rule was used for the transformation of the corresponding patient data.

For expressing the Lab Test Data using the Ontology Terms, the users initially chose to instantiate the Correspondence Pattern implemented for this purpose (i.e., one out of several patterns implemented). Figure 4 presents a screenshot from the Ontology Mapping Tool for specifying the correspondence of a specific laboratory examination (i.e., C4 value) with the corresponding Reference Ontology terms. In the upper part of this figure (Entity 1) the two datatype properties from the Cohort OWL ontology can be seen. The first property is the “C4 value” and the second one is the “sample date” (i.e., year) when the blood sample was collected. On the middle part of this figure (Entity 2) the mandatory and optional properties specified in the Reference Model for capturing the data of a specific Laboratory Examination are to be seen. For each lab test, the code (in our case, the code for C4), the sample date (i.e., the one provided), the outcome value (based on the one provided) and the assessment code that indicates whether the value is within the normal range of values or not need to be specified.

For precisely determining the correspondence among these terms, the procedure that should be followed for expressing the data based on the Reference Ontology terms should be additionally specified. For this purpose, the implementation of a function/service that receives as input the data specified in the left side of the mapping rule (i.e., C4 value and year of date) and produces the values of the parameters presented in the right side (i.e., lab test code, date sample retrieved, outcome amount and assessment code, and normal range of values) is necessary. In this example, the lab test code (as in the Reference Ontology) can be found based on the specific test (i.e., C4 value). Also the year of sample date can be directly found based on the year provided. Additionally, the amount can be calculated based on the specific value (real number) recorded for each patient and the unit of measurement (which is always the same i.e., g/L for all the patients) also considering the preferred unit of measurement of each lab test (specified in the Reference Ontology). Finally, whether the outcome was normal or not can be extracted by taking into account the normal range of values (which should exist in the definition of the C4 value parameter).

Since the above process should be also applied in many other laboratory examinations (and thus reused), the service, (as described above) that receives as input two properties and produces the values of the five properties mentioned, taking into account a) the local name of the lab test (e.g., C4 value), b) the unit of measurement (e.g., g/L) and c) the normal range of values (i.e., upper and lower normal limits), if applicable, was developed and included as part of a correspondence mechanisms repository.

In the document of the patient data (Figure 2), for each patient, the C4 value resides in column S and the year the blood sample was collected in column K (this information also exists in the ontological representation of these parameters). The system examines the mapping rules specified and detects that there is a mapping rule that uses the two aforementioned parameters. Then, it introduces an instance of a “Laboratory Examination” and specifies that this instance has the following five parameters: a) Lab Test Code, b) Sample Date, c) Outcome Amount, d) Outcome Assessment and e) Normal Range of Values.

For producing the values of these parameters the data transformation mechanim uses the service included in the definition of the mapping file. This service detects the Lab Test (i.e., C4 levels in blood, a term from the Reference ontology) based on the name specified and produces an instance of a Date with the year provided as well as an Amount with the appropriate value (in this case, exactly the same with the given one since mg/dL is the same as g/L) with the specific unit recorded for the Reference Ontology Lab Test (i.e., mg/dL). Moreover, it examines whether this value is normal or not and introduces the appropriate terms and finally provides an entity that expresses the normal range of values based on the Upper and/or Lower Normal Limits provided.

Following the above process, the appropriate entities are introduced based on the rest of the parameters specified, including Demographics, Lifestyle Data (i.e., Tobacco Consumption Status), Pregnancy Data, Laboratory Examinations, etc. All the entities introduced based on the data of a row of an Excel File are also linked with the specific patient which is being uniquely determined based on their id. It should be noted that the above process was fully automated and initiated by the clinical experts as soon as the mapping rules specification had been finalized, as presented at the beginning of this section.

