Highlights

Ontology-based integration and querying of heterogeneous rare disease data sources – POLVAS perspective

Wojciech Palacz,Sabina Lichołai,Jacek Musiał,Katarzyna Wawrzycka-Adamczyk,Grażyna Ślusarczyk,Barbara Strug,Beyza Yaman,Michelangelo Tesi,Karl Gisslander,Declan O'Sullivan,Augusto Vaglio,Giacomo Emmi,Mark A. Little,Krzysztof Wójcik

- The FAIRVASC project integrates seven registries concerned with the rare disease of anti-neutrophil cytoplasmic antibody (ANCA) associated vasculitis (AAV).
- Heterogeneous data from registries are converted into unified form defined by an AAV-specific FAIRVASC ontology.
- This article describes the conversion process of the POLVAS registry, which uses a Python preprocessor in addition to an R2RML mapping.
- Each registry has their own SPARQL server with their 'uplifted' data. To combine these datasets, federated queries are used.

Ontology-based integration and querying of heterogeneous rare disease data sources – POLVAS perspective

Wojciech Palacz[a,*], Sabina Lichołai[b], Jacek Musiał[c], Katarzyna Wawrzycka-Adamczyk[c], Grażyna Ślusarczyk[a], Barbara Strug[a], Beyza Yaman[d], Michelangelo Tesi[e], Karl Gisslander[g], Declan O'Sullivan[d], Augusto Vaglio[f,e], Giacomo Emmi[h,i,j], Mark A. Little[d] and Krzysztof Wójcik[c]

[a]*Institute of Applied Computer Science, Jagiellonian University, ul. Łojasiewicza 11, 30-048 Kraków, Poland*
[b]*Division of Molecular Biology and Clinical Genetics, Jagiellonian University Medical College, ul. Skawińska 8, 31-066 Kraków, Poland*
[c]*2nd Department of Internal Medicine, Jagiellonian University Medical College, ul. Jakubowskiego 2, 30-688 Kraków, Poland*
[d]*ADAPT Centre for Digital Content, O'Reilly Institute, Trinity College Dublin, Dublin 2, Ireland*
[e]*Nephrology and Dialysis Unit, Azienda Ospedaliera Universitaria Meyer IRCCS, Firenze, Italy*
[f]*Department of Biomedical, Experimental and Clinical Sciences, University of Florence, Firenze, Italy*
[g]*Department of Clinical Sciences – Rheumatology, Lund University, Lund, SE-221 85, Sweden*
[h]*Department of Medical, Surgical and Health Sciences, University of Trieste, Trieste, Italy*
[i]*Clinical Medicine and Rheumatology Unit, Cattinara University Hospital, Trieste, Italy*
[j]*Centre for Inflammatory Diseases, Monash University Department of Medicine, Monash Medical Centre, Melbourne, Australia*

ARTICLE INFO

Keywords:
rare diseases
ontologies
data integration
federated queries

ABSTRACT

The integration of rare disease medical databases belonging to different countries is an important problem, as a large number of observations are required for reliable statistical inference of patient data in order to facilitate clinical research. Such integration of national registry data, which requires harmonisation of the heterogeneous data sets into a unified view, is facilitated in the European FAIRVASC project by developing a domain-specific ontology. The FAIRVASC project is dedicated to the rare disease of anti-neutrophil cytoplasmic antibody (ANCA) associated vasculitis (AAV). This paper focuses on the practical issues and challenges, encountered during the process of integrating the Polish national database POLVAS into the federated database within the FAIRVASC project. It discusses the use of ontology-based methods for data integration, addressing differences in data formats, language barriers and the importance of ensuring patient privacy and data protection. The modifications of mappings used to 'uplift' national data into the Resource Description Framework (RDF) triplestore are also proposed.

1. Introduction

In Europe, a rare disease is a one which affects less than 500 per million inhabitants [17]. Over 6000 different rare diseases have been identified to date but not all of them are covered by active European registries [17, 12]. The low prevalence of rare diseases means that patient numbers are small in any given country. As a sufficiently large number of observations (patients, symptoms, treatment results, etc.) are required for reliable statistical inference of patient data, the small sizes of data in existing registries are a major barrier to clinical research [24]. Moreover, these data are often fragmented across multiple heterogeneous isolated regional disease registries. Therefore it is important to federate national and local registries that are emerging for many rare diseases. The access to multiple rare disease datasets and the possibility to use them as a single pool of data will lead to new research opportunities.

Several projects striving for integration of heterogeneous resources have been developed. A specific network called AutoInflammatory Disease Alliance (AIDA) [1] has recently been created to develop international registries of rare autoinflammatory diseases. A web platform is used to collect demographic, genetic, clinical and treatment data of patients suffering from nine autoinflammatory diseases, namely Behçet syndrome (BS), PFAPA (Periodic Fever, Aphthous Stomatitis, Pharyngitis and Adenitis), Schnitzler's syndrome, Scleritis, Still disease, Undifferentiated Autoinflammatory Diseases, Uveitis, VEXAS (Vacuoles, E1 enzyme, X-linked, Autoinflammatory, Somatic) and Monogenic Autoinflammatory Diseases [19]. An AIDA registry focusing on BS [33] is a single multicenter shared registry, which allows all participants to use the same platform for data entry and for using the same variables that have a standardized format. In this way, the registry guarantees the uniformity of data, which can be analyzed without discrepancies across databases and facilitates the querying process.

*Corresponding author
wojciech.palacz@uj.edu.pl (W. Palacz); sabina.licholai@cm-uj.krakow.pl (S. Lichołai); jacek.musial@uj.edu.pl (J. Musiał); grazyna.slusarczyk@uj.edu.pl (G. Ślusarczyk); barbara.strug@uj.edu.pl (B. Strug); beyza.yaman@adaptcentre.ie (B. Yaman); michelangelo.tesi@meyer.it (M. Tesi); karl.gisslander@med.lu.se (K. Gisslander); declan.osullivan@adaptcentre.ie (D. O'Sullivan); augusto.vaglio@unifi.it (A. Vaglio); giacomo.emmi@unifi.it (G. Emmi); mlittle@tcd.ie (M.A. Little); krzysztof.wojcik@uj.edu.pl (K. Wójcik)
ORCID(s): 0000-0002-5143-5595 (W. Palacz); 0000-0001-7008-2492 (S. Lichołai); 0000-0002-0955-1808 (J. Musiał); 0000-0002-1439-1474 (K. Wawrzycka-Adamczyk); 0000-0003-1032-1644 (G. Ślusarczyk); 0000-0002-2204-507X (B. Strug); 0000-0003-2130-0312 (B. Yaman); 0000-0001-5500-9991 (M. Tesi); 0000-0003-1915-2656 (K. Gisslander); 0000-0002-3814-9172 (A. Vaglio); 0000-0001-9575-8321 (G. Emmi); 0000-0001-6003-397X (M.A. Little); 0000-0001-9786-2499 (K. Wójcik)

The AVERT model dedicated to anti-Glomerular Basement Membrane disease, which uses RDF [13] to convert multiple files of various formats into a single queryable data source, is described in [28].

One of the rare diseases is anti-neutrophil cytoplasmic antibody (ANCA) associated vasculitis (AAV), a serious systemic autoimmune disorder characterized by blood vessel inflammation and destruction with a European annual incidence of around 20 per million inhabitants [25]. The clinical symptoms of AAV are varied, ranging from skin rash to severe involvement of the kidneys and lungs [24]. Moreover, the disease is often relapsing and has a high morbidity so patients are at risk of serious complications both from the disease and treatment. The European Vasculitis Society (EUVAS) Registries Group (led by Prof. Mark Little, TCD) has identified over 11,000 AAV cases in European national registries. Under the auspices of FAIRVASC, which is a research project [4] of EUVAS and RITA European Reference Network, AAV data from different national registries is integrated to identify features that predict how a patient's illness will develop, and what their major health risks are [24]. The seven registries, which are brought together, are Ireland's Rare Kidney Disease (RKD) registry, the Italian (Italivas) registry, the French Vasculitis Study Group registry (GVEF), the Czech Registry of ANCA-associated vasculitis, the Polish Vasculitis Registry (POLVAS), the GEVAS German/Austrian/Swiss registry and Sweden's Skåne Vasculitis Cohort (SVR).

