Bridging the Gap among Cohort Data Using Mapping Scenarios

Efthymios Chondrogiannis, Efstathios Karanastasis, Vassiliki Andronikou, and Theodora Varvarigou
National Technical University of Athens, Athens, Greece
Email: {chondrog, ekaranas, vandro}@mail.ntua.gr, dora@telecom.ntua.gr

***Abstract*—Disease-specific Cohort data across different healthcare and clinical research entities is of paramount importance for the study of the particular disorder and the development of new clinical and health policies. However, the significant structural and semantic mismatches across these data stemming from their independent development prevent their computer-based processing. The formal expression of the individual Cohorts using a common formalism (Data Harmonization) is the means for producing valid and accurate results, especially for diseases affecting a small percentage of the population, such as is the primary Sjögren's Syndrome (pSS), in which case data analytics on an individual cohort may lead to results not easily generalizable and of low accuracy and trust-worthiness. In this work, the approach followed in the HarmonicSS project for bridging the gap among eight heterogeneous Cohorts from eight different Cohort providers is presented, which was based on the software-aided analysis of their individual data structure and terminology. One of the outcomes of this process was a number of reusable parameterizable correspondence patterns (named Mapping Scenarios) that were accordingly instantiated for the accurate and complete mapping of the Cohort data to the Reference Model elements. The mapping scenarios were incorporated in a Visual Mapping Tool, which was developed for facilitating their use from both ICT experts and non-expert users.**

***Index Terms*—mapping scenarios, correspondence patterns, cohort study, data harmonization, semantic web**

I. Introduction

The patient data collected across different healthcare and clinical research institutes (Cohort Data) can provide valuable insight about the disorder's underlying mechanisms as well as tangible evidence about the effect of a suspected risk factor. For producing valid and unbiased results when analysing the recorded patient data, a large pool of patients should be examined. Especially in the case of not-so-common diseases that affect a limited patient population globally, such as the primary Sjögren's Syndrome [1], it is highly beneficial to use cohort data from different entities, since it boosts the generalizability of the study outcomes.

The ICT-enabled processing, analysis and mining of the cohort data collected and maintained by different institutes is rather challenging due to the legal and ethical implications stemming from the sensitive nature of patient data and the significant structural and semantic mismatches among the data sets. Existing algorithms and tools (presented in Section II) can alleviate these problems by automating the terminology alignment process to a great extent. Nevertheless, for bridging the gap among them, the process for moving from one data representation to the other one should be specified, which is highly affected by the format and structure of the data in each particular cohort as well as additional information that is often essential for the correct interpretation of the data recorded (e.g., normal range of values of lab tests, which are different across laboratories).

In the HarmonicSS project [2], for the expression of the data collected across 8 different institutes using a common formalism (Data Harmonization), a Reference Model was initially designed. Accordingly, mechanisms and tools were developed that facilitate the expression of cohort data using the Reference Model terms in a data-blind manner, i.e., without the patient data being exposed to other entities apart from the data owners and providers. For this purpose, in advance, a software-based analysis of the cohort data took place that was based on the metadata and their linking with the elements specified in the Reference Model. The outcome of this process was a list of different mapping scenarios that may be encountered (i.e., data structures that point to a limited amount of Reference Model classes) across the 8 cohorts, which were then used for bridging the gap among the cohort parameters and the Reference Model terms, based on a software Tool that was also developed.

This paper is structured as follows. In Section II, related work regarding the heterogeneity issues that may be encountered along with the state of the art algorithms, techniques and tools that can be used for alleviating these differences are being presented. In Section III, the semi-automatic process followed for the classification and analysis of data residing in 8 different cohorts is being described as well as the process that should be followed for the expression of such data using the Reference Model. The Mapping Scenarios detected along with a Mapping Tool developed are being presented in Section IV. Relevant issues and next steps are being discussed in Section V. Finally in the last section, our work is summarized.

Manuscript received November 28, 2020; revised April 2, 2021.

doi: 10.12720/jait.12.3.179-188

II. Related Work

A. Data Representation

For the interoperable representation and exchange of patient data there is a considerable amount of work available on web by international standards development organization such as the Clinical Data Interchange Standards Consortium (CDISC) [3] and the Health Level Seven International (HL7) [4]. The standards published by these organizations guide the design of the underlying data structure (e.g., data types specification based on a Reference Information Model - RIM) as well as the methods or protocols used for accessing the data (e.g., a query language or another message exchanged protocol). These standards can be also combined with terminologies published by other standardization bodies about the conceptualization of a particular domain of interest, such as the International Statistical Classification of Diseases [5], the Anatomical Therapeutic Chemical Classification System (ATC) [6] and the Logical Observation Identifiers Names and Codes (LOINC) [7]. Nevertheless, the use of the aforementioned standards across healthcare and clinical research centers is often limited while different codifications, classifications and vocabularies exist for the same concepts (e.g., diseases, drugs, etc.), which perplexes the uniform computer-based processing of their data, since a variety of heterogeneity issues should be dealt with.

The heterogeneity issues are classified in two broad categories, i.e., conceptualization and explication mismatches [8]. In the first category lie the differences related with the conceptualization of a particular domain, which is also depicted in the definition of the relevant entities. For instance, in one data source information about the pharmaceutical drugs prescribed along with the dosage plans may reside, whereas in another data source dosage information may have not be included. Also, in one data source the specific drug prescribed may be recorded (e.g., amiodarone), whereas in another data source only the broader category that the drug belongs to (e.g., anti-arrhythmic drug). Explication mismatches arise from the different ways that the same conceptualization has been specified. For instance, the drug administered along with the dosage plan may be encoded in one field in one case, whereas in another one, the same information may be scattered in two or three separate fields (e.g., one for the drug, another one about the amount of dosage and a third one about the frequency of administration). Also, the names and/or codes used about drugs (as well as other concepts, such as diseases) may be different; a drug can have several trade names, abbreviations and even codes [9], especially when the latter do not stem from an international coding system.

