Computers in Biology and Medicine 145 (2022) 105313

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

FAIRVASC: A semantic web approach to rare disease registry integration

Kris McGlinn [a,*], Matthew A. Rutherford [c], Karl Gisslander [d], Lucy Hederman [a], Mark A. Little [a,b], Declan O'Sullivan [a]

[a] ADAPT, Trinity College Dublin, Ireland
[b] TTMI, Trinity Health Kidney Centre, Ireland
[c] Institute of Infection, Immunity and Inflammation, University of Glasgow, United Kingdom
[d] Department of Rheumatology, Lund University, Sweden

ARTICLE INFO

Keywords:
Knowledge engineering
Linked data
Ontologies
Federated queries
Rare diseases

ABSTRACT

Rare disease data is often fragmented within multiple heterogeneous siloed regional disease registries, each containing a small number of cases. These data are particularly sensitive, as low subject counts make the identification of patients more likely, meaning registries are not inclined to share subject level data outside their registries. At the same time access to multiple rare disease datasets is important as it will lead to new research opportunities and analysis over larger cohorts. To enable this, two major challenges must therefore be overcome. The first is to integrate data at a semantic level, so that it is possible to query over registries and return results which are comparable. The second is to enable queries which do not take subject level data from the registries. To meet the first challenge, this paper presents the FAIRVASC ontology to manage data related to the rare disease anti-neutrophil cytoplasmic antibody (ANCA) associated vasculitis (AAV), which is based on the harmonisation of terms in seven European data registries. It has been built upon a set of key clinical questions developed by a team of experts in vasculitis selected from the registry sites and makes use of several standard classifications, such as Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) and Orphacode. It also presents the method for adding semantic meaning to AAV data across the registries using the declarative Relational to Resource Description Framework Mapping Language (R2RML). To meet the second challenge a federated querying approach is presented for accessing aggregated and pseudonymized data, and which supports analysis of AAV data in a manner which protects patient privacy. For additional security the federated querying approach is augmented with a method for auditing queries (and the uplift process) using the provenance ontology (PROV-O) to track when queries and changes occur and by whom. The main contribution of this work is the successful application of semantic web technologies and federated queries to provide a novel infrastructure that can readily incorporate additional registries, thus providing access to harmonised data relating to unprecedented numbers of patients with rare disease, while also meeting data privacy and security concerns.

1. Introduction

Exchange of healthcare data between people, organizations and services, is seen as a key factor in achieving quality patient care and enabling novel clinical research [2,23]. Yet, health care data remains fragmented and siloed [14,15,25]. Of the data that is digitized, it is rarely standardised from the perspective of data interoperability, meaning it does not adhere to any particular standardised terminologies, schema or syntax [25]. This is true for healthcare registries which store patient data regarding specific diseases, as there is still a lack of standardization across registries [4]. Without a process of harmonisation/normalisation of registries, data analytics over multiple registries is not possible, as the structure and semantics of the data may differ.

Ontologies have been recognised as an effective approach to support interoperability of medical data [8,20] and many have been developed to meet the requirements of different areas of health care [11], such as those made available over BioPortal[1] and the Open Biological and

[*] Corresponding author.
E-mail address: kris.mcglinn@adaptcentre.ie (K. McGlinn).

[1] https://bioportal.bioontology.org/ontologies.

https://doi.org/10.1016/j.compbiomed.2022.105313
Received 21 September 2021; Received in revised form 3 February 2022; Accepted 8 February 2022
Available online 1 March 2022

Biomedical Ontology (OBO) Foundry.[2] Ontologies make use of the Web of Data,[3] which is an initiative to make data open and interconnected, stored and shared across the World Wide Web using a well established architecture, the semantic web stack [32]. Central to this is the concept of Linked Data, a way of structuring and sharing data on the web based on the Resource Description Framework (RDF). By defining data models using semantic web technologies, it becomes possible to make data schemas available using standard web access mechanisms, e.g. Hyper-text Transfer Protocol (HTTP).

Once a data schema is described and published using an ontology (or as an RDF vocabulary), it resides on the web and any data described using LD can be associated with this ontology (or vocabulary) so that the semantics of the data are open and freely available to a global audience. Combined with the SPARQL Protocol and RDF Query Language (SPARQL), an RDF query language, publishing data in this manner provides a powerful tool for sharing and re-using data and has the potential to provide improved support for interoperability across health care data.