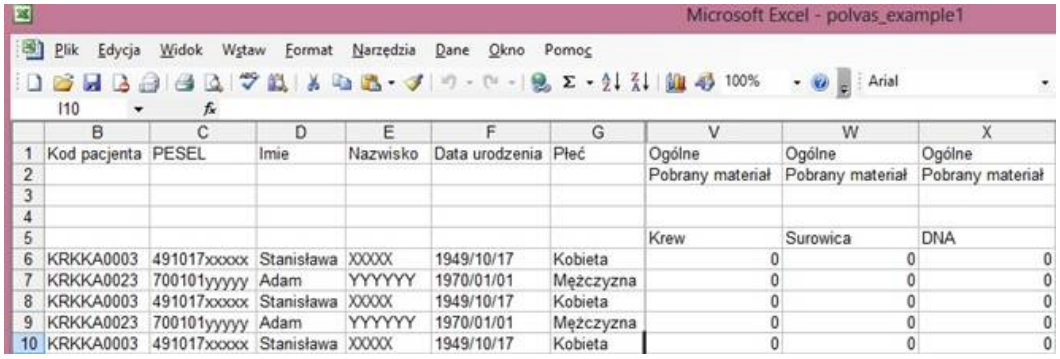
These registries vary with respect to the data elements captured, representation, structure and semantics of the data, and they describe diverse patient populations. As there is still a lack of standardization across registries [15], data analytics over multiple registries is not possible without a process of harmonization and normalisation of registries. Therefore the main goal of the FAIRVASC project is to integrate data at a semantic level, so that it is possible to query over registries and return comparable results. The provided novel infrastructure, which can readily incorporate additional registries, ensures access to harmonised data relating to patients with rare diseases, while also meeting data privacy and security concerns. The management of the data related to AAV adheres to FAIR principles [3, 21] meaning that data be findable, accessible, interoperable and reusable. The FAIRVASC approach is based on a federated architecture of registries, where registry data remain at each partner site, and the approach promotes the sustainability of decentralized components. The proposed solution to data integration has been developed within the context of larger European Union wide initiatives to coordinate activities which overcome challenges related to data fragmentation of European medical registries, such as the European Joint Programme (EJP) on Rare Diseases, which aims to improve healthcare for rare disease patients. EJP has developed a virtual platform architecture, i.e., a service oriented ecosystem of interlinked web services, to foster and facilitate research in the rare diseases domain [32, 30].

An effective approach to support interoperability of medical data is based on ontologies [16, 23] which may be developed to meet the requirements of different areas of health care [18, 27]. Well-known ontologies for medical data include Open Biological Biomedical Ontology (OBO) Foundry [8], the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT) ontology, the Medical Dictionary for Regulatory Activities Terminology (MedDRA) ontology, National Cancer Institute Thesaurus (NCIt) ontology [7] and Orphanet Rare Disease ontology (ORDO), all of which are available on BioPortal [6]. Apart from these general health domain ontologies there are also domain specific ontologies like the domain specific ontology for adverse reactions (DSOAE), developed and applied specifically to chronic kidney disease (CKD) [22], the Chronic Kidney Disease Ontology (CKDO) describing clinical features associated with CKD [16] or the Kidney Tissue Atlas Ontology (KTAO) [26].

The integration of registry data, which requires harmonisation of the heterogeneous data sets into a unified view, is facilitated in the FAIRVASC project by developing the FAIRVASC ontology [5]. The FAIRVASC ontology is domain specific for AAV and makes use of the NCIT and ORDO ontologies, which cover a wide range of terms which are relevant to AAV, but it is light weight compared to the other domain specific ontologies. The ontology has been built upon a set of key clinical questions developed by a team of experts in vasculitis selected from the registry sites. These are expressed as "competency questions" of the ontology. Thus it has been developed with a strong alignment to the existing seven AAV national registries that participate in the federation. The need to involve clinical researchers to identify competency queries in the rare disease registry integration framework is highlighted in [20].

In [24] the methodology for developing the ontology, in particular the harmonisation of data terms across registries and alignment with available ontologies is presented. It also describes the method for adding semantic meaning to AAV data across the registries using the declarative Relational to Resource Description Framework Mapping Language (R2RML) [11, 31, 29]. The mappings used to 'uplift' tabular data (CSV files) into Resource Description Framework (RDF) using the R2RML and SPARQL queries [14] for performing federated queries over the resulting data sets are discussed. By using federated queries which aggregate patient data (i.e. returning counts of patients rather than patients' data), combined with an uplift process into local RDF registries, where data is pseudo-anonymised before it is converted into RDF, the approach ensures sensitive data is not inadvertently exposed in response to queries that may be made.

In this paper we illustrate the subsequent stages of using semantic web technologies in the FAIRVASC project from the perspective of the POLVAS registry. The extensions which have been undertaken with respect to the FAIRVASC ontology and changes of R2RML mappings required to integrate the Polish national database into the European



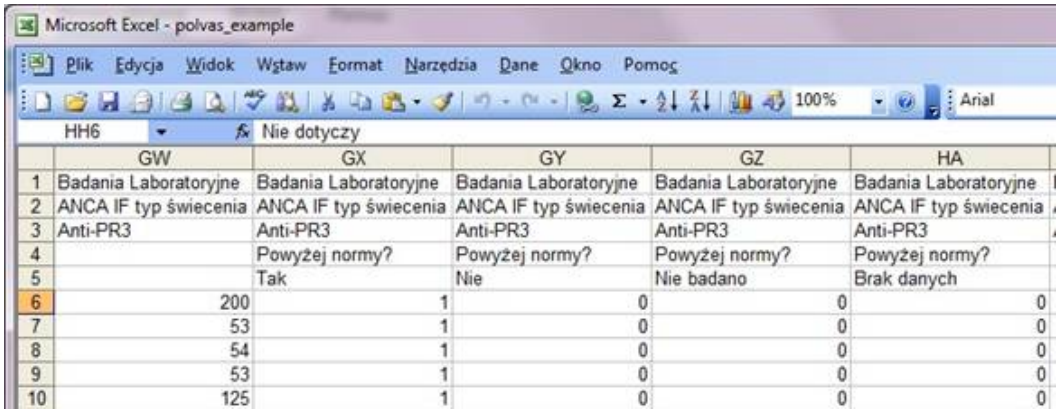
Microsoft Excel - polvas_example1

Plik Edycja Widok Wstaw Format Narzędzia Dane Okno Pomoc

I10

	B	C	D	E	F	G	V	W	X
1	Kod pacjenta	PESEL	Imie	Nazwisko	Data urodzenia	Płeć	Ogólne	Ogólne	Ogólne
2							Pobrany materiał	Pobrany materiał	Pobrany materiał
3									
4									
5							Krew	Surowica	DNA
6	KRKKA0003	491017xxxxx	Stanisława	XXXX	1949/10/17	Kobieta	0	0	0
7	KRKKA0023	700101yyyyy	Adam	YYYYYY	1970/01/01	Mężczyzna	0	0	0
8	KRKKA0003	491017xxxxx	Stanisława	XXXX	1949/10/17	Kobieta	0	0	0
9	KRKKA0023	700101yyyyy	Adam	YYYYYY	1970/01/01	Mężczyzna	0	0	0
10	KRKKA0003	491017xxxxx	Stanisława	XXXX	1949/10/17	Kobieta	0	0	0

Figure 1: Selected entries related to fictional patients' data in the POLVAS registry.



Microsoft Excel - polvas_example

Plik Edycja Widok Wstaw Format Narzędzia Dane Okno Pomoc

HH6 Nie dotyczy

	GW	GX	GY	GZ	HA
1	Badania Laboratoryjne	Badania Laboratoryjne	Badania Laboratoryjne	Badania Laboratoryjne	Badania Laboratoryjne
2	ANCA IF typ świecenia	ANCA IF typ świecenia	ANCA IF typ świecenia	ANCA IF typ świecenia	ANCA IF typ świecenia
3	Anti-PR3	Anti-PR3	Anti-PR3	Anti-PR3	Anti-PR3
4		Powyżej normy?	Powyżej normy?	Powyżej normy?	Powyżej normy?
5		Tak	Nie	Nie badano	Brak danych
6	200	1	0	0	0
7	53	1	0	0	0
8	54	1	0	0	0
9	53	1	0	0	0
10	125	1	0	0	0

Figure 2: Entries related to laboratory tests of ANCA IF pattern in the POLVAS registry.

network are presented. The problems related to data privacy and security are briefly mentioned, as these issues were thoroughly discussed in [24], where also all related work concerning the implemented approach was described.

2. Harmonisation of terms in data registries

The integration of heterogeneous data is possible due to the use of a domain-specific ontology to which different concepts occurring in the national registries are mapped. The FAIRVASC ontology focusing on AAV has been developed in the FAIRVASC project [24] under the direction of domain experts across seven European registries. It is based on other ontologies, like the NCIT and ORDO, which cover a wide range of terms relevant to AAV. The experts in vasculitis disease have developed competency questions, which refer to data existing in national registries. These questions constrain the scope of knowledge to be represented in an ontology. On this basis, classes, properties and relations of the FAIRVASC ontology are defined.