B. Algorithms and Tools

A considerable number of algorithms and techniques [10] exist that can automatically detect possible correspondences among the terms of two different data source and, hence, bridging the gap among them. String-based techniques search for potential matching among the terms of two ontologies based on the sequences of characters being used. For instance, the Edit Distance algorithm counts the minimum number of changes being necessary for transforming one string to another one. Language-based techniques take into the account the language being used for the expression of each term as well as the internal components of each one of them. Stop words elimination (e.g., articles and prepositions) is commonly being used as well as a stemming algorithm in order to get rid of the morphological variations of a term (e.g., presence of a word in singular or plural form). The axioms specified for each ontology term (either explicitly or implicitly through a reasoner), such as the data type constrains and classification of terms, can be taken into consideration in the matching process. External Knowledge such as a general purpose lexical database or a domain specific treasure (e.g., Medical Subject Headings - Mesh [11]) can be also used for matching ontology terms, especially in cases when the previous techniques fail. Machine learning techniques attempt to match ontology terms based on the statistical distribution of features derived from each concept (including their label and axioms specified).

Mapping tools [12] often utilize a combination of the previously mentioned matching techniques which can provide quite satisfying results with the overall f-measure being 0.86 (in the best case) [13]. The outcome of the aforementioned techniques can be further improved with the active participation of the domain experts, who can review the suggested correspondences as well as specify new ones [14]. This is often necessary, particularly in cases when precise and complete mapping of the two ontologies is needed (at design time) so that it can be accordingly used for supporting relevant tasks (e.g., data integration). Hence, the Mapping Tool should be equipped with a user friendly environment that prompts expert users to actively participate in the mapping process and help them along the way. Existing systems either provide simple GUIs or do not provide a GUI at all [13].

Regarding the formal expression of the mapping rules, the Correspondence Patterns [15] enable users to deal with a variety of mismatches that may be encountered, including 1-to-1 correspondence as well as more complicated ones. For instance, Correspondence Patterns allow linking of two different properties in the form of a mapping rule even if a data transformation in their values should be applied (e.g., in case of monthly income expressed in different monetary systems). Still, when it comes to highly heterogeneous data sources in which data scattered along several ontology elements from both source and target ontologies should be combined, the correspondence patterns have proven to be cumbersome or insufficient, as they have been primarily designed for dealing with semantically overlapping domains rather than cases where semantically related concepts could be linked through a well-defined process. The Mapping Rules can be formally expressed using the Expressive and Declarative Ontology Alignment Language (EDOAL) [16]. This is a rather expressive mapping language that enables users to specify all the internal elements of each

mapping rule, including source and target ontology terms, the relation among them, along with the data transformation service(s) often necessary when moving from one data representation to the other one and vice versa. The specified Mapping Rules can then be stored in an XML document, which can be consumed by another software agent.

III. METHODOLOGY

A. Cohort Data Harmonization

For data harmonization purposes, we initially developed a Reference Model (RM) that specifies the parameters of particular interest for the patients diagnosed with primary Sjögren Syndrome (pSS) along with their terminology (e.g., drug prescriptions along with the specific drugs or active substances). Its design was driven by our past work in building a RM for patient data representation which was applied for the formal expression of eligibility criteria [17]. However, it was extended and properly linked with different coding systems (e.g., Symptoms and Codes) explicitly for patients diagnosed with pSS based on close interaction with clinical experts of the HarmonicSS project. The RM was published in the form of an OWL ontology [18] so that it can be used for the formal expression of cohort data.

Accordingly, a data-blind approach was followed that allowed Data Providers to express their Cohort Data using the RM terms (i.e., OWL individuals). For this purpose, the Data Providers initially prepared their cohort data following some general guidelines regarding the data structure, so that it could be further processed by two different software agents implemented; the Metadata Extraction and Data Harmonization tools [19]. The first module (Metadata Extraction) was used for automatically detecting the cohort parameters, along with their data type and possible value range (Fig. 1). The second module (Data Harmonization) was then used for automatically expressing the patient data residing in the initial cohort files using the RM terms based on the Mappings specified.

COHORT ID	COL. ID	COLUMN NAME	DATA TYPE	VALUES (CONSTRAINTS)	CAN BE EMPTY
CHRT_01	U	First visit (year)	DATE	Format: YEAR	YES
	CW	Arthritis (0-1)	INTEGER	One of: { 0 , 1 }	YES
	CX	Arthritis date(-yr)	DATE	Format: YEAR	YES
	FA	WBC Baseline	INTEGER		YES
CHRT_02	C	Inclusion year	DATE	Format: YEAR	YES
	N	Arthritis (0-2)	INTEGER	One of: { 0 , 1 , 2 }	YES
	BB	CRP	REAL NUMBER	In Range: [0 , 274]	YES
	BN	Anti-SSA	INTEGER	One of: { 0 , 1 , 2 }	YES
CHRT_03	G	Year of diagnosis	DATE	Format: YEAR	YES
	L	Xerostomia at diagnosis	INTEGER	One of: { 0 , 1 }	YES
	AH	Anti-Ro-SSA at diagnosis	INTEGER	One of: { 0 , 1 }	YES
	CM	IVIG	INTEGER	One of: { 0 , 1 }	YES
CHRT_04	FQ	Glucocorticoids	INTEGER	One of: { 0 , 1 }	YES
	FR	Glucocorticoids Year Start	DATE	Format: YEAR	YES
	FS	Glucocorticoids Year End	DATE	Format: YEAR	YES

Figure 1. Part of Metadata automatically extracted from four different cohorts.

