MEDINFO 2015: eHealth-enabled Health
I.N. Sarkar et al. (Eds.)

doi:10.3233/978-1-61499-564-7-217

Building a Semantic Interoperability Framework for Care and Research in Fibromuscular Dysplasia

Marie-Christine Jaulent[a], Ariane Assélé-Kama[a], Sébastien Savard[b,d], Alessandra Giavarini[b], Emmanuel Touzé[c], Xavier Jeunemaître[b], Adrien Ugon[a], Pierre-François Plouin[b], Laurent Toubiana[a]

[a] INSERM, U1142, LIMICS, F-75006, Paris, France; Sorbonne Universités, UPMC Université Paris 06, UMR_S 1142, LIMICS, F-75006, Paris, France; Université Paris 13, Sorbonne Paris Cité, UMR_S 1142, LIMICS, F-93430, Villetaneuse, France
[b] Hypertension unit, Department of genetics and Centre for rare vascular diseases; Hopital Europeen G Pompidou (HEGP);Paris-Descartes University ; INSERM U970, Paris Cardiovascular Research Centre, 75015 Paris, France
[c] Université Caen Basse Normandie, CHU Côte-de-Nacre, service de neurologie et unité neurovasculaire, 14000 Caen, France; INSERM U919, GIP CYCERON, 14074 Caen, France.
[d] Hôtel-Dieu de Québec, Service de néphrologie, Quebec, Canada

Abstract

Identifying patients with Fibromuscular Dysplasia (FMD) at the international level will have considerable value for understanding the epidemiology, clinical manifestations and susceptible genes in this arterial disease, but also for identifying eligible patients in clinical trials or cohorts. We present a two-step methodology to create a general semantic interoperability framework allowing access and comparison of distributed data over various nations, languages, formats and databases.
Methods: The first step is to develop a pivot multidimensional model based on a core dataset to harmonize existing heterogeneous data sources. The second step is to align the model to additional data, semantically related to FMD and collected currently in various registries. We present the results of the first step that has been fully completed with the validation and implementation of the model in a dedicated information system (SIR-FMD). We discuss the current achievements for step 2 and the extensibility of the methodology in the context of other rare diseases.

Keywords:

Information Systems; Health Information Exchange; Semantics; Vocabulary; Fibromuscular Dysplasia.

Introduction

Fibromuscular dysplasia (FMD) is a heterogeneous group of idiopathic, non-atherosclerotic, relatively rare vascular diseases, leading to the narrowing of medium-sized arteries, mostly the renal and internal carotid arteries [1], [2]. The etiology and pathogenesis of this disease are unknown and there is no biochemical test for the condition. The diagnosis and classification of FMD relies today on angiographic findings. These data are collected in heterogeneous databases at different points of care. Moreover, additional data about patients are often associated with these findings but not routinely. For instance information about "Anisuporia / non-reactive pupils" is collected in the American registry (FMDSA[1]) [3], but not systematically collected in other databases. Beside this, the large diversity of cultures, laws, regulations and operational implementation regarding personal health data processes (access, gathering, sharing, etc.) across countries induces variation in the nature of the data that are collected. For instance the "race" information will be systematically present in USA registries while in European ones, the information will be "geographical origin of the parents".

Frameworks accessing and comparing distributed and heterogeneous data with consistent semantics are needed to enable cooperative research and progress in the comprehension of the disease. Our main challenge is to develop such a framework that includes an e-registry collecting standardized FMD data across Europe and that gives access and reuse of existing local FMD registries in other regions. E-registries have been developed by the authors in other domains [4][5].

We have designed a generic two-step methodology to set this framework in the context of FMD based on current Information Technology paradigms.

Nowadays, semantic interoperability is a broadly used paradigm to approach the problem of sharing data from heterogeneous data sources [6]. Many semantic interoperability platforms have been developed in various projects to apply this paradigm for data analysis by querying heterogeneous and distributed data sources. Examples are given by the EHR4CR European project [7] and the DebugIT European project [8]. The rationale of such systems is to define the intended meaning (semantics) of the data to ensure coherent interpretation by humans and processing tools, even when data are not coded the same way [9]. Some existing frameworks are based on an ontology to unify structural models and terminologies together with relevant mapping sets. This approach has been tested in the context of the EU Framework Program 7 TRANSFoRm project [10].