Two important areas for AAV, one related to exposure and the second to outcomes, have been identified for focus. Features related to exposure, which have been included in the FAIRVASC ontology, are age, gender, ANCA auto-antibody subtype, AAV diagnosis, affected organ pattern, induction treatment, maintenance treatment and serum creatinine level at presentation. Features related to important clinical outcomes are death, end-stage kidney disease, serious complication such as severe infection, cancer and cardiovascular disease.

After defining the competency questions, the national registries have been analysed and searched for data which can be used to answer these questions. The seven data registries (Ireland, Italy, France, Germany, Sweden, Czech Republic, and Poland) provided data dictionaries and sets of simulated data generated for testing purposes. The data dictionaries typically were represented in spreadsheets or CSV files. Data dictionaries generated on the basis of national registries vary in the types of column names reflecting their own internal database structures. The column names correspond to terms in the data dictionaries.

POLVAS registry collects retrospective data, whereas the majority of FIARVASC registries are prospective observation. It should be noticed, that in retrospective observation each patient has one questionnaire characterizing the whole disease course, whereas in prospective registries there is the set of questionnaires over time. This difference leads to a different approach to data processing for answering queries. For example, in POLVAS registry the CRP at time of diagnosis is defined and this value is taken directly from the questionnaire, whereas in prospective registries, where each patient has more than one encounter, the CRP results must be related to the date of diagnosis to identify the proper value.

```
:Badania Laboratoryjne :ANCA IF typ świecenia:Anti-PR3:::
486 x ' ', 259 x '0', 233 x 'Nie dotyczy', 115 x '200', 38 x '2', 22 x '0.2', 11 x '1', 11 x '3', 9 x '31', 8 x '100', etc.
:Badania Laboratoryjne :ANCA IF typ świecenia:Anti-PR3:Powyżej normy?: Brak danych:
1346 x '0', 232 x 'Nie dotyczy', 155 x ' ', 45 x '1'
:Badania Laboratoryjne :ANCA IF typ świecenia:Anti-PR3:Powyżej normy?: Nie badano:
1307 x '0', 232 x 'Nie dotyczy', 155 x ' ', 84 x '1'
:Badania Laboratoryjne :ANCA IF typ świecenia:Anti-PR3:Powyżej normy?: Nie:
841 x '0', 550 x '1', 232 x 'Nie dotyczy', 155 x ' '
:Badania Laboratoryjne :ANCA IF typ świecenia:Anti-PR3:Powyżej normy?: Tak:
712 x '1', 679 x '0', 232 x 'Nie dotyczy', 155 x ' '
```

Figure 3: Statistics related to laboratory tests of ANCA IF pattern in the real POLVAS registry.

Another example is retrieving information about induction and maintenance treatment. In POLVAS registry these data are explicitly defined, whereas in prospective observation the treatment in encounter must be related to the time of diagnosis, although in some cases it may be difficult to define the induction/maintenance therapy periods. On the other hand, in retrospective observation if patient has more than one induction therapy and many different drugs are used, it is not possible to find out what was the treatment for 1st or 2nd relapse. The prospective type of data collection – many time points with different values of one parameter requires different approach to select the proper encounter as a data source. Again, if we consider the question about CRP at time of diagnosis, the time of diagnosis needs to be correlated with dates of different encounters to choose the closest one.

In the case of the Polish registry the data dictionary has the form of a CSV file. Two fragments of this file, converted to .xls format for clarity of presentation, are shown in Figs. 1 and 2. In Fig. 1 several entries related to fictional patients' data, like patient code, PESEL number, name, surname, date of birth, gender, collected material (blood, serum, DNA), in Polish language can be seen. In Fig. 2 column names are related to laboratory tests of ANCA IF pattern. The first column from the left specifies the result of ANCA IF Anti-PR3 for each patient, while the next four columns indicate whether the result was above the norm, where the possible values 0 and 1 in the subsequent columns evaluate the answers yes, no, not tested and no data available. Moreover, there is another file which gives statistics for data stored in a real registry. A fragment of this file is shown in Fig. 3. From the Figure, it can be seen that out of 1,778 patients, 115 had the result 200 of Anti-PR3, 8 had the result of 100, 9 had the result of 31, etc. The last two rows confirm that 712 patients had the result above the norm, 679 had the result within the norm, 232 people were not affected by this study, and for 155 there was no data to analyse.

The other column names in the Polish registry are:

- patient's address,
- date of consent to the study,
- material collected (blood, serum, DNA, urine, biopsy),
- clinical diagnosis (Takayasu's arteritis, giant cell arteritis, polyarteritis nodosa, Kawasaki's disease, microscopic polyangiitis (MPA), granulomatosis with polyangiitis (Wegener's) (GPA), eosinophilic granulomatosis with polyangiitis (Churg and Stauss)(EGPA), Anti-basement membrane antibody disease (Goodpasture syndrome), cryoglobulinemia-associated vasculitis, IgA-associated vasculitis (Henoch and Schönlein), urticarial vasculitis hypocomplementemic vasculitis (anti-C1q associated vasculitis), Behçet's disease vasculitis, Cogan's syndrome vasculitis, cutaneous leukocytoclastic vasculitis, cutaneous arteritis, primary central nervous system vasculitis, isolated aortitis, lupus vasculitis, rheumatoid vasculitis, sarcoid vasculitis, cryoglobulinemic vasculitis associated with HCV infection, vasculitis associated with HBV infection, syphilis aortitis, drug-related immune complex vasculitis, tumor-related vasculitis, vasculitis of probable known cause, unspecified large-vessel vasculitis, unspecified medium-sized vasculitis, ANCA-associated unspecified small vessel vasculitis, unspecified small vessel vasculitis with immune complex deposition, unspecified miscellaneous vasculitis, other),
- date of diagnosis,
- probability of diagnosis (certain, probable, uncertain),
- diagnosis based on histology, clinical symptoms, radiological tests, laboratory tests, other additional tests,
- date of onset of first symptoms,
- occupation at the beginning of the disease (not working, pension, full-time job, part-time job, retiree, student, housewife),
- smoking (smokes, has smoked in the past, never smoked, no information, number of pack-years),
- clinical manifestation of vasculitis (general symptoms - yes, no, no data; musculoskeletal system - yes, no, no data, skin - yes, no, no data; organ of vision - yes, no, no data; ear/nose/throat - yes, no, no data; respiratory system - yes, no, no data; cardiovascular system - yes, no, no data; digestive system - yes, no, no data; kidneys - yes, no, no data; urogenital system - yes, no, no data; CSN involvement - yes, no, no data; neurological system - yes, no, no data; other symptoms),

- course of the disease (total number of exacerbations, time between the diagnosis of the disease and the first exacerbation (months), exacerbations requiring hospitalization - yes, no, no data; exacerbations requiring a stay in the ICU - yes, no, no data; treatments/surgeries since the onset of the disease - organ transplant, PCI, PTA, vascular surgery, amputation, other, no, no data; other, maximum creatinine concentration (value, unit), maximum CRP concentration [mg/l] at the time of diagnosis, patient's death (yes, no, date of death, cause of death - infection, cancer, vasculitis, cardiovascular event, other, no data),

- treatment (remission-inducing treatment used - no data, oral glucocorticosteroids, IV pulse glucocorticosteroids, cyclophosphamide, plasmapheresis, rituximab, MMF, azathioprine, methotrexate, leflunomide, others; other, maintenance treatment used (from achieving remission or for >6 months) - no data, oral glucocorticoids, cyclophosphamide, rituximab, MMF, azathioprine, methotrexate, leflunomide, cyclosporine, IVIG, other, none; other, pulses >= 0.5 g - 1-3, 3-6, 6-9, > 9, no, no data; cumulative duration of steroid use in years, route of administration - po, iv, po+iv, no data; total cumulative dose [g], total cumulative dose [mg], plzamapheresis - yes, no, no data; dialysis patient - yes, no, temporarily, no data),

- laboratory tests (ANCA IF light type - C, P, atypical, present, absent, not tested, no data; Anti-PR3 - above the norm; Anti-MPO - above the norm - yes, no, not tested, no data; Anti-GBM - not applicable, above the norm - yes, no, not tested, no data; cryoglobulins - not applicable, above the norm - yes, no, not tested, no data; peripheral blood eosinophilia (>10% or >1,560 /ul) - yes, no, not tested, no data, numerical value),

- biopsy (biopsy date, biopsy site - skin, kidney, lung, eye, ear/nose/throat, aorta, temporal artery, other artery, peripheral nerve, muscle, liver, intestine, heart, brain or meninges, urinary bladder, bone marrow, lymph node, other; general result - normal, non-diagnostic/non-specific for vasculitis, supports vasculitis but is not fully characteristic, vasculitis),

- complications of immunosuppressive treatment (cutaneous - yes, no, no data; hematological - yes, no, no data; infection - yes, no, no data; digestive tract - yes, no, no data; autoimmune - yes, no, no data; neurologic - yes, no, no data; malignant tumors - yes, no, no data; organ of vision/ear/nose/throat - yes, no, no data; heart/lungs - yes, no, no data; kidneys - yes, no, no data; general symptoms - yes, no, no data).