Anti-neutrophil cytoplasmic antibody (ANCA) associated vasculitis (AAV) is a serious systemic autoimmune disorder characterized by blood vessel inflammation and destruction. The clinical symptoms are varied, ranging from skin rash to severe involvement of the kidneys and lungs. Despite advances in the treatment of AAV improving the rate of survival, the disease is often relapsing. Patients are at risk of serious complications both from the disease and treatment. It therefore has a high morbidity, placing a heavy burden on the patients and healthcare systems [26]. However, AAV is rare, with a European annual incidence of around 20 per million inhabitants [21]. Patient numbers are therefore small in any given country. As a sufficiently large number of observations are required for reliable statistical inference of patient data, the small cohort sizes in existing AAV registries are a major barrier to clinical research. This makes the publication of data especially relevant for rare diseases, such as AAV.

Added together, over 6000 vasculitis cases in different European clinical registries have previously been identified related to AAV [3], of which the majority are AAV cases. Under the auspices of the FAIRVASC project[4] AAV data from seven registries is being brought together to identify features (clinical and physical characteristics, etc.) that predict how a patient's illness will develop, and what their major health risks are. These registries are; Ireland's Rare Kidney Disease (RKD) registry, the United Kingdoms(UK) UKIVAS 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 Register. The different registries vary with respect to the data elements captured, representation and structure of the data, data about provenance, and they describe diverse patient populations.

The aim of this work is therefore to determine if the application of a methodology and set of technologies can support federated queries over sensitive patient data across multiple rare disease registries. The resulting framework should be flexible and extensible so as to be applicable to any registry. We set out to demonstrate this through the development and implementation of the FAIRVASC ontology, mappings, and infrastructure for supporting federated queries across seven AAV registries. This requires harmonisation of the heterogeneous data sets into a unified view of the data, the FAIRVASC ontology. As the medical data about patients is of a sensitive nature, and data governance legislation discourage its movement across organisational or national boundaries, only anonymised data within each registry can be exposed. A combination of local access control and project-wide code-of-practice for data governance is therefore required. This is partially addressed during the uplift process, by pseudo-anonymising data before it is converted into RDF and also at the query level through the use of federated queries, and by limiting the types of queries, so that subject level data is aggregated.

In this paper we present the first iteration of the FAIRVASC ontology, which is an ontology for managing data related to AAV and which adheres to FAIR principles,[5] meaning that data be findable, accessible, interoperable and reusable. It 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 on Rare Diseases (EJP RD), 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. They identified five key services related to; authentication and authorization, services for discovery and elaboration of rare disease data, services to request and access data, services for dataset enhancement (e.g. pseudonymization) and services for making resources FAIR for federated use, e.g. catalogues of ontological models for RD data and metadata. One of the core design principles of EPJ is to support a federated architecture (so registry data can remain on site) which a: adheres to General Data Protection Regulation (GDPR) from a resource perspective and b: fosters sustainability of the decentralized components. EPJ relies on use case driven design, and a strong emphasis on the use of common data representations, such as ontologies and their reuse. The work within this paper fits well into the overall architecture of the EPJ.

The FAIRVASC ontology development is built upon a set of core competency questions to drive research into AAV. These were identified by experts in AAV across the pilot registries. The paper presents the methodology for developing the ontology, in particular the harmonisation of data terms across registries and alignment with available ontologies. It also presents discussion of the mappings used to 'uplift' tabular data (CSV files) into RDF using the Relational to Resource Description Framework Mapping Language (R2RML) along with a set of SPARQL queries for performing federated queries over the resulting data sets. By using federated queries which aggregate patient data (i.e. returning counts of patients rather than subject level data), combined with an uplift process into local RDF registries which is tailored towards anonymization, the approach ensures sensitive data is not exposed in queries. Finally we give our findings regarding the development of the ontology, discussion and future work, and finally, conclusions.

2. Background and related work

To address the two major challenges identified to enable queries over large cohorts of rare disease patient data, this section first introduces some of the background related to clinical registries from the perspective of data interoperability and examples of generic and domain specific ontologies in the biomedical domain. Next it explores efforts to enable federated queries over clinical data using semantic web technologies.

2.1. Clinical registries and data interoperability

Clinical registries are databases for collecting and storing information about the health of patients. They are important for monitoring disease and recording evidence regarding treatment and service delivery on health outcomes. They are generally developed for specific types of diseases, e.g. cancers, cardiovascular disorders, etc. and are increasingly used to support medical research, providing data for randomized clinical trials. The collection of real world data supports research through the generation of research hypotheses, facilitating descriptive studies and health service research [16].