In the following, we describe the two-step method to develop a framework that comprises first, a European e-registry based on a pivot multidimensional model to merge existing local FMD databases in France and second, mapping resources to make the registry usable at the international level by sharing data semantics with the USA FMD registry. We present the resulting multidimensional model and the current status of the project.

Finally, we discuss some issues raised by this two-step methodology and its generalisability.

[1] http://www.fmdsa.org/research_network/fmd_registry accessed on December 14, 2014

Materials and Methods

Material

Since 2006, the French authorities have implemented a network of 131 rare disease centers of expertise in the country. The reference center for FMD is located in the "Georges Pompidou European Hospital" (HEGP), Paris, France. In collaboration with the Hypertension Unit of HEGP, the reference center started the collection of FMD data in 2009. Clinical and biological data are collected at the first visit. For patients referred to the center after FMD has been diagnosed elsewhere, clinical and biochemical data are collected at the first visit in the center if it occurred within 1 year of the diagnosis of FMD and if there had been no renal artery intervention during this period.

A French network of nephrologists, neurologists, radiologists and specialists in hypertension is responsible to document phenotypic and genetic traits of the disease and the progression of FMD lesions in patients with renal and/or cervical artery FMD. Information regarding >500 patients with FMD is currently scattered in several non-standardized, redundant databases (Figure 1).

1. The FMD local database of the HEGP reference center in Paris contains about 400 cases described each by *300 items*.
2. A database validated by the HEGP reference center and dedicated to activity reporting at the national level includes about *50 items* mostly redundant with the local database.
3. The ARCADIA/PROFILE program[2] was completed in October 2014. A specific Case Report Form was defined in this programme. The resulting database contains 500 patients with renal artery or cervical artery FMD from 12 different clinical sites (including HEGP) and described each by *200 items*[3].

In 2007, the Fibromuscular Dysplasia Society of America (FMDSA) decided to begin a registry to better understand the disease and its treatment. A template has been defined to collect *318 items*. Clinical data include elements such as date of diagnosis, types of tests conducted and results of these tests, past medical history, family history, subsequent clinical events and any clinical outcomes [2].

Besides this, several smaller studies exist in Europe to answer specific questions concerning the patients and reported in scientific papers. For these studies, specific databases, most of the time Excel files, are locally built for data management and analysis [1]. Such databases are redundant with ARCADIA/PROFILE but may include interesting additional items.

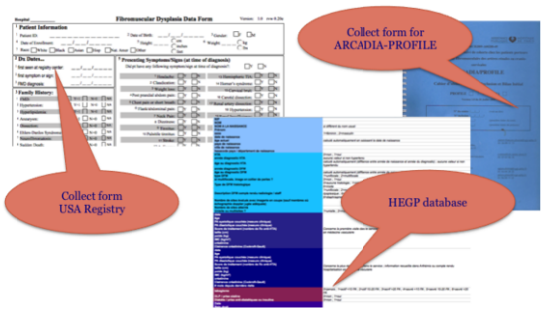

Figure 1 – Heterogeneous databases to collect FMD patients in France and USA

The HEGP FMD center wants to propose a national French e-registry based on the Case Report Form used in the ARCADIA/PROFILE cohort. It also aims at extending this registry to European countries and to making it compatible, as far as possible, with the USA registry.

Methods

Ideally, data from multiple sources are converted to standardized formats using well-characterized data vocabularies. Such approach currently requires huge efforts from experts to reach a consensus about the semantics of the domain and the standard vocabulary. It is for instance the case with the National Database for Autism Research[4] who reaches consensus about the semantics of the data.

In the first step of our methodology, a multidimensional model is set for the definition of a standard European FMD e-registry through a close collaboration between clinical experts and ontologists. This model is based on the identification of a set of core data elements for which a consensus is reachable with reasonable efforts. The second step consists of mapping all the items of the local database to the USA model. Figure 2 draws the two-step methodology.

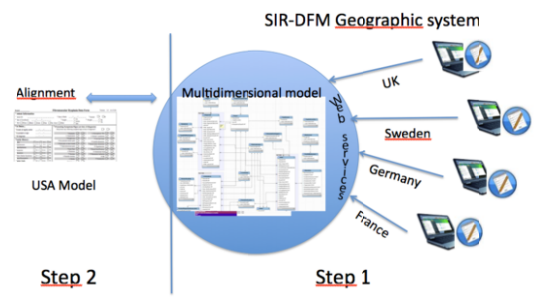

Figure 2 – The two-step methodology to set the framework

A bottom-up approach to build a multidimensional model

A data set includes data elements (DE) corresponding to some specifications which stipulate the sequence of inclusion of the DEs, whether they are mandatory, what verification rules should be employed, and the scope of the collection. A core data element (CDE) is a DE that is used in various data sets and recognized by the experts as a standard information. A specific DE (SDE) is a DE specifically defined for a given purpose (eg, the "race" is a SDE used as a public health indicator in the US) [11].