The first step in analysis of the data dictionaries was to determine in what way the terms represented align with the competency questions gathered. Having a good understanding of the column names in the simulated data and the types of values that were being stored in corresponding cells was vital. The members of the FAIRVASC project (the Harmonisation Implementation Team (HIT) described in [24]) have identified a set of harmonised terms across the registries, expressed via the FAIRVASC ontology.

3. FAIRVASC ontology for AAV

Harmonisation of nomenclature used in national registries enabled the design and implementation of the FAIRVASC OWL ontology [9]. Classes, their relationships and data properties were defined using the Protégé tool [10]. The version of the ontology described in [24] includes 9 top level classes, 9 object properties, and 24 data properties. The FAIRVASC ontology also imported the Birmingham Vasculitis Activity Score (BVAS) classes from a related ontology developed in parallel with FAIRVASC. Top level classes were *Patient, PatientOverview, Diagnosis, ClinicalOutcomes, Encounter, ClinicalTest, OrganPattern, Complication* with four subclasses (*ComplicationCardiovascular, ComplicationInfection, ComplicationMalignancy, ComplicationThromboembolism*), and *ANCA* (anti-neutrophil cytoplasm antibodies). Object properties were: *hasPatientOverview, hasClinicalOutcome, hasEncounter, hasDiagnosis, hasANCA, hasOrganPattern, hasBVAS, hasClinicalTest, hasComplication*.

In the next version of the ontology, the class *Encounter* and the property *hasEncounter*, describing the relation between the *Patient* and his *Encounter*, were removed. Therefore, the domain of the relations *hasBVAS* and *hasClinicalTest* has been changed to *Patient*. Moreover, new classes, like *Treatment* (with two subclasses *TreatmentMaintenance* and *TreatmentInduction*), *Comorbidities, BiobankConsent, RegistryConsent*, and *FutureContactConsent*, new object properties, like *hasComorbidities, hasTreatment, treatmentType*, and new data properties like *biobankName* have been added. In the current version, the FAIRVASC ontology [5] contains 20 top level classes, 22 subclasses, 27 object properties, and 38 data properties.

The key subject in FAIRVASC for all data registry records is still the *Patient*. Each patient has a relation to their *PatientOverview, Complication, ClinicalTest*, and *ClinicalOutcomes*. The class *Patient* has data properties *patientID, yearOfBirth*, and *dateOfEncounter*, as well as object property *gender*. The *PatientOverview* represents an overview of the patient health in relation to vasculitis, which consists of the patient's current diagnosis, their ANCA, the way in which their organs have been affected over the course of their disease, comorbidities, and the received treatment. It has data properties *BMI* (body mass index), *height* and *weight*, and is related to the classes *Diagnosis* (through the object property *hasDiagnosis*), *ANCA* (through the object property *hasANCA*), *OrganPattern* (through the object property *hasOrganPattern*), *Comorbidities* (through the object property *hasComorbidities*), *Treatment* (through the object property *hasTreatment*), and *CreatinineTest* (through the object property *hasCreatinineAtDiagnosis*), which is a

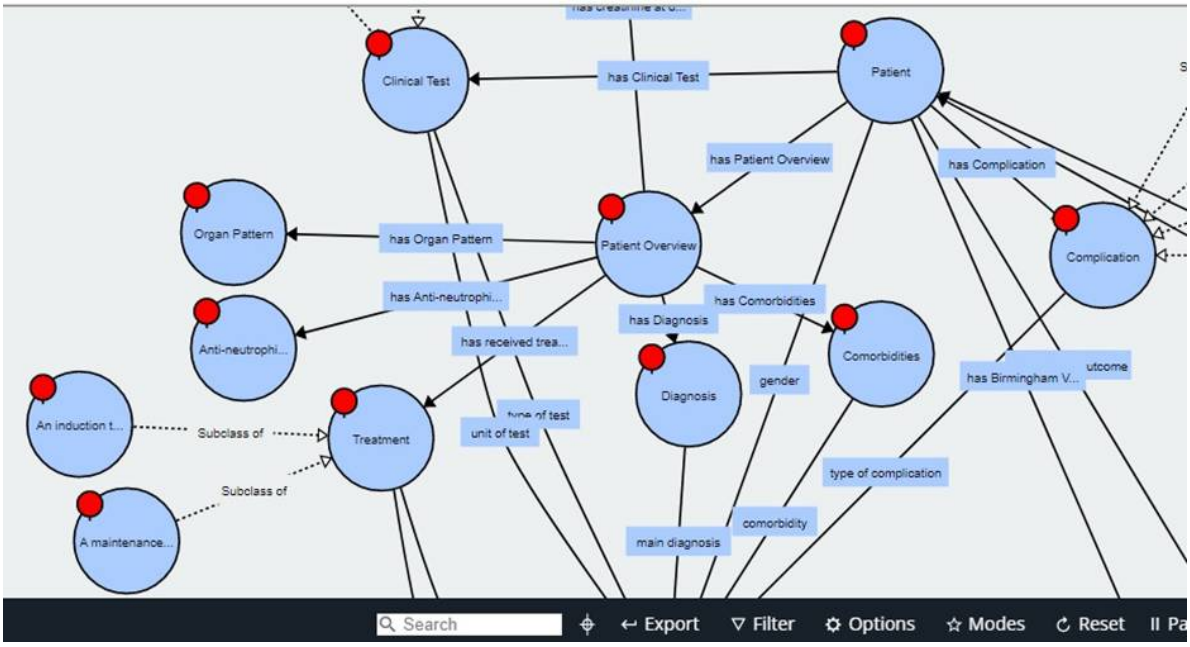


Figure 4: A fragment of the FAIRVASC ontology.

subclass of the class *ClinicalTestNorm*, which in turn is a subclass of a top level class *ClinicalTest*.

The top level class *Diagnosis* describes a patient's diagnosis of vasculitis. It has eight data properties *ageAtDiagnosis, ageAtOnset, dateOfCarePathwayStart, dateOfDiagnosis, dateOfOnset, diagnosisClassification, diagnosisConfidence, inclusionType* and one object property *mainDiagnosis*. The top level class *ANCA* describes autoantibodies that target neutrophils, which are a type of human white blood cells important for fighting infection. In ANCA-associated vasculitis, patients usually have autoantibodies against proteinase 3 (PR3-ANCA) or myeloperoxidase (MPO-ANCA). The class *OrganPattern* describes organs which have been affected by vasculitis. The class *Comorbidities* describes comorbidities a patient had when they have been enlisted to the registry. The class *Treatment* describes the treatments for AAV received by a patient, suppressing the immune system, which is inappropriately activated in AAV. This immunosuppressive treatment is typically considered as either 'induction' or 'maintenance' therapy. The class *CreatinineTest* describes creatinine levels determined by creatinine tests conducted at an encounter. Blood levels of creatinine are used as a measure of kidney function, as creatinine is a waste product produced by muscles and is excreted from the body by the kidneys.

In Fig. 4 a fragment of the FAIRVASC ontology is presented. This figure illustrates the relationships between top level classes, where the class *Patient* is related to *PatientOverview, ClinicalTest*, and *Complication*, while *PatientOverwiev* is related to *Diagnosis, Comorbidities, ANCA, OrganPattern*, and *Treatment* having two subclasses (only the relations which can be seen in the figure are listed). The Figure was generated via the Protégé tool.

4. Data mapping into RDF instances

In order to assign semantic meaning to AAV data in the registries, CSV files containing tabular data are converted into RDF files using R2RML (Relational to Resource Description Framework Mapping Language) language [24]. This mapping language allows to relate the structure of the tabular data to the ontology vocabulary, by specifying how data from the non-RDF resources should be transformed into RDF graphs. In this way data of each national registry are transformed into RDF instances and added to an RDF triplestore. These stores are then used as data sources for SPARQL queries that support implementation of the competency questions. The returned results are aggregated, so only counts of patients with specific information are given, thus preserving the privacy of registry patient data.

R2RML mappings were developed by the FAIRVASC Implementation Team (FIT), and then tested on simulated data, as the different versions of the ontology were constructed. As each registry differs in types of data and data formats, the mappings needed to be locally modified to work with the local registry.

The first important modification for the POLVAS registry was the use of a Python preprocessor to reorganize the original FIT mapping. The mappings in other registries apply pre-processing of the data with the use of SQL queries before mapping the selected values from the CSV data to ontology subjects and predicates. In POLVAS the mapping does not use SQL queries, but all condition checking and calculation of values required by the FAIRVASC ontology is executed by the Python preprocessor.

