Towards data integration automation for the French rare disease registry

Meriem Maaroufi, MSc1,2, Remy Choquet, PhD1,2, Paul Landais, MD, PhD1,3, Marie-Christine Jaulent, PhD2

1 Banque Nationale de Données Maladies Rares, Hôpital Necker Enfants Malades, Assistance Publique des Hôpitaux de Paris, Paris, France; 2 INSERM, U1142, LIMICS, Paris, France; 3 Montpellier University, EA2415 & BESPIM, University Hospital, Nîmes, France

Abstract

Building a medical registry upon an existing infrastructure and rooted practices is not an easy task. It is the case for the BNDMR project, the French rare disease registry, that aims to collect administrative and medical data of rare disease patients seen in different hospitals. To avoid duplicating data entry for health professionals, the project plans to deploy connectors with the existing systems to automatically retrieve data. Given the data heterogeneity and the large number of source systems, the automation of connectors creation is required. In this context, we propose a methodology that optimizes the use of existing alignment approaches in the data integration processes. The generated mappings are formalized in exploitable mapping expressions. Following this methodology, a process has been experimented on specific data types of a source system: Boolean and predefined lists. As a result, effectiveness of the used alignment approach has been enhanced and more good mappings have been detected. Nonetheless, further improvements could be done to deal with the semantic issue and process other data types.

Introduction

Data analysis is important for health care systems optimization. Studying medical administrative data helps institutions to take right decisions toward health policy so as to lower costs and improve healthcare quality and access. In this context, the French health ministry is funding the BNDMR project (Banque Nationale de Données Maladies Rares) in order to build a medical data warehouse for rare diseases.

Currently, the French rare diseases (RD) care network is based on 131 rare disease centers of excellence. Each center is composed of one or more healthcare RD units, called sites, located in hospital structures that are known for their expertise and specific skills in managing and treating particular rare diseases.

In their daily activity, professionals from those sites use different systems to enter patients’ data such as administrative data, biological/imagery results and reports in the hospital information system; disease specific data in national or international registries; genetic or biological data in specific applications; customized data in local data bases, etc.

In order to avoid duplicating data entry for health professionals, the BNDMR will propose the deployment of connectors to retrieve data from the different source systems and allow interoperability at a national level.

The first step to ensure interoperability was to define a minimum data set built on common information to all rare diseases. This minimum data set was built and standardized following a methodology1. The next step to collect data from all the different systems used by health professionals is to align the data schemas of source systems with the target data schema, the BNDMR minimum data set, in order to detect the mappings linking data elements. To manually align the data schemas is a tedious task. Even though we are familiar with the target data schema and its sixty data elements, there is a huge number of potential source systems, about 500 systems, with heterogeneous datasets that are more or less voluminous. The automation of this task becomes necessary.

The literature describes many automated alignment tools based on different approaches2–4. Those tools have been used in different contexts to discover mappings between heterogeneous data sources e.g. BiobankConnect5. Obtained results are promising and reliable but they are not sufficient to proceed directly to data exchange. An important consideration is missing: the values of data elements. If the value element level is not taken into account, first we can miss some mappings that exist at that level, and second, we cannot define a concrete mapping between data elements that can be directly implemented in the data integration process.

In this context, we previously proposed a mapping formalization6 in order to take into account the value element level that can be used to generate some parts of the program in charge of data import or export. In this paper we describe how this formalization contributes in a whole data integration process. Then, we present a concrete application by aligning a sub dataset of a data source schema, CEMARA7, an academic rare disease data base, with the target data schema, the minimum data set of the BNDMR.
Background and proposal

Authors of different comparative studies on alignment or matching approaches agree to say that the efficiency of those approaches depends on data types and the application context. Furthermore, the aim of those tools is to align two sets of concepts (concepts of a thesaurus or an ontology). Using a certain measure, they detect pairs of concepts and assign them a similarity level. The obtained results are mappings in this format: \( \{C_1, C_2, s(C_1, C_2)\} \), e.g. \( \{\text{Cystinosis}, \text{Cystin}, 0.8\} \).