[2] http://www.obofoundry.org/.

[3] https://www.w3.org/2013/data/.

[4] https://fairvasc.eu/.

[5] https://www.go-fair.org/fair-principles/.

Modern clinical registries can have sophisticated data dictionaries which describe the organization and logical structure of patient data, allowing structured data to be captured in a standardised manner. These tend to be developed for single registries in an ad hoc manner and so there is a broad lack of standardization across registries [4]. Therefore, to support, for example, data analytics across multiple registries, a process must be undertaken to harmonise and normalize the registry information at a semantic level. This is essential if one wishes to use and integrate data accurately from multiple registries in a meaningful way.

Two critical tasks towards this goal are first, to harmonise the definition of data elements and their corresponding values and second, to uplift/convert the data into a structure that adheres to this harmonised schema or enable queries that adhere to that schema [24]. To support standardised definitions of terminologies in clinical registries several terminologies and ontologies have been developed. Terminologies give definitions for terms in a specific domain, such as biomedical information. Ontologies share a similar goal as terminologies, i.e. to enable communication and knowledge sharing between human (or computer) agents, but differ in that they include a specification of terms and their meaning along with additional semantics such as relationships between concepts, defined using a formal (or semi-formal) language, such as the Web Ontology Language (OWL).[6] Ontology formality has two advantages over terminologies, firstly the use of reasoners which can infer knowledge and check the ontology consistency and second the use of ontology query languages (e.g. SPARQL) which allows data to be returned on relationships via predicates, inherent graph structure and triples [35]. An important concept in ontology engineering (see section 3) is reuse of existing ontologies. In the next section, several popular terminologies and ontologies relevant to clinical registries and AAV are explored.

2.2. Clinical terminologies and ontologies

A key resource within the biomedical informatics domain which brings together several important terminologies, classifications and coding standards to support interoperability between computer systems, is the Unified Medical Language Systems (UMLS) repository, developed by the United States National Library of Medicine. Examples of popular terminologies for clinical repositories are the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) and the International Classification of Diseases (ICD) [5]. Several other popular standard terminologies are also available on BioPortal,[7] such as the Medical Dictionary for Regulatory Activities Terminology (MedDRA), the Orphanet Rare Disease ontology (ORDO) and others. Here we describe some of the more popular terminologies and ontologies that are relevant to clinical registries which host clinical registries for rare diseases, such as AAV registries. These terminologies and ontologies are currently in use in some of the registries which have been analysed as part of the development of the FAIRVASC ontology, for example RKD and the European Vasculitis Society (EUVAS)[8] model registry.

SNOMED-CT is a comprehensive collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting. It has seen wide use in over 50 countries.[9] The SNOMED-CT ontology on BioPortal contains over 350 thousand concepts. These cover a wide range of classes, such as body structures, clinical findings, environment and geolocation, events, organisms, etc. There is no OWL version of SNOMED-CT available for download, as it is imported directly into BioPortal as RDF from the UMLS format [5] using Python scripts[10], so it is not an easy process to analyse the ontology. It is possible, though, with the appropriate UMLS license, to download the SNOMED Release Format 2 (RF2) file, and convert this using an open source toolkit.[11]

Several challenges have been identified when implementing SNOMED-CT. These are related to, for example, its accuracy, such as lack of coverage of terms, ambiguity of terms, syntactic consistency and issues related to missing relations and concepts combined with a complex hierarchy [18]. For successful implementation, Lee et al. [18] have highlighted the need for engagement with clinicians during the development phase, and the inclusion of terminology experts, analysts, clinicians and programmers to ensure both clinical and technical viewpoints are represented. This type of engagement when integrating services, can help inform the harmonisation process, as ultimately the data required is driven by the needs of these different stakeholders. Due to SNOMED-CT's large coverage of clinical terms, along with its popularity, it remains a good candidate standard for supporting interoperability. It should be noted though that use of SNOMED-CT is subject to the International Health Terminology Standards Development Organization (IHTSDO) Affiliate license provisions. This makes it free to use in IHTSDO Member territories, in low income countries, and for Qualifying Research Projects in any country.[12]