For bridging the semantic and structural gap among the Cohort Fields and the RM terms, the Metadata automatically extracted by the software agent were analysed through an interactive, iterative process including both data providers and technical experts in order to capture the meaning of both the cohort Fields and their Values. For instance, the analysis indicated that the values "0" and "1" for the "CHRT_01" cohort parameter "Arthritis" stand for "no" and "yes" respectively, whereas in the "CHRT_02" cohort the values "0", "1", "2" stand for "current", "past" and "never". Then, the correspondence among the cohort parameters and the RM terms was precisely specified, in the form of several Mapping Rules, so that they could be used for the expression of Cohort Data using RM terms. For this purpose, in advance, a software based analysis of the cohorts' metadata took place (described in the following two sections) that highlighted the type of data residing in the given cohorts, the relation among them and especially the process that should be followed for the potential expression of the information residing in such fields using the RM terms. Since different data structures and patterns detected could be linked with the same RM elements even in the same cohort, different Mapping Scenarios were developed that were summarized in a document and formally expressed using JSON [20] so that they could finally be used for automatically mapping cohort fields with RM terms.

B. Terminology Alignment

For the meaningful description of the data captured by each particular cohort, the corresponding RM terms for both Fields and their Values were initially specified through a semi-automatic process. For this purpose, the Ontology Alignment Tool (OAT) [21] was used. This tool supports the whole mapping process by enabling users to upload source (cohort metadata) and target (reference model) ontologies, manage the automatically detected mapping rules (accept/reject them), manually specify those missing and, finally, export the mapping rules specified in the appropriate format. The service used in the background for detecting the similarities among the terms was revised and updated using a plethora of algorithms and techniques, including but not limited to, string matching techniques (e.g., Levenshtein Distance [22]), language based techniques (e.g., Porter Stemming Algorithm [23]), axioms specified in the OWL ontology (i.e., classification of the Reference Model terms) as well as additional knowledge obtained from the literature and incorporated in the Reference Model (including synonyms and abbreviations).

Fig. 2 presents the suggested correspondences for the "CHRT_03" fields (part of which is being presented in Fig. 1). The tool has automatically detected that the corresponding RM term for "IVIG" is the pharmaceutical drug "Intravenous Immunoglobulin" (long form). The tool has also detected that the corresponding RM term for "Xerostomia" is the symptom "Dry Mouth" as it has exactly the same meaning (specified in the RM). Moreover, the tool has detected that the term "Anti-Ro-SSA" is probably the same with the RM Lab Test "Anti-Ro/SSA". Nevertheless, some correspondences have not been automatically detected, since the criteria used for

similarity detection purposes were quite strict (hence, provided high precision but low recall) in order to avoid inappropriate suggestions that may confuse the end users during the mapping process.

The suggested mapping rules were accepted or rejected based on the similarity, being a real number in range [0, 1] – the greater the value is the more similar the terms are, calculated by the tool. For avoiding potential errors the suggested mapping rules with similarity below "0.9" were manually examined and were accepted or rejected using the buttons existing in the left and right side of each rule. Regarding the mapping of the remaining fields, the highly interactive graphical environment provided by this tool enables users to quickly specify 1-to-1 correspondences (e.g., mapping of values "0" and "1" with the RM confirmation terms) by selecting the corresponding terms and then pressing the appropriate button (e.g., equivalent terms) or introduce more complicated ones (e.g., the term Sicca stands for Dry Mouth and Dry Eyes) by instantiating the appropriate Ontology Pattern(s) [24] and then specifying the RM terms. In the second case, the offered auto-complete functionality speeded up the process and limited typing errors given the plethora of RM terms especially about drugs, diseases and lab tests.

Once the corresponding RM terms were specified, the Mapping Rules were exported in JSON format so that they could be further processed by a software agent. It should be noted that the whole terminology alignment process took less than 1 hour for each cohort provided that the metadata analysis had been completed and the appropriate clarification regarding the meaning of cohort fields and their values had been provided by the data providers. Also, 150 Mapping Rules were specified for each cohort on average, which was perfectly normal taking into account that there were about 120 fields in each cohort and their values were often a confirmation term. Nevertheless, not all of the existing cohort fields could be linked with the RM terms since, in some cases, there was no relevant entity.

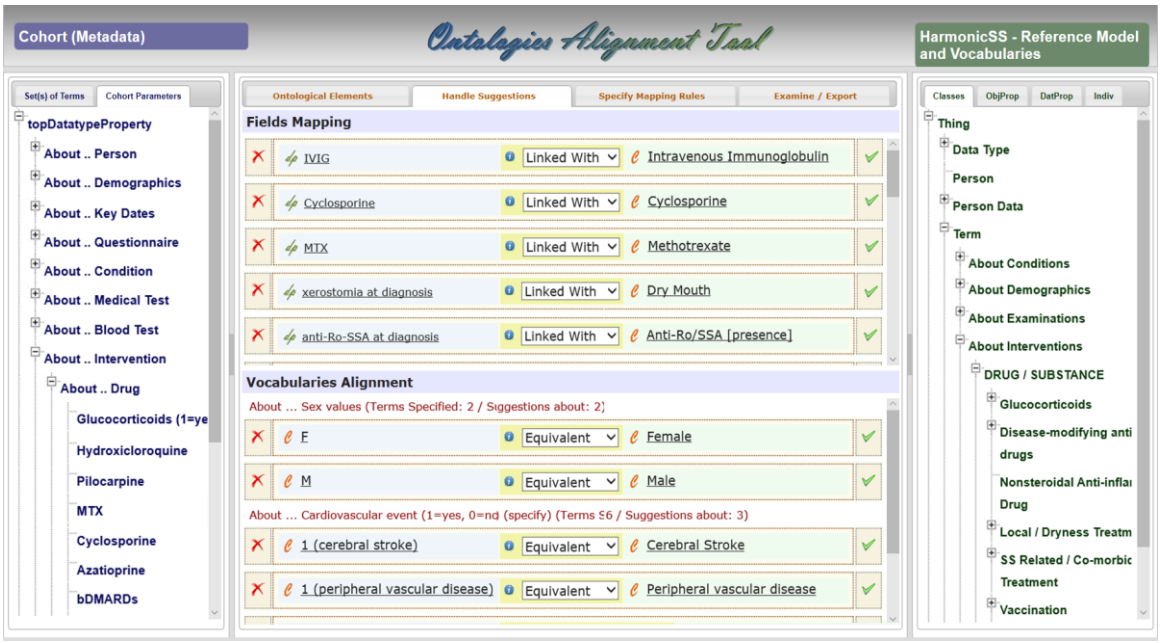

Figure 2. Accepting/Rejecting automatically detected mapping rules for "CHRT_03" cohort.

