e-Health – For Continuity of Care
C. Lovis et al. (Eds.)

doi:10.3233/978-1-61499-432-9-283

Formalizing Mappings to Optimize Automated Schema Alignment: Application to Rare Diseases

Meriem MAAROUFI[a,b,1], Rémy CHOQUET[a,b], Paul LANDAIS[a,c], Marie-Christine JAULENT[b]
[a] *Banque Nationale de Données Maladies Rares, Hôpital Necker Enfants Malades, Assistance Publique des Hôpitaux de Paris, Paris, France*
[b] *INSERM, U1142, LIMICS, Paris, France;*
Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1142, LIMICS, Paris, France;
Université Paris 13, Sorbonne Paris Cité, UMR_S 1142, LIMICS, Villetaneuse, France
[c] *Montpellier1University, EA2415 & BESPIM, University Hospital, Nîmes, France*

Abstract. In the era of data sharing and systems interoperability, the automation of data schema alignment has become a priority. Discovering data mappings is the aim of many alignment approaches that have been described in the literature and the effectiveness of which depends on data specifications. In this context, we propose a method for mappings formalization that allows automated data integration processes optimization. This method, involving both data element level and value element level, allows an automated inference of mappings expressed by rules. In this paper, we start by describing the methods used to achieve this mappings formalization. Then, we explain how it has been validated by characterizing data from two use cases. We end up by discussing the objectives of the proposed formalization.

Keywords. Data integration, schema alignment, interoperability, mappings formalization, medical informatics

Introduction

Building a French Database for Rare Diseases (BNDMR) that identifies patients suffering from rare diseases at a national level is an ambitious ongoing project [1]. This database should allow institutions to evaluate the adequacy between the healthcare supply and demand, but should also allow researchers to identify eligible patients for clinical trials or rare diseases cohorts. The project deals with 7000 [2] different diseases and the data is collected at a national scale from multiple sources. Data sources are 131 Rare Disease Reference Centers and 501 Competence Centers that cover 30 rare disease groups.

Connecting all these sources to the target database requires huge efforts in order to deal with data heterogeneity. All the sources do not collect the same data, and even if they share similar data elements [3], syntactic and semantic heterogeneity may persist.

[1] meriem.maaroufi@bndmr.fr

As a first effort to deal with the interoperability issue, BaMaRa, a communicating application, was designed to collect a Rare Disease Minimum Data Set (MDS) [4]. Standardized elements were used to build the MDS and mappings[2] have to be set to connect the databases that don't use those standards.

Automated schema alignment[3] approaches have been released to avoid spending time on manual alignment to detect similarity between data elements building different data schemas (e.g. database schemas, xml schemas, ontologies…). The way to classify these alignment approaches may differ in the literature [5,6]. However, we can notice the existence of four major classes: linguistic approaches, structure level approaches, approaches based on constraints and instance level approaches.

As analyzed in some evaluation studies [5,7,8] on alignment techniques, the effectiveness of the different approaches depends on the inherent characteristics of data, schemas and coding. In this paper, we propose a methodology for characterizing mappings that will allow a data pre-analysis, and optimize the appliance of each automatic alignment approach. We thus investigated the nature of the experimental mappings using a heuristic approach to derive the characterization.

1. Methods

1.1. First attempt of a mappings' classification

Our first experimentation was to integrate data coming from the National Alzheimer Database [9] (BNA). After a manual alignment, we obtained less than 50% of recovery rate. A study of the results allowed the identification of five different mappings linking the source (BNA) and target (BaMaRa) data elements:

- *Exact match:* the source element was mapped to the target element and their domains values matched perfectly.

  E.g. the source element "birth name" was mapped to the target element "patronymic name", there was no coding transformation.
- *Partial match:* the source element was mapped to the target element but the domains values matched only partially.

  E.g. the source element "patient sent by" was mapped to the target element "patient addressed by" but their coding lists just overlapped.
- *Conditional match:* the source element could be mapped to the target element only if a condition was verified.

  E.g. the source element "name of use" could be mapped to the target element "marital name" only if it was different from "birth name".
- *Aggregation:* Two or more source elements were mapped to the target element.

  E.g. the source elements "department code" and "commune code" were aggregated to give the target element "birth country code".
- *Split-up:* the source element was mapped to two or more target elements.

[2] Mapping = the relationship indicating a similarity according to a given measure between two elements of two data schemas.

[3] Alignment = (also called matching) the process of detecting mappings between elements of different data schemas.

E.g. the source element “type of act” was mapped to three target elements “activity context”, “activity objective” and “profession of the personnel performing the activity”.