The US Office of Rare Diseases Research (ORDR) developed a set of CDEs at a national level as a tool for global collection of rare disease patient data [12]. In France, a minimum dataset

[2] ARCADIA, Assessment of Renal and Cervical Artery DysplasIA
[3] PROFILE, PROgression of FIbromuscular Lesions; Programme Hospitalier de Recherche Clinique, French Ministry of Health 2009-2014

[4] http://ndar.nih.gov. Accessed January 28, 2014.

for rare diseases has been set to respond to public health and epidemiological queries across all rare diseases [13]. These initiatives demonstrate the value of building CDE data sets. In [13], the authors argue the necessity to define core minimum data sets for each group of rare diseases in order to continue the efforts to make data interoperable for research.

Excel files and paper-based clinical Research Forms were identified and analysed by both FMD experts and ontologists to reach a consensus on a conceptual model for FMD descriptions. The target FMD model is a fixed set of CDEs. In terms of standardisation, the coding of the medical domain values can be adapted to a country's existing practices. For example, the rare disease diagnosis can be coded in Orphanet[5], OMIM, or SNOMED CT depending on the context and the required granularity of the coding. In France, diagnostic coding relies on Orphanet codes and is done by clinicians. Academic portals such as UMLS[6] or CISMeF[7] provide alignments between coding systems.

We built a set of fake cases, each case corresponding to a virtual patient, and asked a physician to describe the case using the model in order to confirm its accuracy. An information system implementing the model has been developed and is currently routinely used to feed the e-registry (Figure 2, step 1).

Using reference terminologies to align SDEs

One limit of using CDE come from the fact that, very often, data are collected at a local level for specific studies and may include specific data, semantically related to FMD but not explicitly expressed in the multidimensional model. For instance, the item "Summary_Smoking" is collected in [1] to answer a given query, it is semantically related to the item "number-of-cigarettes_daily" of the multidimensional domain but some alignment is needed. In this case, semantic interoperability can be achieved through data integration guided by an ontology [9] and in particular the approach of the type "global as view" in which an overall ontology is used as a source of mediation. However, and although there are starting efforts in the field, a domain ontology does not exist yet for FMD. In this situation, reference terminologies developed by Standard Development Organisations may be used to explicit the semantics of the data. The important initiative CTS2[8] is a specification for representing, accessing and disseminating terminological content. Moreover, many services exist to provide correspondences between these reference ressources [14].

Such an approach is used in the second step of our methodology to define mappings between reference terminologies and local interface terminologies used in local registries.

Results

Step 1 : e-registry

A standardized set of Core Data elements

The design of the methodology resulted in the following steps: (i) collecting a flat list of items as the union of the data elements and data values available from the material; (ii) performing a systematic review of the items by a group of experts; (iii) setting the different tables and value sets corresponding to the items and (iv) standardization by the eHealth team using Standard Development Organisations resources to support semantic interoperability and facilitate data re-use.

This step allows us to update the FMD nomenclature (e.g. unifocal vs. multifocal.) and to standardize the definition of FMD and FMD subtypes in order to harmonize clinical practices and enable cooperative research for this pathology.

Several face-to-face meetings were conducted with experts to reach a compromise and obtain the set of CDEs. The USA registry items were considered during this process in order to cover as many items as possible that were collected from the material. The CDE set is composed of 194 items.

The standardization of the data elements consists of:

1. Identifying concepts in reference terminologies that correspond to the items. This allows us to standardize the value sets associated to the items.
2. Choosing the prefered terms for the items and the elements in the value sets.

For example, the item "Artères" is linked to the concept "Set of Arteries" in the FMA reference ontology[9] and all elements in the value set of "Artères", such as "Coronaire Droite" are also associated to concepts in FMA (see Figure 3).