C. Patterns Detection and Analysis

The Cohort Metadata automatically extracted by the system and the Mapping Rules specified by the end user were further processed by another software system developed, which highlighted the type of data captured by each cohort (e.g., Medical Conditions) and especially the patterns used for the expression of such data (e.g., the Value of a Field about a particular Medical Condition is a Confirmation Term). For this purpose, the system was based on the classification of the RM terms for detecting the broader category of both cohort fields and their values, especially when the later came from a controlled set of terms. For instance, the system automatically detected that the "CHRT_03" Field "CM" (Fig. 1) refers to a pharmaceutical drug and its value is always a confirmation term (if not being empty), since the corresponding RM terms for both field and its value had been already specified (terminology alignment). Taking into account the fact that in some cases, the same RM terms were presented in more than one Fields, more complicated patterns were also detected. For instance, the system automatically detected that two separate "CHRT_01" fields (i.e., Fields "CW" and "CX") refer to the same Medical Condition (i.e., Arthritis). However, their values were different (in Field "CW" was a Confirmation term whereas in Field "CX" an Integer – Year).

Table I presents the most commonly used data patterns along with the total number of occurrences in the 8 cohorts. Each pattern (aka template) consists of one or more simple patterns (separated by addition operators) each of which specifies the corresponding RM class for both field and its value (separated by a semicolon). As

can be noticed, more than one pattern is often used for the expression of medical conditions diagnosed, pharmaceutical drugs prescribed, and laboratory tests performed (including blood/urine tests). For each one of them there is often a separate field with a confirmation term. Regarding the Date that an event took place (e.g., date of diagnosis, drug administration start date, etc.), it can be an integer (year) or, in general, a sequence of characters with the year and optionally the month and day. It should be noted that there is more often information about the date of a particular patient visit rather than the date that an event took place. Demographic characteristics such as Sex and Ethnicity are not presented in this Table since there are only a few fields in each cohort (e.g., one Field about Ethnicity) and hence the corresponding patterns have limited occurrences.

TABLE I. COMMONLY USED DATA PATTERNS ACROSS THE 8 COHORTS

ID	Mnemonic Name	Count
DIS-10	Disorder: Confirmation Term	124
DIS-11	Disorder: Confirmation Term + Date: String	32
DIS-20	Disorder: Current/Past/Never	7
DRG-10	Drug: Confirmation Term	32
DRG-11	Drug: Confirmation Term + Start Date: String	12
DRG-12	Drug: Confirmation Term + Start Date: String + End Date: String	16
DRG-20	Drug: Current/Past/Never	8
LAB-10	Lab Test: Numeric Value	96
LAB-11	Lab Test: Numeric Value + Test Date: String	14
LAB-20	Lab Test: Confirmation Term	79
LAB-21	Lab Test: Confirmation Term + Test Date: String	13
DT-01	Date: String	22

Accordingly, the detected patterns were manually examined and the relation among them (if not automatically detected) was specified. Moreover, the process that should be followed for the formal expression of data residing in the corresponding fields using the RM terms was determined. For this purpose, the mandatory and optional parameters of the respective RM data types were taken into consideration, as well as their linking with other RM entities. For instance, about a Diagnosis, the medical condition should be definitely known, and, optionally, the date of diagnosis (among others), while it can be linked with a Person using positive or negative assertions [25] for respectively indicating that the person has been diagnosed with that medical condition or not. Consequently, in case of pattern "DIS-11" an entity with the given disorder and the date of diagnosis (if not empty) should be created, and this entity should be then linked with the person using a positive or negative assertion based on the value of the first field (i.e., confirmation term). Meanwhile, in case of "DIS-20" the medical condition diagnosed and the date (or period of time) that the diagnosis was made could also be recorded. However, for the correct interpretation of the values of this field (i.e., current, past and never) the date that the person visited the healthcare provider should be additionally known, which was captured by another field that follows the data pattern "DT-01".

IV. RESULTS

A. Mapping Scenarios

The previous analysis indicated that the process to be followed for the precise and accurate mapping of the existing cohort parameters with the corresponding RM terms is highly affected by the class and format of data residing in a particular cohort as well as the corresponding RM entities. Nevertheless, since the data often follow a limited number of patterns, this process can be automated to a great extent with the role of the end user being limited to the parameterization and configuration of well-defined patterns and processes. For this purpose, a considerable number of mapping scenarios was developed, which the end user can use for mapping the appropriate cohort fields with the RM terms. For each mapping scenario the following three components were specified: a) the number of cohort fields needed and especially the simple data pattern that each of them should follow, b) the corresponding RM Data Type (including its properties) and c) the process that should be followed for the expression of the cohort data using RM terms, as well as additional information necessary for the correct interpretation of the cohort data residing in the given fields.