For example, in the POLVAS file the creatinine level can be stored both in umol/l and mg/dl units, while FAIRVASC requires umol/l. In Listing 1 a fragment of the POLVAS

```
# POLVAS stores the maximal creatinine level in
# umol/l or in mg/dl, FAIRVASC requires umol/l.
if orig.creati_value == "":
    prepr.creatinine_umol_l = ""
else:
    try:
        x = float(orig.creati_value)
        match orig.creati_umol, orig.creati_mg:
            case "1", ("0"|""|"Nie dotyczy"):
                prepr.creatinine_umol_l = str(x)
            case ("0"|""|"Nie dotyczy"), "1":
                x = 88.4 * x
                prepr.creatinine_umol_l = str(x)
    except ValueError:
        pass
```

Listing 1: A fragment of the processor code for converting the test results into umol/l units.

```
# POLVAS collects only one creatinine test result per
# patient. It has no date, because it represents the
# maximal result obtained during all tests the patient
# was subjected to.
rr:logicalTable [
    rr:sqlQuery """
        SELECT * FROM polvas_preprocessed WHERE
        creatinine_umol_l IS NOT NULL
    """ ;
] ;
rr:predicateObjectMap [
    rr:predicate fvc:testValue ;
    rr:objectMap [
        rr:column "CREATININE_UMOL_L" ;
        rr:datatype xsd:double ;
    ] ;
] ;
```

Listing 2: Fragments of the R2RML mapping related to the value of the creatinine level converted to umol/l.

processor code, which is responsible for converting test results into the appropriate units, is presented. In Listing 2 fragments of the R2RML mapping produced by the preprocessor and related to the value converted using code from Listing 1 can be seen. Another example is related to the proper matching of patient's diagnosis. In the FAIRVASC ontology only three types of diagnosis are acceptable, i.e, GPA, MPA and EGPA, while POLVAS contains also other types of disorder for some patients. In this case, the preprocessor discards records of patients with impermissible diagnosis ensuring that all records sent to the RDF store contain only acceptable data for FAIRVASC purposes.

In some cases there is no information in the national registry corresponding to the concepts present in the FAIRVASC ontology. The POLVAS Python preprocessor deals with missing information by aggregating other data available within the database. Some columns are created as the result of aggregating some other columns from the POLVAS csv file have been added to the database. For example, the ANCA ELISA property is not present in the POLVAS registry, but an appropriate column has been added to the csv file to calculate this information. Also the date of the last contact with a patient is not available in the POLVAS registry. Therefore the data of the last contact is calculated as either the date of the death of the patient or the last time the patient's information has been updated.

Another modification for the POLVAS registry case is related to the format of the generated URLs being a format for addressing resources on the Internet. As illustrated in Listing 3, each mapping shows the required prefixes being name spaces of the different vocabularies used in the mapping process. As far as prefix *data* is concerned (for Poland it is `http://w3id.org/FAIRVASC/resource/polvas/`), instead of hierarchically building the ending, like patient/ID or patient/patientoverview/ID, in our mapping the endings are constructed only from letters, numbers and the underscore, like patient_ID or overview_ID. This modification allows to naturally use the @prefix directive in the resulting file, which makes it much more readable, as in the RDF instances there are strings, for example "patient_7" and "patient_12", instead of the previous "7" and "12". In Listing 3 a fragment of the mapping of patient data in the POLVAS registry is shown.

The next modification concerns the simplification of the generated RDF graph. If many patients are treated by the same medicine with the same route of administration, in mappings of other registries, like RKD or SVR, as many copies of one subgraph will be generated in the RDF graph as there are such number of patients. In the POLVAS mapping case, no such duplication is present, and each treatment type is treated as a singleton with no patient number in the URLs identifying them. In Listing 4 a fragment of the mapping which relates the patient to the inductive treatment by cyclophosphamide taken orally is shown. The second part of the mapping in this listing allows to specify relation between cyclophosphamide, its treatment type (ncit:C405), and its administration route (ncit:C38288). In Listing 5 a fragment of the RDF graph obtained by mapping data from our CSV file is presented. It shows the overview of two patients. It can be seen that none of them took a cyclophosphamide orally in the induction treatment. In the first case a cyclophosphamide has been administered intravenously in the induction treatment, while in the second case a cyclophosphamide has been also taken orally, but in the maintenance treatment.

```
@prefix rr: <http://www.w3.org/ns/r2rml#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix data: <http://w3id.org/FAIRVASC/resource/polvas/> .
@prefix fvc: <http://w3id.org/FAIRVASC#> .
@prefix ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#> .
@prefix dpv: <https://w3id.org/dpv/dpv-owl#> .

<#join_on_id>
        rr:child "id" ;
        rr:parent "id" ;
.
<#patient>
        rr:logicalTable [
                rr:tableName "polvas_preprocessed" ;
        ] ;
        rr:subjectMap [
                rr:template "http://w3id.org/FAIRVASC/resource/polvas/patient_{POLVAS_ID}" ;
                rr:class fvc:Patient , dpv:Patient ;
        ] ;
        rr:predicateObjectMap [
                rr:predicate fvc:patientID ;
                rr:objectMap [
                        rr:column "POLVAS_ID" ;
                        rr:datatype xsd:string ;
                ] ;
        ] ;
        rr:predicateObjectMap [
                rr:predicate fvc:yearOfBirth ;
                rr:objectMap [
                        rr:column "BIRTH_YEAR" ;
                        rr:datatype xsd:gYear ;
                ] ;
        ] ;
        rr:predicateObjectMap [
                rr:predicate fvc:gender ;
                rr:objectMap [
                        rr:template "http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#{GENDER_NCIT}" ;
                ] ;
        ] ;
        rr:predicateObjectMap [
                rr:predicate fvc:hasPatientOverview ;
                rr:objectMap [
                        rr:parentTriplesMap <#overview> ;
                        rr:joinCondition <#join_on_id> ;
                ] ;
        ] ;
        rr:predicateObjectMap [
                rr:predicate fvc:hasClinicalOutcome ;
                rr:objectMap [
                        rr:parentTriplesMap <#outcome> ;
                        rr:joinCondition <#join_on_id> ;
                ] ;
        ] ;
```

Listing 3: A fragment of the mapping of patient data in the POLVAS registry.

```
<#overview>
        rr:logicalTable [
                rr:tableName "polvas_preprocessed" ;
        ] ;
        rr:subjectMap [
                rr:template "http://w3id.org/FAIRVASC/resource/polvas/overview_{POLVAS_ID}" ;
                rr:class fvc:PatientOverview ;
        ] ;
        rr:predicateObjectMap [
                rr:predicate fvc:hasTreatment ;
                rr:objectMap [
                        rr:parentTriplesMap <#ind_cyclophosphamide_oral> ;
                        rr:joinCondition <#join_on_id> ;
                ] ;
        ] ;

<#ind_cyclophosphamide_oral>
        rr:logicalTable [
                rr:sqlQuery """
                        SELECT id FROM polvas_preprocessed WHERE
                        ind_cyclophosphamide_oral = 1
                """ ;
        ] ;
        rr:subjectMap [
                rr:constant data:ind_trt_cyclophosphamide_oral ;
                rr:class fvc:TreatmentInduction ;
        ] ;
        rr:predicateObjectMap [
                rr:predicate fvc:treatmentType ;
                rr:object ncit:C405 ;
        ] ;
        rr:predicateObjectMap [
                rr:predicate fvc:routeToAdministration ;
                rr:object ncit:C38288 ;
        ] ;
```

Listing 4: A fragment of the mapping which relates the patient to the inductive treatment by cyclophosphamide taken orally.

5. Federated questions over RDF triple stores

The generated RDF data is uploaded into a triplestore which supports SPARQL federated queries. The Apache Jena Fuseki [2] was chosen by FAIRVASC team as a triplestore because it is a free and open source Java framework, is easy to install and provides a web client for testing queries. Patients' data are private and they are not shared without the explicit authority given from the registry, as RDF triplestores are located on registries' local networks.

The interface was implemented using Django, NGINX, HTML5, JavaScript and Bootstrap. It currently supports basic federated querying of simulated data on the registries, which includes queries related to counts of patients stratified by gender, ANCA specificity and diagnosis. A given federated query is sent from a central server to each registry triplestore and the results are aggregated, to return only counts.