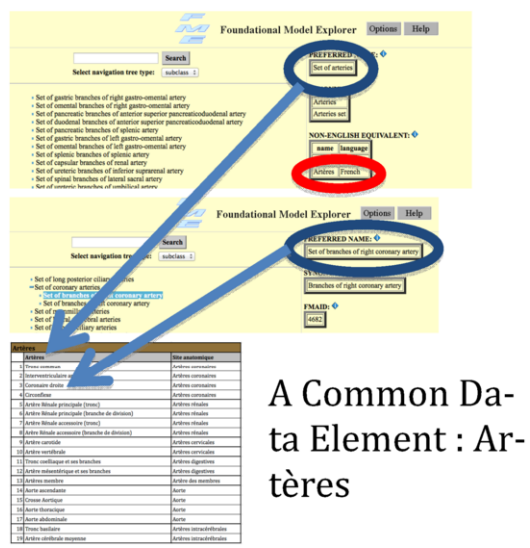

Figure 3 - Example of a common Data Element "Artères" resulting of an experts consensus and standardized with FMA concepts

The resulting multidimensional model is composed of three fact tables corresponding respectively to "Diagnosis", "Past Medical History" and "Current Situation". It is composed of 20 dimensions which are: Patients, Geographical Origin; Link; Symptoms; Signs; Vascular Events, Vascular Event Origin; Vascular Risk Factors; Type of FMD injury; Type of FMD histology; Surgery/Intervention; Type of intervention; Type of exam; Imagery Interpretation; Anatomical site, Arteries; Localisation; Drug Family; Drugs; Physical Exam; Biochimie, Familial History. Two groups of items have been designed in addition (exams; therapies) to take into account temporal and iterative information.

The Figure 4 highlights the table Diagnosis and the relations with the other tables.

[5] http://www.orpha.net/consor/cgi-bin/index.php
[6] http://www.nlm.nih.gov/research/umls/
[7] http://www.chu-rouen.fr/cismef/
[8] http://informatics.mayo.edu/cts2/index.php/Main_Page
[9] http://sig.biostr.washington.edu/projects/fm/AboutFM.html

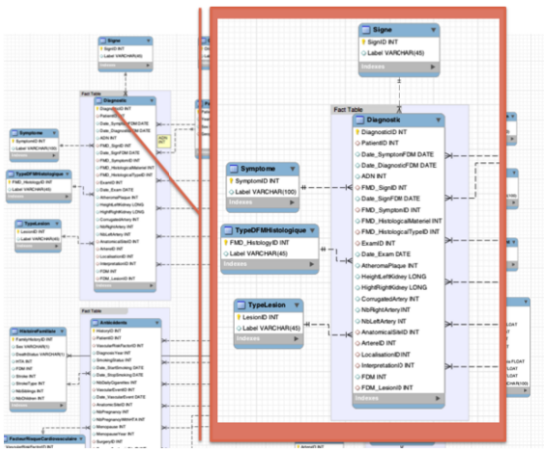

Figure 4 - Sample of the multidimensional model: the table Diagnosis

Development of the e-registry

The SIR-FMD information system was developed in 2013 with the "MetaSurv" generator described elsewhere [15]. It implemented the multidimensional model described above. A secure platform has been made available as an evaluation platform since the beginning of November 2013. On the user side, SIR-DFM relies on existing local Internet networking facilities. Via a web browser, the user connects to the interface, which is connected to the databases. This application is organized into multiple specifications that provide access to intelligent electronic forms. These electronics forms are scalable, i.e., it is possible at any time to add new items that can be used for any other patients.

SIR-DFM fulfils several requirements: scalability, portability, reliability, accessibility and cost-effectiveness oriented toward non-proprietary software. After the validation of the model, medical records of 471 patients from the databases presented in the Material section of this paper were included and are accessible through a secure user account. Users are organized into a collaborative group, and can access patient groups. The system has been presented recently in the context of the international Fibromuscular Dysplasia Research Network symposium in Cleveland, Ohio USA [16]. At first, this platform is only accessible by members of the French network, but by design, the extension of this registry to European countries is possible and depends only on the adoption of the multidimensional model at the European level.

Step 2 : Alignment process for SDEs

The proposed e-registry model includes a set of Core Data Elements with unambiguous definitions. It would be complex, due to legal and regulatory differences between Europe and USA, to settle on a single international e-registry. Some items present in the USA forms are not present in the FMD e-registry, like for instance, race or religion. One important aspect is to be able to query both data sources (to include patients in a cohort for instance) and then to establish meaningful alignments between items of the different databases that are not in the CDE dataset.