Overall 67 different mapping scenarios were implemented that enable users to specify every possible correspondence for the fields of the 8 cohorts and the RM terms. The Mapping Scenarios were organized under broader categories based on the meaning of data captured by the respective fields (Table II). As can be noticed, a considerable number of mapping scenarios was specified about Laboratory Examinations since the outcome of a Lab Test may be a number, a Boolean value or even another RM term (e.g., ANA pattern detected). Also, the numeric outcome of a laboratory examination may not be available but only the assessment, i.e., whether its value is normal or not (high/low) or even whether it lies within a predefined range of values (e.g., above a cut-off value) other than the normal range. Additionally, the date recorded may not be available, but rather the period of time that it belongs to (e.g., before the date of the first visit). A few mapping scenarios were also specified about Questionnaires and Biopsies. Regarding the medical conditions diagnosed and drugs prescribed, the cohort data may also contain a list of diseases or drugs (separated by comma) in different fields. Demographic characteristics such as Sex and Ethnicity often follow a common data pattern and hence a limited number of relevant patterns were detected. The same also stands for Smoking Status and Pregnancies.

TABLE II. MAPPING SCENARIOS CREATED FOR EACH ONE OUT OF SEVEN RM CLASSES OF DATA

Reference Model (RM) Class	Mapping Scenarios Count
Demographics	5
Smoking Status and Pregnancies	5
Medical Conditions	14
Interventions (e.g., Medications)	11
Lab Tests	21

Reference Model (RM) Class	Mapping Scenarios Count
Biopsies	7
Questionnaires & Other Data	4

For enabling software agents to further process the elements specified in each mapping scenario and use them for introducing one or more Mapping Rules, the scenarios were formally expressed in JSON format. Fig. 3 presents the formal expression of a mapping scenario for bridging the gap among two separate cohort fields, i.e., the numeric outcome of laboratory examination and the date that it took place, with the corresponding RM terms. In this example, it should be noted that the data patterns have been specified for each one of the two cohort parameters, whereas the specific RM class and its properties have been recorded. Regarding the data transformation service that should be used for moving from one data representation to the other one (in our work only from source to target), a unique ID of the appropriate service has been provided (in our work, the JAVA class name) along with the parameters that the end user should (mandatory) or could (optional) provide during the mapping process. The data transformation service would be responsible for specifying the values of the properties existing in the right side based on a) the data in the given fields, b) additional parameter(s) provided as well as c) other mapping rules specified (in this case, the corresponding RM Lab Test term). Since the unit of measurement of lab tests outcome has already been specified in the RM, in case the data is expressed in a different unit of measurement, the appropriate unit conversion formula should be used. A detailed description of the functionality that this service should provide along with the additional parameters being necessary is given in the formal expression of each mapping scenario. The actual implementation of this service in a procedural language can take place at a later stage.

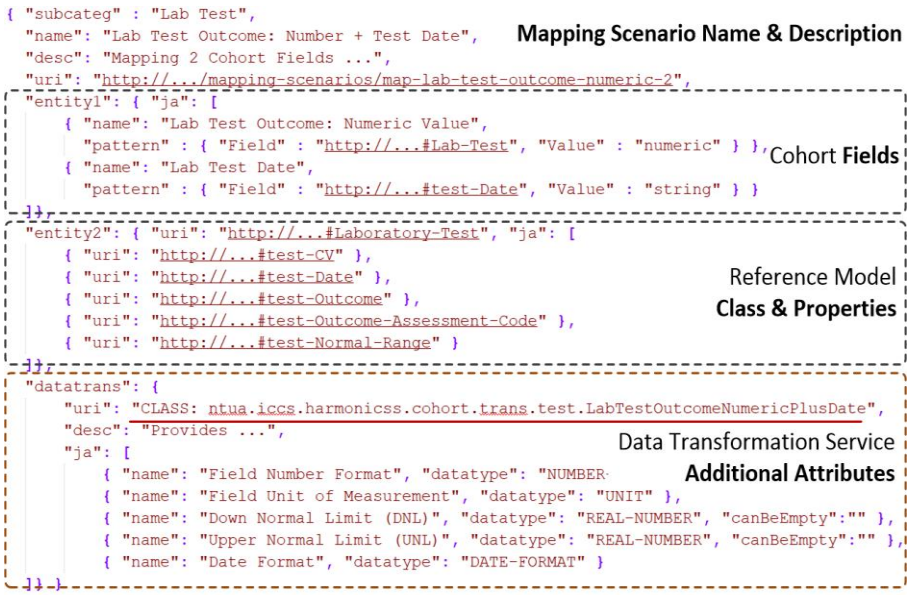

Figure 3. Formal expression of a mapping scenario.

B. Mapping Tool and Mapping Rules

For facilitating the use of the Mapping Scenarios for specifying the correspondence among the Cohort Fields and the ones specified in the RM, the functionality provided by the OAT was extended so that it can be used for instantiating the developed Mapping Scenarios. For this purpose, the appropriate services were implemented, which provide the list of mapping scenarios available (i.e., a mnemonic name, a brief description and the category that they belong to) as well as the internal elements of each specific scenario (i.e., source data patterns, RM elements, service and attributes). Also, the User Interface (especially the 3rd tab about manually specifying Mapping Rules) was updated so as to enable users to instantiate the appropriate mapping scenario and accordingly specify a Mapping Rule (Fig. 4).

Initially, all the available mapping scenarios are presented in OAT (organized in broader categories). Then, depending on the mapping scenario selected by the user, the GUI is populated with the appropriate HTML elements based on the JSON data specified for the particular mapping scenario (step 1). More precisely, the appropriate input fields are introduced for capturing the specific cohort parameters as well as the additional attributes that are essential for the correct interpretation of the data residing in the given fields. Also, the corresponding RM terms as well as the data transformation service used are presented. Accordingly, the user can select the appropriate cohort fields, among the ones that comply with the given patterns, and provide additional attributes, if being necessary (step 2).