In Fig. 5 a FAIRVASC interface with a query related to counts of patients with selected options dividing the results according to gender and ANCA specificity, and selecting them from the POLVAS registry is shown, while the results of this query are shown in Fig. 6. In Fig. 7 options selected for a treatment information query are shown. This query is related to counts of patients having cyclophosphamide administered orally in the induction treatment and stored in GFEV (France) and Italivas (Italy) registries. The results of this query, which are stratified by ANCA specificity, are presented in Fig. 8. Listing 6 illustrates the SPARQL query which selects and returns these data from the chosen registries. It should be noted that results below the count of 10 are ignored, so as to ensure anonymity of patient data. In Fig. 9 the results of the same query, i.e., related to counts of patients having cyclophosphamide administered orally in the induction treatment, but returned only from the POLVAS registry are shown. There are only 14 patients listed out of 37 stored in the Polish registry, who have cyclophosphamide administered orally in the induction treatment. In all these 14 cases the ANCA specificity is 'PR3 positive'. The number

```
data:ind_trt_cyclophosphamide_iv a fvc:TreatmentInduction ;
        fvc:routeToAdministration ncit:C38276 ;
        fvc:treatmentType ncit:C405 .

data:ind_trt_cyclophosphamide_oral a fvc:TreatmentInduction ;
        fvc:routeToAdministration ncit:C38288 ;
        fvc:treatmentType ncit:C405 .

data:mnt_trt_cyclophosphamide_oral a fvc:TreatmentMaintenance ;
        fvc:routeToAdministration ncit:C38288 ;
        fvc:treatmentType ncit:C405 .

data:overview_KRKKA0003 a fvc:PatientOverview ;
        fvc:hasANCA data:anca_if_c_spec_pr3 ;
        fvc:hasDiagnosis data:diagnosis_KRKKA0003 ;
        fvc:hasTreatment data:ind_trt_cyclophosphamide_iv ;
        fvc:hasTreatment data:ind_trt_glucocorticoids_oral ;
        fvc:hasTreatment data:ind_trt_rituximab_iv ;
        fvc:hasTreatment data:mnt_trt_glucocorticoids_oral .

data:overview_KRKKA0023 a fvc:PatientOverview ;
        fvc:hasANCA data:anca_if_c_spec_pr3 ;
        fvc:hasDiagnosis data:diagnosis_KRKKA0023 ;
        fvc:hasTreatment data:ind_trt_cyclophosphamide_iv ;
        fvc:hasTreatment data:ind_trt_glucocorticoids_iv ;
        fvc:hasTreatment data:mnt_trt_cyclophosphamide_oral ;
        fvc:hasTreatment data:mnt_trt_glucocorticoids_oral .
```

Listing 5: A fragment of the RDF graph obtained by mapping data from the POLVASC registry.

of patients with other ANCA types is less than 10, therefore they are not included in the obtained results.

The development of formal queries is a challenging task as it requires aligning clinical knowledge of the local registry experts with query capabilities in order to define queries which are relevant and important for physicians. Knowing what data clinical experts expect from the queries drives both the development of the ontology and of the front end of the query interface.

6. Discussion

Launching the rare disease registry is a complex process that requires not only clinical skills and expertise, but among others requires strict cooperation with IT and legal/ethical experts. POLVAS collects questionnaire data since 2016. However, process of legal authorization of the registry by university and national authorities started in 2014. It should be also noticed that during last 8 years we were experiencing several legal changes which required from us to comply with the new regulations. Changes in data governance rules required, among others, continuous modification and update in the patient information and consent forms. Based on our experience and discussions with our EU partners following major problems while launching a registry were identified:

- Legal issues related to the permission to run a registry for research purposes by an academic entity with a data set tailored to the research purposes encompassing a specific disease/group of strictly related diseases, e.g. AAV.
- Corresponding IT infrastructure, with regard to safety and accessibility of sensitive health records. Properly equipped and skilled hospital/university dedicated IT unit.
- Strict protection of the acquisition and storage of subject level data with the clear and specific rules to share aggregated data for research purposes. Aggregated data generated under FAIRVASC infrastructure are one of the possible answers which may help to overcome EU countries limitations in this regard.

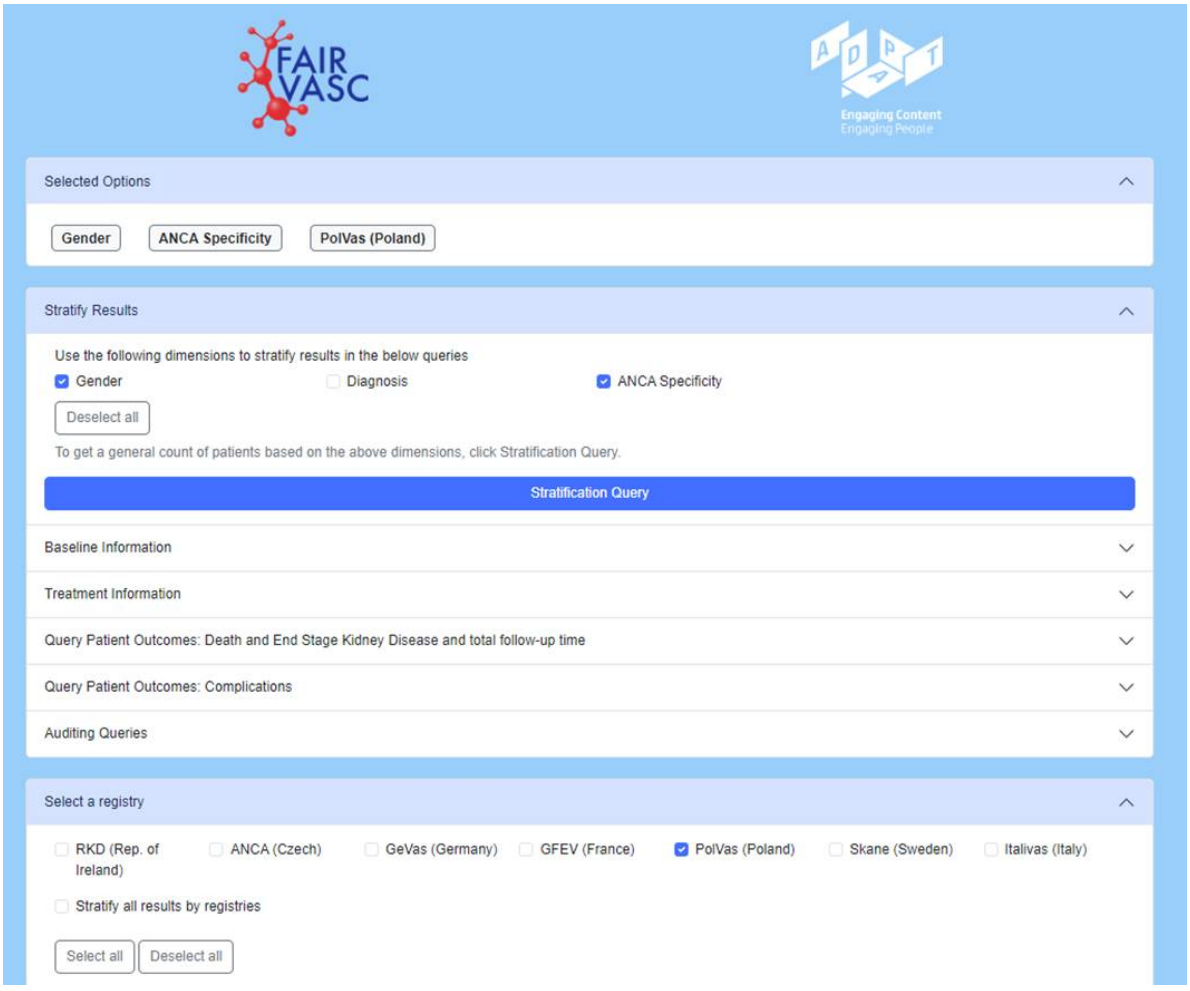
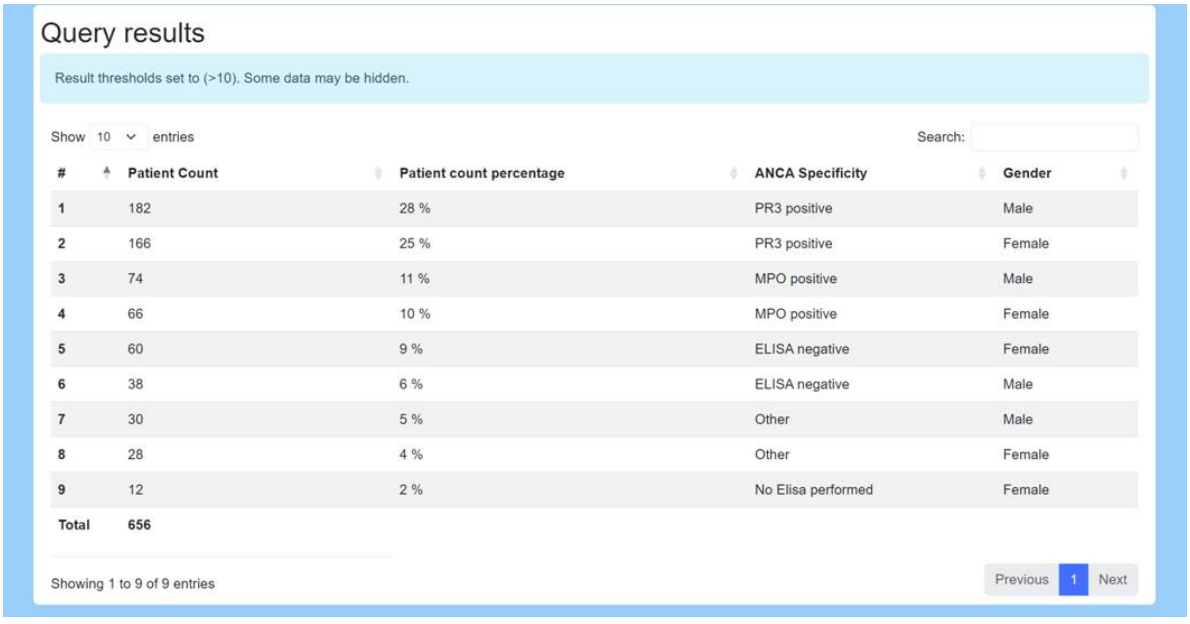


Figure 5: A FAIRVASC interface with selected options of a query related to gender and ANCA specificity of patients selected from the POLVAS registry.