Automated schema alignment approaches have been released to avoid spending time on manual alignment to detect similarity between data elements[10]. The way to classify these alignment approaches may differ in the literature [17][18]. In our current context, the alignment cannot be automatic due to the lack of an Ontology in the FMD domain. We have analysed the potential alignments on a case by case basis. Some propositions, like the "race" item in the USA and the "geographical origin" item in France,are identified as semantically linked. At first, we manually fixed ad-hoc the similarity degree between the identified correspondences.

The mapping process has three levels of correspondences: 1) the table section (21 sections in France and 10 sections with a lot of sub-sections in USA); 2) the items level with almost double items in the USA registry; 3) the value sets.

Figure 5 shows an example of the mapping effort at the level of items and value sets. We followed the methodology described in [18]. The work is in progress and accounts for the building of the semantics of the domain.

HEGP			
Table	**num**	**items**	***Description***
Patient	1	NIP	
	2	NOM	
	3	NOM A LA	*si différent du nom usuel*
	4	Prénom	
	5	sexe	*1=féminin ; 2=masculin*
	6	date de naissance	
	7	âge actuel	
	8	pays de naissance	
	9	ville de naissance	
	10	hexacode pays / département de naissance	
USA			
Table	**num**	**items**	***Description***
Patient	1	Patient ID	
	3	Gender	*F or M*
	2	Date of Birth	*dd/mm/yyy*
	4	Date of Enrollment	*dd/mm/yyy*
	5	Height	*units: cm, inches, feet*
	6	Weight	*units:kg, lbs*
	7	Race	*white, black, asian, hisp, nat. Amer, other*

HEGP			**USA**		
Type d'examen	1	Angio-Scanner	*Exam Type*	1	Angiogram
	2	Angio-IRM		2	CCA/ICA Plaque
	3	Arthérographie		3	Ultrasond
	4	DIVA		4	Transcranial Doppler
	5	Echographie			

Figure 5 - Sample of correspondances set between the e-registry and the USA registry

Discussion and Conclusion

Fibromuscular dysplasia is a heterogeneous group of rare vascular diseases, leading to the narrowing of medium-sized arteries, mostly the renal and internal carotid arteries. Patients data are today collected in heterogeneous databases at different points of care. In order to enable cooperative research and progress in the comprehension of the disease, our challenge is to develop a framework to access and compare such distributed and heterogeneous data with consistent semantics. We propose a generic two-step methodology to

[10] ISO/EIC 11179 s Information technology - Metadata registries (MDR) - Part 3:Registry metamodel and basic attributes, Third edition 2013-02-15

develop this framework. We have fully completed the first step of the methodology. A multidimensional model was developed from existing material and working sessions with experts and ontologists. An information system was implemented, SIR-FMD, based on this dimensional model to set an e-registry. Currently, all pre-existing patients from the ARCADIA/PROFILE program are included in the e-registry. A prospect of this work is to extend this e-registry to European countries as they will increase their use of the SIR-FMD platform [19]. One important further step will be to support the integration of EHR data in the SIR-FMD platform.

The second step is in progress and needs a large group of experts. A collaboration between the French team and the FMDSA team started recently with two objectives: 1) validate the correspondences between SDEs proposed by the French team and 2) agree on semantics. More work has to be done at the level of the SIR-FMD platform to integrate all needed alignments for performing new research queries across the European e-registry and the FMDSA registry.

Finally, the methodology can be easily extended in other contexts. For instance, we are currently developing an international cohort for the ehlers danlos syndrome, a different rare disease, re-using the methodology and the "MetaSurv" generator.

Acknowledgments

We thank *Jeffrey Olin*, Professor of Medicine at the Icahn School of Medicine at Mount Sinai Hospital and Director of Vascular Medicine and the Vascular Diagnostic Laboratory in the Zena, New York, as well as *Pam Mace*, President of the Fibromuscular Dysplasia Society of America for giving us access to the USA FMDA questionnaire.

References

[1] Savard S, Steichen O, Azarine A, Azizi M, Jeunemaitre X, Plouin PF. Association between 2 angiographic subtypes of renal artery fibromuscular dysplasia and clinical characteristics. Circulation. 2012 Dec 18;126(25):3062-9.