In the example presented in Fig. 4 the user has instantiated the aforementioned mapping scenario (Fig. 3) for specifying the correspondence of the two "CHRT_01" Fields, i.e., the Field "FA" with the number of White Blood Cells (WBC) counted based on the blood sample drawn during the first visit (baseline) and the Field "U" with the date that the patient visited that institute, with the appropriate RM terms. For the meaningful description of

the data residing in the given fields, the user has also specified the unit of measurement of the lab test outcome (i.e., "#/mm^3") as well as the normal range of values (i.e., from 4000 up to 11000 #/mm^3), which applies for all the patients independently of their demographic characteristics. Regarding the date, it's a four digit number that indicates the year that the visit (and hence the test) took place. All the data provided by the user were internally stored by the system in the form of a Mapping Rule so that they could be finally exported in the appropriate format, including JSON, EDOAL XML and HTML.

The Mapping Tool was used for bridging the gap among the fields of each of the 8 cohorts and the RM. Apart from the Mapping Rules already specified regarding the meaning of the cohort fields and their values (terminology alignment), additional mapping rules regarding the process that should be followed for moving from one data representation to the other one were introduced, by instantiating the appropriate mapping scenarios. Overall, more than 600 mapping rules were introduced based on the 67 mapping scenarios implemented. The mapping rules specified were carefully examined by the data providers (e.g., corresponding RM terms, normal range of values) and finally used for the harmonization of their cohort data.

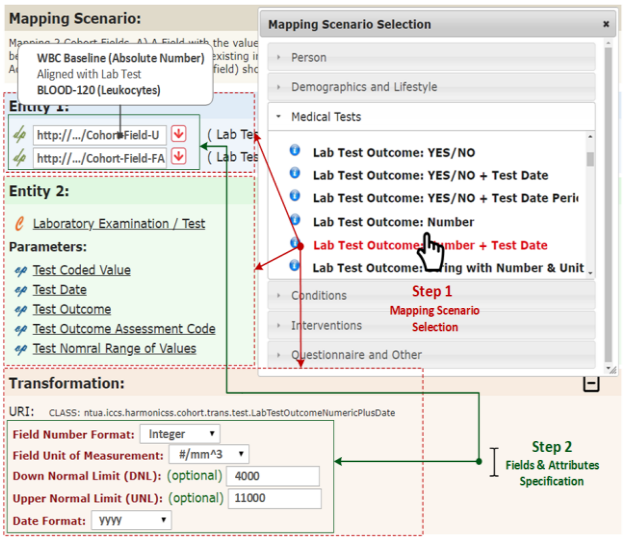

Figure 4. Graphical user interface for introducing a mapping rule using mapping scenarios.

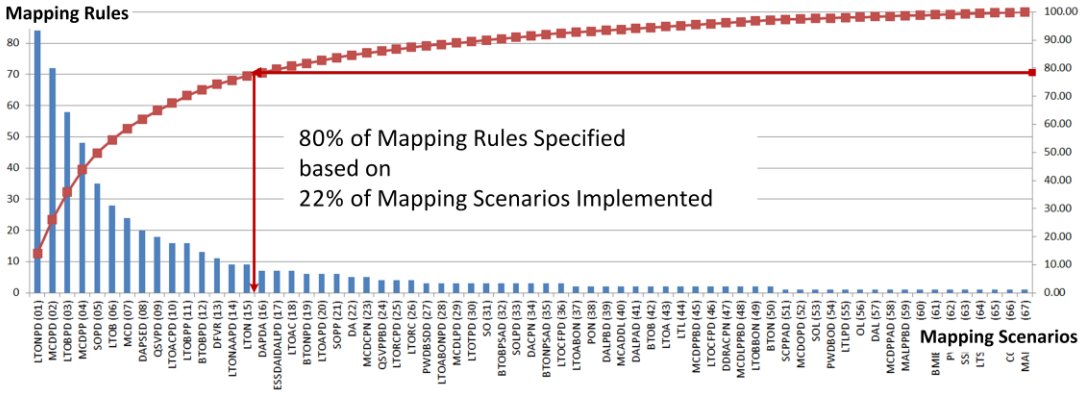

Figure 5. Mapping rules specified based on the mapping scenarios implemented.

V. Discussion and Next Steps

A. Correspondence Specification and Consumption

The approach followed enables the technical experts to accurately map the cohort parameters and their values with the corresponding RM terms so that they can be accordingly used for the harmonization of the patient data. For this purpose, several mapping scenarios were specified despite the fact that some of them are being used a limited number of times (Fig. 5). In particular, it was noticed that almost 80% of the mapping rules specified was based on less than 22% of the mapping scenarios implemented (quite close to Pareto law [26]), which clearly indicates that the existing mapping scenarios can be used for aligning the vast majority of fields of any newly introduced cohort about patients with Sjögren's Syndrome. Nevertheless, the list of mapping scenarios can be easily extended with new ones, which can be directly adapted by the Mapping Tool without any other change being necessary.

The data providers have a distinctive role in the mapping process. They should clarify the meaning of the used cohort terms as well as provide additional information that it is often necessary for the correct interpretation of data residing in their cohort fields (e.g., normal range of values). This is a time consuming process, since there was often the need for communication with the data providers several times in order to identify how the cohort fields and their values could be linked correctly with the RM terms. Since the majority of possible mapping scenarios are now known, the interaction with the data providers could be accelerated through the formulation of pertinent questions, not only taking into account the meaning of terms but also the attributes that would be necessary for the instantiation of the corresponding mapping scenarios.

For facilitating the technical experts in the mapping process, the Mapping Tool should not only provide the corresponding RM terms (i.e., Terminology Alignment) but also the process (i.e., the Mapping Scenario) that should be followed for moving from one side to the other one. More precisely, the system should suggest possible

Mapping Scenarios that can be instantiated based on the meaning of cohort terms and especially the pattern being used, so that the end user can accordingly refine and update them. This is necessary because, in many cases, a few parameters (e.g., units of measurement, normal range of values) should be additionally known which cannot be automatically detected by the system, since they are either presented in free text (e.g., in the description of each field) or they are not provided at all at that time (before contacting the data providers).