#	Patient Count	Patient count percentage	ANCA Specificity	Gender
1	182	28 %	PR3 positive	Male
2	166	25 %	PR3 positive	Female
3	74	11 %	MPO positive	Male
4	66	10 %	MPO positive	Female
5	60	9 %	ELISA negative	Female
6	38	6 %	ELISA negative	Male
7	30	5 %	Other	Male
8	28	4 %	Other	Female
9	12	2 %	No Elisa performed	Female
Total	656			

Figure 6: The results of the query related to counts of patients with options selected in Fig. 5.

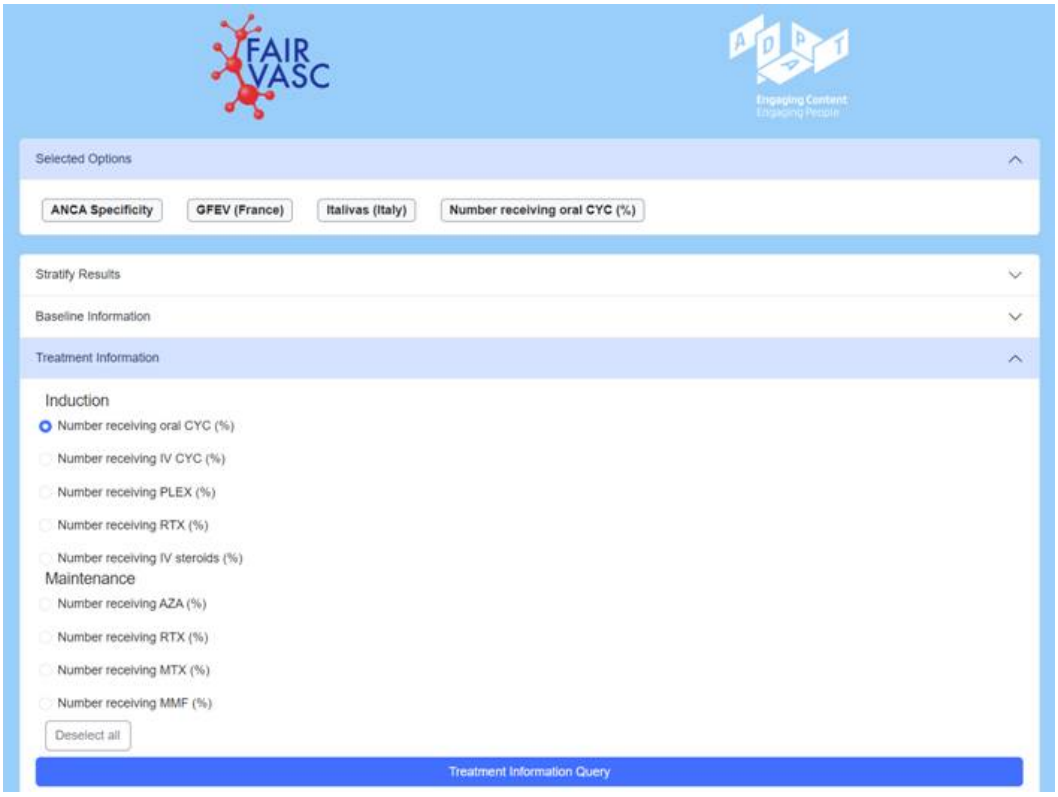


Figure 7: Options selected for a query related to counts of patients having cyclophosphamide administered orally in the induction treatment and stored in GFEV (France) and Italivas (Italy) registries.

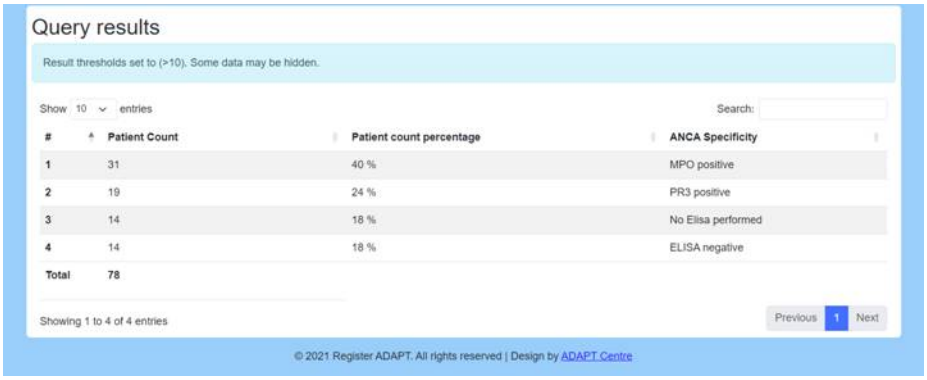


Figure 8: The results of the query related to the number of patients having cyclophosphamide administered orally (Fig. 7) stratified by ANCA specificity.

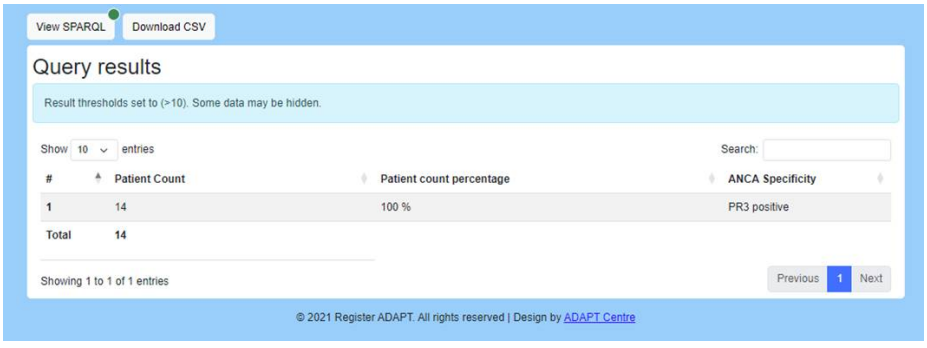


Figure 9: The results of the query related to the number of patients having cyclophosphamide administered orally (Fig. 7) returned from the POLVAS registry.

```
SELECT  (SUM(?total_n) AS ?totalPatientCount)  ?anca

WHERE{
  {
    SERVICE <https://fairvasc.tentelemed.com/FAIRVASC_SIMULATED_DS1/query>
    {
      SELECT (COUNT(DISTINCT ?patient) AS ?total_n)   ?anca

      WHERE{
        BIND( http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C405 AS ?type)
        BIND( http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C38288 AS ?route)
       ?patient a fvc:Patient.
       ?patient fvc:hasPatientOverview ?patientOverview.
       ?patientOverview fvc:hasTreatment ?Treatment.
       ?Treatment a fvc:TreatmentInduction.
       ?Treatment fvc:treatmentType ?type.
       ?Treatment fvc:routeToAdministration ?route.
        BIND('GFEV(France)' AS ?registry)
       ?patientOverview fvc:hasDiagnosis ?dia.
       ?dia fvc:mainDiagnosis ?mainDiagnosis.
       ?patientOverview fvc:hasANCA ?hasAnca.
       ?hasAnca fvc:ancaSpec ?anca.

      }  GROUP BY ?total_n ?anca
    }
  }
  UNION
  {
    SERVICE <https://fairvasc.meyer.it:3043/italivas/query>
    {
      SELECT (COUNT(DISTINCT ?patient) AS ?total_n)   ?anca

      WHERE{
        BIND( http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C405 AS ?type)
        BIND( http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C38288 AS ?route)
       ?patient a fvc:Patient.
       ?patient fvc:hasPatientOverview ?patientOverview.
       ?patientOverview fvc:hasTreatment ?Treatment.
       ?Treatment a fvc:TreatmentInduction.
       ?Treatment fvc:treatmentType ?type.
       ?Treatment fvc:routeToAdministration ?route.
        BIND('Italivas(Italy)' AS ?registry)
       ?patientOverview fvc:hasDiagnosis ?dia.
       ?dia fvc:mainDiagnosis ?mainDiagnosis.
       ?patientOverview fvc:hasANCA ?hasAnca.
       ?hasAnca fvc:ancaSpec ?anca.

      }  GROUP BY ?total_n ?anca
    }
  }
}  GROUP BY ?totalPatientCount ?anca

HAVING (?totalPatientCount > 10)
```

Listing 6: The SPARQL query for returning patients having cyclophosphamide administered orally in the induction treatment stratified by ANCA specificity and stored in GFEV (France) and Italivas (Italy) registries.