VI. Discussion and Next Steps

In our work, the clinical data were provided in the form of an Excel document following some general guidelines, so that they could be processed through one or more software modules (i.e., metadata extraction and data transformations). This format was selected by the clinicians given that most of the datasets available for harmonization were already in this format. However, in several cases the parameters recorded for each patient were not well-organized, which led to various difficulties during the data harmonization process. For example, there were several instances of the same medical test for the same patient in case it had been performed more than once in the past (e.g., biopsy 1, 2, 3, etc.) and hence it would be necessary to specify several mapping rules for the same test. Also, in case more than one terms were included in the same field (e.g., a field containing all the symptoms of a patient) it was difficult (during the metadata extraction process) to automatically detect the multiple values, especially when the terms were not separated by a common delimiter across the dataset or even the specific field values. In this case, clinical experts were requested to provide additional information about these fields. Moreover, since the data were provided in the form of a flat list it was difficult to organize it in broader categories and especially link it with the corresponding elements specified in the Reference ontology. For instance, there were data about several different lab tests and, hence, (during the correspondence specification process) several mapping rules (one for each lab test) had to be introduced, even if all of them followed the same approach. Nevertheless, the correspondence among these terms and the ones specified in the Reference ontology was specified accurately.

It should be noted that the approach followed would remain the same even if the data were provided in a different file format. In this case the Metadata Extraction and Ontological Representation tools should be updated so that they could process the data, according to the specific file format. In fact, in case of a Relational Database an Ontological Representation can be directly exported based on the Database Schema [25] and the controlled set of terms (often stored in separate tables in the database). In this case, the correspondence specification with the Reference Ontology terms would be easier, since the mapping rules are expected to be limited in number.

Concerning the patient data, ideally it should include detailed information such as the exact drug dosage along with the specific period of time of administration, the exact result of each laboratory examination and when it was performed, information about clinical studies in which the patient participated in the past, etc. However, the above information may not be available. For instance, the exact drug or active substance prescribed may not be known, but rather just its broader category (e.g., Disease-Modifying Arithmetic Drug, abbreviated to DMARD). Also, the exact date an event occurred (e.g., blood sample collection) may not be known, but rather the year or even decade. For this purpose, during the design of the Reference ontology the appropriate terms were introduced so that the above scenarios can be covered. Consequently, the patient data may be expressed by using an already specified term/category or another semantically-related term. Regarding temporal information, the exact date of an event or the period of time when it happened (e.g., before the date that the data was recorded) can be specified. The more specific the patient data are the more questions can be answered and in a more accurate and reliable manner. For example, in case a person received a DMARD before a specific date, a question that could not be answered is whether the person received Methotrexate (i.e., one of the most commonly used DMARDs) or whether the person received a DMARD before a specific date which is prior to the one specified. What can be answered is whether the person received a DMARD in the past.

VII. Conclusion

The plethora of heterogeneity issues along with the sensitive nature of patient data renders the harmonization of and big data analytics on clinical data a rather challenging issue and poses limitations to the services that can be offered within the context of a smart city for citizens, healthcare providers, business and government. In this work we have presented an ontology-based approach for the expression of patient data using a common language based on the correspondence specified among the Reference Ontology terms and the ones used for the expression of patient data at each data provider. During this process, we have developed and used several tools that facilitate the analysis, mapping and transformation of patient data in the form of an OWL ontology. It should be noted that, for avoiding erroneous interpretation of patient data the analysis and, especially, the correspondence specification process should be driven by clinical experts who should actively participate in this process.

Acknowledgement

This work has been partially funded by the European Commission's activity of the Horizon 2020 project HarmonicSS under contract number 731944. We would like to offer our special thanks to the clinical partners of the HarmonicSS for their guidance in developing the Reference Model. 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] C.L. Ventola, Big Data and pharmacovigilance: data mining for adverse drug events and interactions, Pharmacy and Therapeutics, vol. 43, no. 6, 2018, pp. 340-351.

[2] D.W. Bates,A. Heitmueller, M. Kakad, and S. Saria, Why policymakers should care about "big data" in healthcare, Health Policy and Technology, 2018, pp. 211-216.