Regarding the Mapping Rules specified, they should be further processed by another software system (mentioned in Section III.A) for the expression of the Cohort Data based on the RM terms. This process is quite straightforward, since the source and target elements along with the process that should be followed have already been defined. However, the fact that the mapping rules were specified based on an ontological representation of the respective data model (i.e., Cohort Metadata) should be considered. Also, this module should deal with the mapping scenarios used by each mapping rule, the instantiation and configuration of the data transformation services and the relations among entities produced by system. The background mechanisms used by the Data Harmonization module will be presented in our future work.

B. Application to a Different Domain

The approach followed is highly affected by the structured data adapted for the storage of patient-specific information as well as the formalism developed for the expression of the harmonized data. In our work, the cohort data were stored in spreadsheet documents and the patient data were scattered across several columns, with a particular name/label for each one of them. Regarding the formal expression of harmonized patient data, a RM (OWL ontology) was developed, where each patient was linked with several entities that belong to one out of approximately 25 different data types, which were also organized under broader categories based on the domain that they cover. Consequently, the mapping scenarios developed for bridging the gap among cohort parameters and RM data types was based on the data patterns detected (also reflected in the ontological representation of cohort metadata) and the RM data types specified. Since the design of the RM was driven by the one developed in the past for the representation of Eligibility Criteria [17], which is independent of study drug and disorder, the mapping scenarios implemented can also be used for the harmonization of data from a different clinical domain, with minor changes, provided that the cohort data follow the same format.

In case the cohort data are expressed in a different format (e.g., a relational or graph database), a similar approach can be followed for its analysis (e.g., pattern detection) and the development of the appropriate mapping scenarios. For instance, the tools and services already developed [27] can be used for automatically creating an ontological representation of a relational database schema and the controlled sets of terms used, and accordingly search for commonly used data patterns and identify how they can be linked with the RM terms. The correspondence among source and target models is expected to be more complicated since the data scattered across several cohort data types should probably be joined or links should be followed, since the value of a parameter may point to another entity. Nevertheless, the mapping among source and target ontologies is still feasible through the instantiation of the appropriate Ontology Pattern and the development of the necessary services.

VI. Conclusion

Harmonization of patient data across different healthcare and clinical research entities sets the ground for the ICT-based processing of a massive pool of patient data for clinical research, healthcare and policy making purposes. The significant structural and semantic mismatches among cohorts make the data harmonization problem a rather challenging issue, especially in the health domain, given the sensitive nature of patient data. In our work, a data-blind approach was followed for the formal expression of patient data using a common formalism based on a rich RM developed. The whole approach was based on the precise and complete mapping of the cohort parameters with the RM using a plethora of mapping scenarios, developed for software-based analysis of cohort metadata. Further examination of the Mapping Rules indicated that the existing mapping scenarios can adequately cover the vast majority of heterogeneity issues that should be dealt with for the harmonization of data of any newly introduced cohort. Additionally, the approach followed and the tools developed can be easily extended in order to cover new mismatches, while they can be also adapted to cover different domains.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Efthymios Chondrogiannis conceived the overall approach followed and implemented the tools presented. The first version of this document was prepared by Efthymios Chondrogiannis, but it was further revised and updated by both Efstathios Karanastasis and Vassiliki Andronikou. The research was conducted under the supervision of Theodora Varvarigou. All authors had approved the final version of this document.

Acknowledgment

This work has been partially funded by the European Commission's activity of the Horizon 2020 project HarmonicSS under contract number 731944. 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] M. Ramos-Casals, P. Brito-Zerón, A. Sisó-Almiral, and X. Bosch, "Primary Sjögren syndrome," *BMJ*, vol. 344, no. e3821, June 2012.

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

[3] T. Souza, R. Kush, and J. P. Evans, "Global clinical data interchange standards are here!" *Drug Discov. Today*, vol. 12, no. 3-4, pp. 174-181, Feb. 2007.

[4] C. N. Mead, "Data interchange standards in healthcare IT - Computable semantic interoperability: Now possible but still difficult, do we really need a better mousetrap?" *J. Healthc. Inf. Manag.*, vol. 20, no. 1, pp. 71-78, 2006.

[5] International statistical classification of diseases and related health problems (ICD). [Online]. Available: https://www.who.int/classifications/icd/en/

[6] Anatomical Therapeutic Chemical (ATC). [Online]. Available: https://www.whocc.no/atc_ddd_index/

[7] Logical Observation Identifiers Names and Codes (LOINC). [Online]. Available: https://loinc.org/

[8] P. R. S. Visser, D. M. Jones, T. J. M. Bench-Capon, and M. J. R. Shave, "An analysis of ontology mismatches; heterogeneity versus interoperability," in *Proc. AAAI 1997 Spring Symposium on Ontological Engineering*, 1997, pp. 164-172.

[9] E. Chondrogiannis, V. Andronikou, E. Karanastasis, and T. Varvarigou, "Semantically-Enabled context-aware abbreviations expansion in the clinical domain," in *Proc. 9th ICBBT*, 2017, pp. 89-96.

[10] 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, pp. 238-258, Sept. 2010.

[11] Medical Subject Headings (MeSH). [Online]. Available: https://www.nlm.nih.gov/mesh/meshhome.html

[12] L. Otero-Cerdeira, F. J. Rodríguez-Martínez, and A. Gómez-Rodríguez, "Ontology matching: A literature review," *Expert Systems with Applications*, vol. 42, no. 2, pp. 949-971, Feb. 2015.

[13] 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, pp. 158-176, Jan. 2013.