These results are not sufficient to define a complete data integration process. In fact, value elements are only considered when the alignment tools use instance-based approaches, but they are not integrated in the mappings definitions, they are only used to detect mappings between data elements. Data elements from the different schemas are often not coded the same way. Data transformation will be needed in order to complete the harmonization and proceed to data import. Data transformation includes: aligning lists of data values (thesaurus, specific lists…) and other operations like arithmetic operations to harmonize units or simple concatenation.

Rahm and Bernstein quickly mentioned some important issues, they pointed out the importance of “mapping expressions” to specify how the elements of two schemas are related. They also underlined the fact that, in practice, the criteria matching the elements are based on heuristics and are not mathematically precise to be used when implementing the match. On the basis of those considerations, it is necessary to: (i) optimize the use of the different alignment approaches in the data integration process while taking into account data types; (ii) generate well-defined mappings, integrating a value element level that can be directly exploitable.

(i) Integrate schema alignment approaches in a whole process

The alignment algorithms intervene in the second step of a four phased data integration process:

- First phase: data selection

As depicted in most studies, some alignment approaches can be considered as appropriate or inappropriate depending on data types. This phase will allow data selection according to the data types that will be addressed: numeric types, string and text types, Boolean type, lists and enumerations, etc. Data selection can be carried further by considering some constraints on data like string length, bounds and units of numeric types, or source terminologies of some enumerations. It is also important to specify the context and domain of study in order to adapt the external resources that will be used: dictionaries and terminologies.

- Second phase: mappings detection

In this phase, automatic alignment tools will be used to detect mappings in the way that have been designed for the selected data. This can be done by applying directly one alignment approach or by defining a complex path that will be taken by the data. External resources can be used to semantically enrich the data schemas and, for example, be able to map synonyms.

- Third phase: mappings validation

Validation is generally a human validation involving persons familiar with the source and target data schemas.

- Fourth phase: code generation

Once the mappings validated, they are integrated in the program in charge of data transformation and exchange. This implies that mapping expressions have to be readable by programming languages.

(ii) Mapping formalization

As proposed in our previous publication, a mapping from a source schema \( S \) to a target schema \( T \) can be characterized by the triplet \( \{E^S_i ; E^T_j ; r\} \):

- A binary relation \( E^S_i ; E^T_j \) between a source data element and a target data element. A data element is a unit of data identified by a name having a definition and that can be permissive to only some values.
- A binary relation \( e^S_i ; e^T_j \) between a source value element of \( E^S_i \) and a target value element of \( E^T_j \). A value element is a value out of a set of permissible values pertaining to a data element.
- A rule \( r \) expressed in the “if … then …” format defining the exact relation linking the different data and value elements.

This formalization is not a bijection; it links the source elements to the target elements and the reverse path is not always true.

Example: \( \{\text{“Coming CPC”} \text{– “Patient addressed by”}; \text{“Y” – “CPC”}; \text{if “Coming CPC” = “Y” than “Patient addressed by” = “CPC”}\} \)
Experimentation

We experimented our methodology by aligning a sub dataset of our first data source schema, CEMARA, with the target data schema, the minimum data set of the BNDMR.

• Data selection

The process we defined is able to detect simple mappings between source and target elements from Boolean and predefined lists data. Among the 72 data elements from CEMARA schema, 39 are Boolean data and 15 are predefined lists. As of the BNDMR schema, it contains 15 Boolean and 16 predefined lists among its 62 data elements.

Since the Boolean value elements (“true” and “false”) are not enough informative to infer mappings, we settled for considering only the data element level. For instance, we will try to map the data element’s label “the patient is deceased” and we will not do so for the value elements “true” or “false”. However the value elements will be integrated in the formalization of mappings into the rules during the final phase.

On the other side, considering the value element level becomes more important than considering the data element level when we are dealing with predefined lists. Furthermore, value elements from two lists can be mapped even though their respective data elements are not semantically similar.

E.g. the source data element “act type” and the target data element “participant profession” are two different concepts however their respective value elements, “nurse intervention” and “nurse” are closely linked.