All the mappings obtained after the BNA and BaMaRa schema alignment fitted into the previous classification. This classification was similar to the one described in [10] and provided an overview on the relations linking the source data elements to the target data elements and the involved cardinalities.

The classes defined above were not always disjoint: a mapping between source and target elements could be a split-up matching with a condition.

Our main contribution is to propose a new formalization of mappings processable by the machine and operating at a value elements level. Being given a source schema description, the objective is to fulfill the maximum of the target schema elements. It is a one-way process: to integrate the source data in the target data model. However, with this approach, the reverse path can be used to get a bidirectional alignment.

1.2. Mappings formalization

In the proposed method for characterizing mappings, formalization considers the data elements, the value elements and the exact relation that links the source and the target elements which is described below:

Let $S=\{E^S_i; i=1..n; n=card(S)\}$ be the *source* dataset and $T=\{E^T_j; j=1..m; m=card(T)\}$ the *target* dataset, with E^S_i and E^T_j their constitutive *data elements*. E^S_i and E^T_j domain values can be either finite (e.g. a predefined list of values) or infinite (e.g. a textual or an integer entry). We note e^S_{ik} and e^T_{jl} the respective *value elements* of E^S_i *and* E^T_j. A value element can represent one item from the finite value domain (e.g. for e^S_{ik}, k=1..p with p=card(E^S_i)) or the different possible values of the infinite domain (e.g. for e^S_{ik}, we set k=0 and e^S_{i0} is treated regardless of the value it takes).

E^S_i can be mapped to E^T_j by one or more binary relations e^S_{ik}-e^T_{jl}. Each binary relation e^S_{ik}-e^T_{jl} is defined by one or more rules r. A mapping is then defined for each value element pair and not for each data element pair.

To summarize, a mapping from *S* to *T* can be characterized by the triplet $\{E^S_i$-E^T_j; e^S_{ik}-e^T_{jl}; $r\}$:

- A binary relation E^S_i-E^T_j between a source data element and a target data element.
- A binary relation e^S_{ik}-e^T_{jl} between a source value element of E^S_i and a target value element of E^T_j.
- A rule *r* expressed in the “ if … then …” format.

2. Results

The mappings formalization described above is a result that we validated by characterizing a set of mappings from two use cases. As a first use case, we chose to characterize data from CEMARA [11], a database that collects data of about 240,000 patients suffering from rare diseases. CEMARA and BaMaRa schemas are not too different and share some of their data elements. The triplet data elements pair, value elements pair and rule was relevant for each mapping obtained after schema alignment.

Our second use case was to align BNA schema (1.1) and BaMaRa schema and we could also characterize all the mappings by the proposed formalization (see Table 1).

Table 1. Examples of triplet mappings from the two use cases

E^S_i - E^T_j	e^S_{ik} - e^T_{jl}	r
"Death" – "Vital status"	"Y" - "Y"	If $\mathbf{e}^S_{ik}$ ="Y" then $\mathbf{e}^T_{jl}$ ="Y"
"Death" – "Vital status"	"N" - "N"	If $\mathbf{e}^S_{ik}$ ="N" then $\mathbf{e}^T_{jl}$ ="N"
"Coming CPC" – "Patient addressed by"	"Y" - "CPC"	If $\mathbf{e}^S_{ik}$ ="Y" then $\mathbf{e}^T_{jl}$ ="CPC"
"Name of use" – "Marital name"	String - String	If $\mathbf{e}^S_{i0} \neq \mathbf{e}^S_{c0}$ then $\mathbf{e}^T_{j0} = \mathbf{e}^S_{i0}$ (E^S_c= "Birth name")

The mappings formalization allowed the definition of automatic alignment processes where the two first pairs of the triplet (data elements and data values) were inputs and the third element, which was the rule describing the exact relation between them, was the output. The methodology was the following:

- Data pre-analyses: Getting the source and target schema description and identifying data elements (E^S_i, E^T_j) and corresponding value elements (e^S_{ik}, e^T_{jl}). Homogeneous data groups' creation according to data types and domain values.
- Processes definition: Defining the strategies (processes or algorithms involving one or more schema alignment approaches) to come along for each group of data, and route the data to the suitable treatment.