[14] N. F. Noy, "Ontology mapping," in *Handbook on Ontologies*, Springer, Berlin, Heidelberg, 2009, pp. 573-590.

[15] F. Scharffe and D. Fensel, "Correspondence patterns for ontology alignment," in *Proc. 16th EKAW*, 2008, pp. 83-92.

[16] Expressive and Declarative Ontology Alignment Language (EDOAL). [Online]. Available: http://alignapi.gforge.inria.fr/edoal.html

[17] E. Chondrogiannis, V. Andronikou, A. Tagaris, E. Karanastasis, T. Varvarigou, and M. Tsuji, "A novel semantic representation for eligibility criteria in clinical trials," *JBI*, vol. 69, pp. 10-23, May 2017.

[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, pp. 309-322, Nov. 2008.

[19] E. Chondrogiannis, V. Andronikou, E. Karanastasis, and T. Varvarigou, "A novel approach for clinical data harmonization," in *Proc. BigComp*, 2019, pp. 1-8.

[20] JavaScript Object Notation (JSON). [Online]. Available: https://www.json.org/

[21] E. Chondrogiannis, V. Andronikou, E. Karanastasis, and T. A. Varvarigou, "An intelligent ontology alignment tool dealing with complicated mismatches," in *Proc. 7th International Workshop SWAT4LS*, 2014.

[22] L. Yujian and L. Bo, "A normalized Levenshtein distance metric," *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1091-1095, June 2007.

[23] P. Willett, "The porter stemming algorithm: Then and now," *Program: Electronic Library and Information Systems*, vol. 40, no. 3, pp. 219-223, July 2006.

[24] O. Šváb-Zamazal, V. Svátek, F. Scharffe, and J. David, "Detection and transformation of ontology patterns," in *Proc. IC3K*, 2009, pp. 210-223.

[25] W3C Web Ontology Language (OWL) 2 assertions. [Online]. Available: https://www.w3.org/TR/owl2-syntax/#Assertions

[26] R. C. Craft and C. Leake, "The Pareto principle in organizational decision making," *Management Decision*, vol. 40, no. 8, pp. 729-733, Oct. 2002.

[27] DB to OWL tools. [Online]. Available: http://ponte.grid.ece.ntua.gr:8080/DbToOWL/

Dr. Efthymios Chondrogiannis received his diploma from the School of Electrical and Computer Engineering (SECE) of the National Technical University of Athens (NTUA), in 2008. Since then he has been a researcher at the Institute of Communication and Computer Systems (ICCS) of the NTUA. In 2017, he received his PhD from the SECE of the NTUA. In his thesis, he has focused on the development of innovative semantically-enabled mechanisms for improving the design of a clinical study. During the course of his military service he worked as a software engineer in the Directorate of Military and Technological Support of the Greek Army. Also, he has participated in many EU funded projects such as PONTE, OpenScienceLink, ACTIVAGE and HarmonicSS. His research interests include ontology design, semantic web, health informatics, system interoperability, data harmonization and service oriented architectures.

Efstathios Karanastasis received his Diploma on Electrical and Computer Engineering from the University of Patras, Greece, in 2007. In the past has undertaken IT and web projects in the private sector. In 2004 he worked for Athens Olympic Broadcasting, in the production and archiving of the broadcasted program of the XXVIII Summer Olympic Games. During the course of his military service he worked for the Center of Informatics of the Greek Army (KEPYES) as a software developer. In addition, he qualified as a consultant IT-specialist at the e-government team assembled by the Greek Ministry of Administrative Reform and e-Governance in cooperation with the Ministry of Defense. Currently, he is a PhD candidate in the department of Electrical and Computer Engineering of the National Technical University of Athens (NTUA) and has been employed since 2006 as a researcher at the Institute of Communications and Computer Systems (ICCS). He has participated in numerous EU-funded IT projects including BEinGRID, PONTE, OpenScienceLink, HarmonicSS and ACTIVAGE, mainly involved with SOA platforms design and implementation, and heterogeneous data integration. He is fluent in Greek, English and German. His research interests include service oriented architectures, knowledge modeling, data integration, cloud computing, IoT, and web interfaces.

Dr. Vassiliki Andronikou received her diploma from the Electrical and Computer Engineering School of the National Technical University of Athens in 2004. She has worked in the National Bank of Greece and the Organization of Telecommunications of Greece, while since 2004 she has been a research associate and PhD candidate in the Telecommunications Laboratory of the NTUA. In 2005 she was given the Ericsson award for her thesis on "Mobile IPv6 with Fast Handovers". In 2009, she received her PhD in the area of Biometric Systems focusing on innovative techniques for the improvement of their efficacy and effectiveness at fusion and resources level from the school of Electrical and Computer Engineers of NTUA. Her research has involved her participation in many European projects, such as HarmonicSS, ACTIVAGE, OpenScienceLink, PONTE, BEinGRID, POLYMNIA, FIDIS and

AKOGRIMO, with her interests focusing on the knowledge modeling and data harmonization in the biomedical domain.

Prof. Theodora A. Varvarigou received the B. Tech degree from the National Technical University of Athens, Athens, Greece in 1988, the MS degrees in Electrical Engineering (1989) and in Computer Science (1991) from Stanford University, Stanford, California in 1989 and the Ph.D. degree from Stanford University as well in 1991. She worked at AT&T Bell Labs, Holmdel, New Jersey between 1991 and 1995. Between 1995 and 1997 she worked as an Assistant Professor at the Technical University of Crete, Chania, Greece. Since 1997 she was elected as an Assistant Professor while since 2007 she is a Professor at the National Technical University of Athens. Prof. Varvarigou has great experience in the area of embedded systems and Cloud computing. She has published more than 150 papers in leading journals and conferences. She has participated and coordinated several EU funded projects such as PONTE, OpenScienceLink, ALLADIN, COSMOS, SocIoS, CONSENSUS and SCOVIS.