7. Conclusions

In this paper the methodology of using semantic web technologies in order to integrate the POLVAS registry with other registries in the FAIRVASC project was presented. The modifications of the R2RML mapping needed to cope with different types of data and different data formats in which information is stored in the POLVAS registry were described.

The method of integrating national registries for rare diseases is based on the FAIRVASC ontology. Federating heterogeneous data sources is important as a sufficiently large amount of observations is required for reliable statistical inference needed to conduct clinical research. The information coming from several national registries may inform on disease characteristics and potential pathogenic mechanisms. An easy access to a large amount of information through one interface can expand current clinical knowledge on AAV [21], as well as inspire new clinical trials to test novel treatments.

Semantic Web technologies allow human-computer interoperability and facilitate the automation of data analysis. The use of a common ontology onto which data from the registers of individual countries are mapped solves both the problem of semantic discrepancies across databases and of the diverse nomenclature used in different countries. Moreover, the proposed methods allow to easily incorporate additional national registries into the federated database. Presenting the modifications that, from the perspective of the Polish registry, should be introduced to simplify the mapping process, may help teams from other countries that would like to add their registries to the FAIRVASC project.

References

[1] AIDA Network – autoinflammatory diseases alliance, accessed 6 Mar 2024. URL: https://www.aidanetwork.org/.

[2] Apache Jena Fuseki – a SPARQL server, accessed 6 Mar 2024. URL: https://jena.apache.org/documentation/fuseki2/.

[3] FAIR Principles, accessed 6 Mar 2024. URL: https://www.go-fair.org/fair-principles/.

[4] FAIRVASC – building registry interoperability to inform clinical care, accessed 6 Mar 2024. URL: https://fairvasc.eu/.

[5] The FAIRVASC ontology, accessed 6 Mar 2024. URL: http://w3id.org/FAIRVASC.

[6] NCBO BioPortal – the most comprehensive repository of biomedical ontologies, accessed 6 Mar 2024. URL: https://bioportal.bioontology.org/ontologies.

[7] NCIt – NCI (National Cancer Institute) thesaurus, accessed 6 Mar 2024. URL: https://ncithesaurus.nci.nih.gov/.

[8] OBO Foundry – open biological and biomedical ontology foundry, accessed 6 Mar 2024. URL: https://obofoundry.org/.

[9] OWL: Web ontology language, accessed 6 Mar 2024. URL: https://www.w3.org/TR/owl2-overview/.

[10] Protégé ontology editor, accessed 6 Mar 2024. URL: https://protege.stanford.edu/.

[11] R2RML: RDB to RDF mapping language, accessed 6 Mar 2024. URL: https://www.w3.org/TR/2012/REC-r2rml-20120927/.

[12] Rare disease registries in Europe (Orphanet report series, December 2021), accessed 6 Mar 2024. URL: https://www.orpha.net/orphacom/cahiers/docs/GB/Registries.pdf.

[13] RDF: Resource Description Framework, accessed 6 Mar 2024. URL: https://www.w3.org/TR/rdf11-concepts/.

[14] SPARQL: Query language for RDF, accessed 6 Mar 2024. URL: https://www.w3.org/TR/sparql11-query/.

[15] Blumenthal, S., 2018. Improving interoperability between registries and EHRs, in: AMIA Summits on Translational Science Proceedings, pp. 20–25.

[16] Cole, N.I., Liyanage, H., Suckling, R.J., Swift, P.A., Gallagher, H., Byford, R., Williams, J., Kumar, S., de Lusignan, S., 2018. An ontological approach to identifying cases of chronic kidney disease from routine primary care data: a cross-sectional study. BMC Nephrology 19, 85. doi:10.1186/s12882-018-0882-9.

[17] Deliverska, M., 2016. Patient registries for rare diseases. Journal of IMAB 22, 1166–1168. doi:10.5272/jimab.2016222.1166.

[18] Dimitrieski, V., Petrović, G., Kovačević, A., Luković, I., Fujita, H., 2016. A survey on ontologies and ontology alignment approaches in healthcare, in: Trends in Applied Knowledge-Based Systems and Data Science, pp. 373–385. doi:10.1007/978-3-319-42007-3_32.

[19] Gelain, E., Tesi, M., Mazzariol, M., Vaglio, A., 2023. Registries of rare diseases: current knowledge and future perspectives. Internal and Emergency Medicine 18, 19–21. doi:10.1007/s11739-022-03151-1.

[20] Gisslander, K., Mohammad, A.J., Vaglio, A., Little, M.A., 2023. Overcoming challenges in rare disease registry integration using the semantic web – a clinical research perspective. Orphanet Journal of Rare Diseases 18, 253. doi:10.1186/s13023-023-02841-z.

[21] Gisslander, K., Rutherford, M., Aslett, L., Basu, N., Dradin, F., et al., 2024. Data quality and patient characteristics in European ANCA-associated vasculitis registries: data retrieval by federated querying. Annals of the Rheumatic Diseases 83, 112–120. doi:10.1136/ard-2023-224571.

[22] Kang, Y., Fink, J.C., Doerfler, R., Zhou, L., 2018. Disease specific ontology of adverse events: Ontology extension and adaptation for chronic kidney disease. Computers in Biology and Medicine 101, 210–217. doi:10.1016/j.compbiomed.2018.08.024.

[23] McGlinn, K., Hussey, P., 2020. An analysis of demographic data in Irish healthcare domain to support semantic uplift, in: Computational Science – ICCS 2020, pp. 456–467. doi:10.1007/978-3-030-50423-6_34.

[24] McGlinn, K., Rutherford, M.A., Gisslander, K., Hederman, L., Little, M.A., O'Sullivan, D., 2022. FAIRVASC: A semantic web approach to rare disease registry integration. Computers in Biology and Medicine 145, 105313. doi:10.1016/j.compbiomed.2022.105313.

[25] Mohammad, A.J., 2020. An update on the epidemiology of ANCA-associated vasculitis. Rheumatology 59, iii42–iii50. doi:10.1093/rheumatology/keaa089.

[26] Ong, E., Wang, L.L., Schaub, J., O'Toole, J.F., Steck, B., et al., 2020. Modelling kidney disease using ontology: insights from the Kidney Precision Medicine Project. Nature Reviews Nephrology 16, 686–696. doi:10.1038/s41581-020-00335-w.

[27] Rath, A., Olry, A., Dhombres, F., Brandt, M.M., Urbero, B., Ayme, S., 2012. Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. Human Mutation 33, 803–808. doi:10.1002/humu.22078.

[28] Reddy, B.P., Houlding, B., Hederman, L., Canney, M., Debruyne, C., O'Brien, C., Meehan, A., O'Sullivan, D., Little, M.A., 2019. Data linkage in medical science using the resource description framework: the AVERT model. HRB Open Research 2019, 1:20. doi:10.12688/hrbopenres.12851.2.

[29] Rodríguez-Muro, M., Rezk, M., 2015. Efficient SPARQL-to-SQL with R2RML mappings. Journal of Web Semantics 33, 141–169. doi:10.1016/j.websem.2015.03.001.

[30] dos Santos Vieira, B., Bernabé, C.H., Zhang, S., Abaza, H., Benis, N., et al., 2022. Towards FAIRification of sensitive and fragmented rare disease patient data: challenges and solutions in European reference network registries. Orphanet Journal of Rare Diseases 17, 436. doi:10.1186/s13023-022-02558-5.

[31] Sequeda, J.F., 2013. On the semantics of R2RML and its relationship with the direct mapping, in: ISWC 2013 Posters & Demonstrations Track, pp. 193–196.

[32] Sernadela, P., González-Castro, L., Carta, C., van der Horst, E., Lopes, P., et al., 2017. Linked registries: Connecting rare diseases patient registries through a semantic web layer. BioMed Research International 2017, 8327980. doi:10.1155/2017/8327980.

[33] Vitale, A., Della Casa, F., Ragab, G., Almaghlouth, I.A., Lopalco, G., et al., 2022. Development and implementation of the AIDA International Registry for patients with Behçet's disease. Internal and Emergency Medicine 17, 1977–1986. doi:10.1007/s11739-022-03038-1.