The number of elements to map moves from 15 data elements to 86 value elements for the source CEMARA and from 16 data elements to 106 value elements for the target BNDMR.

• Mappings detection

A linguistic approach is used to operate a cross alignment between the different types of elements. This linguistic approach is based first on a semantic enhancement of both source and target data schemas and second on a lexical matching between the elements using the Levenshtein algorithm (also called the Levenshtein distance used to measure the differences between two sequences using the number of characters’ deletions, insertions and substitutions)(8). For the semantic enhancement, synonyms and English translations have been assigned to the elements.

OnAGUI8, a tool for aligning ontologies integrating different algorithms, has been used to apply the Levenshtein algorithm on the source and target schemas that have been translated in OWL format.

The same information can be structured differently from one schema to another. Cross alignment between different typed data is necessary. Four alignments have been done, operating either at the same level (data element or value element level) or at different levels (both data element and value element levels) depending on the type of data that we are aligning, as explained in the data selection phase (Figure 1).

<table>
<thead>
<tr>
<th>Source schema: CEMARA</th>
<th>Target schema: BNDMR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data element level</strong></td>
<td></td>
</tr>
<tr>
<td>type:Boolean</td>
<td>type:Boolean</td>
</tr>
<tr>
<td>Matching 1</td>
<td>Matching 2</td>
</tr>
<tr>
<td>type:List</td>
<td>type:List</td>
</tr>
<tr>
<td>Matching 3</td>
<td>Matching 4</td>
</tr>
<tr>
<td>List element</td>
<td>List element</td>
</tr>
</tbody>
</table>

**Figure 1.** Alignment process: Cross alignment for Boolean data and predefined lists

• Mappings validation

For our experimentation validation, a person familiar with both source and target databases has done a manual alignment between the different elements to establish a gold standard. Two experts of both databases validated this alignment. This liable set of mappings allowed the validation of mappings that have been detected by the defined process.
In routine use, a simple selection of the valid mappings would have been sufficient for this phase.

- Code generation

The table below (Table 1) shows that four simple types of rules will be generated by this process, linking Boolean value elements (true and false) and lists value elements.

Table 1. Generated mappings expressions

<table>
<thead>
<tr>
<th>Source elements</th>
<th>Target elements</th>
<th>Resulting mappings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data type</td>
<td>Element level</td>
<td>Data type</td>
</tr>
<tr>
<td>Matching 1</td>
<td>boolean</td>
<td>boolean</td>
</tr>
<tr>
<td>Matching 2</td>
<td>list</td>
<td>boolean</td>
</tr>
<tr>
<td>Matching 3</td>
<td>boolean</td>
<td>list</td>
</tr>
<tr>
<td>Matching 4</td>
<td>list</td>
<td>VE</td>
</tr>
</tbody>
</table>

* DE: data element, VE: value element

Below some examples of the generated mapping expressions:

- If PropLink=propositus [source] then Propositus=true [target]
- If ConfCyto=true [source] then ConfirmationMode=cytogenetic [target]

Results

The similarity threshold for the Levenshtein distance has been set to 0.9 to contain the number of false positive mappings. Counting is relative to the target elements: How many target data elements and value elements have been mapped? How many have been correctly mapped or not mapped?

In Matching 1 and Matching 2, the algorithm is used to search mappings with source elements for the 16 Boolean target data elements. In Matching 3 and Matching 4, the algorithm is used to search mappings with source elements for the 106 target value elements building the predefined lists of some data elements (Table 2).

Table 2. Alignment results

<table>
<thead>
<tr>
<th></th>
<th>Good mappings</th>
<th>Wrong mappings</th>
<th>Missed mappings</th>
<th>Good non mappings</th>
<th>Total</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TP*</td>
<td>FP*</td>
<td>FN*</td>
<td>TN*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matching 1</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>9</td>
<td>16</td>
<td>0.43</td>
<td>1.00</td>
</tr>
<tr>
<td>Matching 2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>16</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Matching 3</td>
<td>22</td>
<td>4</td>
<td>4</td>
<td>76</td>
<td>106</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>Matching 4</td>
<td>35</td>
<td>8</td>
<td>25</td>
<td>38</td>
<td>106</td>
<td>0.58</td>
<td>0.81</td>
</tr>
</tbody>
</table>