[2] Olin JW, Froehlich J, Gu X, Bacharach JM, Eagle K, Gray BH, Jaff MR, Kim ES, Mace P, Matsumoto AH, McBane RD, Kline-Rogers E, White CJ, Gornik HL. The united states registry for fibromuscular dysplasia: Results in the first 447 patients. Circulation. 2012;125:3182-3190

[3] Poloskey SL, Olin JW, Mace P, Gornik HL. FMD Patient Page. Circulation 2012;125:e636-e639.

[4] Gimenez-Roqueplo AP, Lehnert H, Mannelli M, Neumann H, Opocher G, Maher ER, Plouin PF; European Network for the Study of Adrenal Tumours (ENS@T) Pheochromocytoma Working Group. Phaeochromocytoma, new genes and screening strategies. Clin Endocrinol (Oxf). 2006 Dec;65(6):699-705.

[5] Landais P, Messiaen C, Rath A, Le Mignot L, Dufour E, Ben Said M, Jais JP, Toubiana L, Baujat G, Bourdon-Lanoy E, Gérard-Blanluet M, Bodemer C, Salomon R, Aymé S, Le Merrer M, Verloes A; CEMARA task force. CEMARA an information system for rare diseases. Stud Health Technol Inform. 2010;160(Pt 1):481-5.

[6] Martínez-Costa C, Kalra D, Schulz S. Improving EHR semantic interoperability: future vision and challenges. Stud Health Technol Inform. 2014;205:589-93.

[7] Moor GD, Sundgren M, Kalra D, Schmidt A, Dugas M, Claerhout B, Karakoyun T, Ohmann C, Lastic PY, Ammour N, Kush R, Dupont D, Cuggia M, Daniel C, Thienpont G, Coorevits P. Using Electronic Health Records for Clinical Research: the Case of the EHR4CR Project. J Biomed Inform. 2014 Oct [Epub ahead of print]

[8] Lovis C, Colaert D, Stroetmann VN. DebugIT for patient safety - improving the treatment with antibiotics through multimedia data mining of heterogeneous clinical data. Stud Health Technol Inform. 2008;136:641-6

[9] Assélé Kama A, Choquet R, Mels G, Daniel C, Charlet J, Jaulent MC. An ontological approach for the exploitation of clinical data. Stud Health Technol Inform. 2013;192:142-6.

[10] Ethier JF, Dameron O, Curcin V, McGilchrist M, Verheij R,Arvanitis T,Taweel A, Delaney B, Burgun A. A unified structural/terminological interoperability framework based on LexEVS: application to TRANSFoRm. J Am Med Inform Assoc 2013;20:986–994.

[11] Winget MD, Baron JA, Spitz MR, et al. Development of common data elements: the experience of and recommendations from the early detection research network. 2003;5056:41–8.

[12] Forrest CB, Bartek RJ, Rubinstein Y, et al. The case for a global rare-diseases registry. Lancet 2011;377:1057–9.

[13] Choquet R, Maaroufi M, de Carrara A, Messiaen C, Luigi E, Landais P. A methodology for a minimum data set for rare diseases to support national centers of excellence for healthcare and research. J Am Med Inform Assoc. 2014

[14] Grosjean J, Merabti T, Soualmia LF, Letord C, Charlet J, Robinson PN, Darmoni SJ. Integrating the human phenotype ontology into HeTOP terminology-ontology server. Stud Health Technol Inform. 2013;192:961.

[15] Toubiana L, Moreau S, Bonnard G. MetaSurv: Web-platform generator for the monitoring of health's indicators and interactive geographical information system. Stud Health Technol Inform. 2005;116:989-93.

[16] P.-F.Plouin, L. Toubiana, M.-C. Jaulent, X. Jeunemaitre French FMD registry: organizational structure, data elements, summary of findings. International Fibromuscular Dysplasia (FMD) Research Network symposium May15th and 16th 2014 Cleveland, Ohio USA

[17] Rahm E, Bernstein PA. A survey of approaches to automatic schema matching, The VLDB Journal 10, 334–350, 2001.

[18] Maaroufi M, Choquet R, Landais P, Jaulent MC. Formalizing mappings to optimize automated schema alignment: application to rare diseases. Stud Health Technol Inform. 2014;205:283-7.

[19] Persu A, Giavarini A, Touzé E, Januszewicz A, Sapoval M, Azizi M, Barral X, Jeunemaitre X, Morganti A, Plouin PF, de Leeuw P; ESH Working Group Hypertension and the Kidney. European consensus on the diagnosis and management of fibromuscular dysplasia. J Hypertens. 2014 Jul;32(7):1367-78