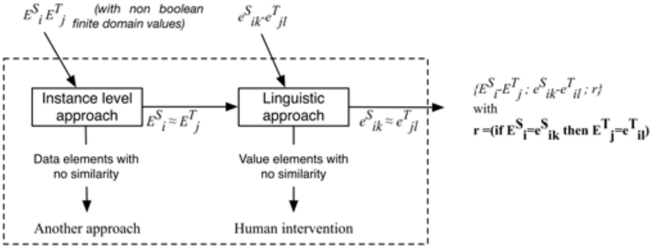
E^S_i E^T_j (with non boolean finite domain values)
e^S_{ik}-e^T_{jl}
Instance level approach
$E^S_i \approx E^T_j$
Linguistic approach
$e^S_{ik} \approx e^T_{jl}$
{E^S_i-E^T_j ; e^S_{ik}-e^T_{il} ; r}
with
r =(if E^S_i=e^S_{ik} then E^T_j=e^T_{il})
Data elements with no similarity
Value elements with no similarity
Another approach
Human intervention

Figure 1. Example of an optimized alignment process.

An example (see figure 1) of an optimized process of schema alignment might be the following:

- As input process, consider subsets of {E^S_i} and {E^T_j} data elements that have a finite and non-Boolean domain of values.
- Apply an instance level approach on the two sets of data elements {E^S_i} and {E^T_j} in order to detect similarity based on instance redundancy.
- For the similar pairs E^S_i-E^T_j, a linguistic approach will be applied to the corresponding value sets {e^S_{ik}} and {e^T_{jl}} to detect the e^S_{ik}-e^T_{jl} pairs.
- The output mappings can be characterized by the triplet {E^S_i-E^T_j ; e^S_{ik}-e^T_{jl} ; r} with: r = (if E^S_i=e^S_{ik} then E^T_j=e^T_{jl}).

3. Discussion

To summarize, the methodology we describe in this paper is based on:

- Data pre-analysis and the importance of the dualities: source data/target data and data element level/value element level. This will allow the construction of homogeneous data groups.

- Possibility of proposing reliable automated data integration processes for each data group that will infer the third element of the triplet presented in section 1.2: the rule that specify the mapping.

Adopting this new methodology will not improve the effectiveness of the automated approaches that are applied on data. We do not introduce a new approach or a new algorithm that will infer mappings that were previously undetectable without human intervention because of semantic issues. Indeed, this proposition aims to optimize the use of alignment approaches and to limit human intervention.

In a real data integration experience, aligning schemas using automated alignment tools remains a human supervised task, not only to validate the inferred mappings but also to make a decision toward the "good" result. In fact, different approaches, that proved their effectiveness in previous schemas alignments, used to be applied on all data then the obtained results are compared and weighted according to data specificities. Following the methodology that we propose in this paper and standardizing the minimal data description would allow the reusability of effective processes, for each suitable data group and would introduce some confidence in the previous works and reliance on the proposed results.

Acknowledgments

We acknowledge A. de Carrara and C. Messiaen for their comments. This work was funded by the French Ministry of Health.

References

[1] Banque Nationale de Données Maladies Rares [Internet]. 2014 [cited 2014 Jan 14]. Available from: http://www.bndmr.fr/

[2] Orphanet [Internet]. 2012 [cited 2014 Jan 14]. Available from: http://www.orpha.net/

[3] ISO, IEC. ISO/IEC 11179-3. 2013.

[4] Choquet R, Messiaen C, Priouzeau A, de Carrara A, Landais P. Un jeu de données minimum pour faciliter l'interopérabilité des bases de données pour les maladies rares. Paris; 2012. p. 1–6.

[5] Rahm E, Bernstein PA. A survey of approaches to automatic schema matching. VLDB J. 2001 Dec 1;10(4):334–50.

[6] Euzenat J, Shvaiko P. Classifications of Ontology Matching Techniques. Ontology Matching. Springer Berlin Heidelberg; 2013. p. 73–84.

[7] Euzenat J, Shvaiko P. Evaluation of Matching Systems. Ontology Matching. Springer Berlin Heidelberg; 2013. p. 285–317.

[8] Kaza S, Chen H. Evaluating Ontology Mapping Techniques: An Experiment in Public Safety Information Sharing. Decis Support Syst. 2008 Nov;45(4):714–28.

[9] Le Duff F, Duport N, Gonfrier S, Lafay P, Texier N, Schück S, et al. Plan national Alzheimer 2008-2012 - Mesure 34 Mise en place du recueil épidémiologique national et premières tendances. Rev Gériatrie. 2010;35(8):575–82.

[10] Savasere A, Sheth A, Gala S, Navathe S, Marcus H. On applying classification to schema integration. , First International Workshop on Interoperability in Multidatabase Systems, 1991 IMS '91 Proceedings. 1991. p. 258–61.

[11] Messiaen C, Le Mignot L, Rath A, Richard J-B, Dufour E, Ben Said M, et al. CEMARA: a Web dynamic application within a N-tier architecture for rare diseases. Stud Health Technol Inform. 2008;136:51–6.