* TP: true positives, FP: false positives, FN: false negatives, TN: true negatives

For a comparative evaluation, we ran the Levenshtein algorithm on the complete source and target data schemas as a reference test. We kept the same threshold of 0.9 of similarity. Fifteen mappings have been detected linking 14 CEMARA data elements to 14 BNDMR data elements. Among the 15 mappings, 5 were false mappings, e.g. Diagnostic status – Patient status. Operating on a data element level can be misleading especially when dealing with predefined lists. The 10 good mappings are distributed as shown below (Table 3).
Table 3. Comparative evaluation: number of good mappings for the reference test and the experimentation

<table>
<thead>
<tr>
<th></th>
<th>Reference Test</th>
<th>Experimentation</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matching 1</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Matching 2</td>
<td>0</td>
<td>1</td>
<td>Not the same total number of elements. In the reference test, only data elements are considered</td>
</tr>
<tr>
<td>Matching 3</td>
<td>1 (2 missed)</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>Matching 4</td>
<td>6 (2 missed)</td>
<td>35</td>
<td></td>
</tr>
</tbody>
</table>

Besides the missed mappings, mappings that have been detected by the reference test are linking data elements and not value elements. Thus, seven of the ten “good” mappings are not precise enough to be considered as good mappings.

Discussion

To summarize, the proposed method is based on a new way to formalize mappings with rules specifying the exact relationship between source and target data elements and their values. The alignment process is designed to suit the data types chosen as inputs in order to generate well-defined type of rules as output. In our experimentation we studied two data types: Boolean and predefined lists, we operate a cross-alignment at both data element and value element levels and detected four types of rules defining the mappings.

The comparative evaluation shows that using the same data schemas (CEMARA and BNDMR) and the same alignment tool (OnAGUI) as the reference test, more mappings could have been detected when using the proposed methodology. Further, value elements form integral part of the mapping expressions.

Moreover, results show that globally we obtained good recall and precision. Wrong mappings (false positives) that move away the precision from 1 can be explained by the generic nature of some value elements, e.g. “other”. Their occurrence in different list data elements explains their wrong mapping to similar elements several times. Such cases can be processed differently in the alignment process by linking them to their data element or siblings value elements. However, precision remains acceptable, greater than 0.8. The recall is more related to semantic issues. Some mappings have not been detected (false negatives) this is mainly due to the chosen alignment approach that is based on a simple string similarity metric. Moreover, and despite the efforts made to build and enrich the schemas, the algorithm does not use at the moment an external semantic resource to search more synonyms or vocabulary variants, like the difference between “medically assisted procreation” and “medical assistance to procreation” and thus improve on the recall.

If the data integration process were totally automated with no human intervention to validate or reject the proposed mappings, the registry data quality will be negatively affected. A bad recall would mean a non-efficient alignment with missing mappings and therefore missing data. A bad precision would mean wrong mappings and therefore importing false information in the target data base. The priority in our work is to maintain the precision as high as possible while trying to improve the recall making the task easier for the user which validation remains essential to not affect the data base quality and maintain the reliability on the ensuing studies. Thus, among the next evolutions of our work, an automatic semantic enrichment will be integrated to the process. The first step would be to request the MeSH terminology to retrieve data elements and value elements synonyms to catch some undetected mappings. It would also be interesting to consider other data types and define the suitable alignment process to detect their mappings. E.g. in order to integrate some measures like the patient weight and height, numeric data alignment may be addressed with taking into account the unit conversion. And finally, a global validation will be conducted using other source systems.

Conclusion

The proposed methodology aims to optimize the use of existing alignment approaches and to limit human intervention in the data integration process. Our work takes part of a global effort for the healthcare systems interoperability. Data schemas alignment can be an alternative to address the lack of standardization in the existing systems or as a complement if different standards are implemented.

Acknowledgments

We acknowledge the BNDMR project team for their comments. This work was funded by the French Ministry of Health.
References