[3] Fiscal Sustainability of Health Systems: Bridging Health and Finance Perspectives, OECD Publishing, Paris, 2015. doi=https://doi.org/10.1787/9789264233386-en.

[4] HARMONIzation and integrative analysis of regional, national and international Cohorts on primary Sjögren's Syndrome (pSS) towards improved stratification, treatment and health policy making (HarmonicSS), available at https://www.harmonicss.eu/

[5] Extensible Markup Language (XML), available at https://www.w3.org/XML/

[6] JavaScript Object Notation (JSON), available at https://www.json.org/

[7] E. Miller, An Introduction to the Resource Description Framework, Bulletin of the American Society for Information Science and Technology, vol. 25, no. 1, 1998, pp. 15-19.

[8] B. Smith and W. Ceusters, HL7 RIM: An Incoherent Standard, Studies in health technology and informatics, vol. 124, 2006, pp. 133-8.

[9] T. Souza, R. Kush, and J.P. Evans, Global clinical data interchange standards are here!, Drug Discovery Today, vol. 12, no. 3, 2007, pp. 174-181.

[10] World Health Organization (WHO), available at http://www.who.int/

[11] Food and Drug Administration (FDA), available at https://www.fda.gov/

[12] Logical Observation Identifiers Names and Codes (LOINC), available at https://loinc.org/

[13] R. Kolli and P. Doshi, OPTIMA: Tool for Ontology Alignment with Application to Semantic Reconciliation of Sensor Metadata for Publication in SensorMap, In Proceedings of the IEEE International Conference on Semantic Computing, Santa Clara, CA, 2008, pp. 484-485.

[14] I.F. Cruz, F.P. Antonelli, and C. Stroe, Agreementmaker: Efficient Matching for Large Real-World Schemas and Ontologies, In Proceedings of the VLDB Endowment, vol. 2, 2009, pp. 1586-1589.

[15] P. Shvaiko and J. Euzenat, Ontology Matching: State of the Art and Future Challenges, IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 1, 2013, pp. 158-176.

[16] M. Granitzer, V. Sabol, K. W. Onn, D. Lukose, and K. Tochtermann, Ontology Alignment—A Survey with Focus on Visually Supported Semi-Automatic Techniques, Future Internet, vol. 2, no. 3, 2010, pp. 238-258.

[17] D. Aumueller, H.H. Do, S. Massmann, and E. Rahm, Schema and ontology matching with COMA++, In Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, Baltimore, Maryland, 2005, pp. 906-908.

[18] B.C. Grau, I. Horrocks, B. Motik, B. Parsia, P. Patel-Schneider, and U. Sattler, OWL 2: The next step for OWL, Journal of Web Semantics, vol. 6, no. 4, 1008, pp. 309-322.

[19] Microsoft Excel, available at https://products.office.com/el-gr/excel

[20] E. Chondrogiannis, V. Andronikou, A. Tagaris, E. Karanastasis, T. Varvarigou, and M. Tsuji, A novel semantic representation for eligibility criteria in clinical trials, Journal of Biomedical Informatics, vol. 69, 2017, pp. 10-23.

[21] Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT), available at https://www.nlm.nih.gov/healthit/snomedct/

[22] Anatomical Therapeutic Chemical (ATC) Classification System, available at https://www.whocc.no/atc_ddd_index/

[23] E. Chondrogiannis, V. Andronikou, E. Karanastasis, and T. Varvarigou, An intelligent ontology alignment tool dealing with complicated mismatches, In Proceedings of the 7th International Workshop on Semantic Web Applications and Tools for Life Sciences, Berlin, Germany, 2014.

[24] O. Šváb-Zamazal, V. Svátek, F. Scharffe, and J. David, Detection and Transformation of Ontology Patterns, Communications in Computer and Information Science, vol. 128, 2011, pp. 210-223.

[25] R. Gwani, and N. Cullot, Database-to-Ontology Mapping Generation for Semantic Interoperability, In Proceedings of the Third International Workshop on Database Interoperability, 2007.