Another ontology of note is the Medical Dictionary for Regulatory Activities Terminology (MedDRA), developed in 1994 by the pharmaceutical industry is a standard for recording and reporting adverse reactions to drugs, and applies to all phases of drug development for humans, and to the health effects and malfunction of devices. The MedDRA ontology on BioPortal lists over 75,000 classes, including signs and symptoms, diseases, diagnoses and, medical procedures, etc. As with SNOMED-CT, terms in MedDRA may have multiple classifications. There is an issue therefore with the choice of a specific MedDRA classification for an adverse reaction as it may not be self evident. As with other medical classifications, subjectivity plays a role and different users may interpret the classifications differently. Another issue with MedDRA is that infections may be either coded by the type of microorganism responsible or the bodily location of infection. For example, in our exploration of MedDRA some ambiguity was found when classifying "Pneumocystis infection" which is a subclass of "Fungal infectious disorders", as this has no relationship to its location in the lungs. This becomes a problem then if one wishes to count the number of lung infections, for example, that occurred for patients, as a query would not pick up Pneumocystis. It is important to consider these limitations when using standards such as MedDRA to ensure that classifications and relationships that may be relevant, are not lost.

The Orphanet nomenclature (ORPHAcode) is a resource to improve the diagnosis, care and treatment of patients with rare diseases and was developed by the French National Institute for Health and Medical Research (INSERM) in 1997. It is organised in a multi-hierarchical classification system arranged in three hierarchical levels: Group of disorders, Disorder, and Subtype of a disorder. It is also available as the Orphanet Rare Disease ontology (ORDO) on BioPortal, which includes over 14,000 classes and can be downloaded as OWL. It integrates a nosology (classification of rare diseases), relationships (gene-disease relations, epidemiological data) and connections with other terminologies (SNOMED CT, UMLS, MedDRA), and classifications (ICD-11), as well as databases. The WHO's commitment to including Orphanet into ICD-11, demonstrates its applicability going forward as a source for data related to rare diseases. Together, these standards and ontologies have a large scope and cover a wide range of terms which are relevant to AAV and can therefore play an important role in supporting interoperability with other data sets.

[6] https://www.w3.org/OWL/.

[7] https://bioportal.bioontology.org/.

[8] https://vasculitis.org/.

[9] https://web.archive.org/web/20120210230029/http://www.ihtsdo.org/news/article/article/meaningful-use-quality-performance-measures-benefit-from-new-snomed-ct-public-good-use-policy.

[10] https://github.com/ncbo/umls2rd.

[11] https://github.com/IHTSDO/snomed-owl-toolkit.

[12] https://www.nlm.nih.gov/healthit/snomedct/snomed_licensing.html.

While previous ontologies have been created to address general health domains such as disease and diagnosis, body structures, adverse reactions, drugs, etc., there have been efforts to develop domain specific ontologies. An example of this is the domain specific ontology for adverse reactions (DSOAE), developed and applied specifically to chronic kidney disease (CKD) [33]. DSOAE builds on the Ontology of Adverse Events (OAE) by providing formal definitions of adverse reactions in relation to a specific disease, e.g. hypoglycemia as experienced by CKD patients. This domain specific data may not be available in existing more general ontologies without domain adaptation. DSOAE is designed to help support both data analysis, for example using ontology inference, and data integration, for example by providing clear semantics when converting free text or to integrate heterogeneous data sources with mappings and translations. DSOAE demonstrates the need to develop domain specific semantics, and how existing ontologies can be reused and extended. DSOAE though is developed without the input of domain experts, and has yet to be evaluated with real world data. Also, it does not explore any of the potential issues when working with patient data related to privacy and security of data, which is an important consideration for patient level data.

The Chronic Kidney Disease Ontology (CKDO) is an ontology primarily to describe clinical features associated with CKD [8]. It was developed from a study of the Royal College of General Practitioners (RCGP) Research and Surveillance Centre (RSC) database in the United Kingdom, and has primary care data from 109 General Practitioner (GP) practices across the United Kingdom, comprising a total population of around 1.8 million patients. It includes concepts specific to CKD organised according to the following categories: diagnosis, clinical examination finding, investigation, treatment, procedure, complication of care, and process of care. CKDO is also available on BioPortal and lists 280 classes. These classes were identified from an analysis of the UK National Health Service Read v2 hierarchy of terms and Read Code Browser Version 20.0, and reviewed by three consultant nephrologists, and two clinical informatics experts. The CKDO provides a detailed taxonomy of terms related to kidney disease, although the vast majority of terms (240) lack any definition and there are no relations defined between the different concepts. This has the potential to lead to ambiguity, which could result in wrong classification for those wishing to use this ontology. The Kidney Tissue Atlas Ontology (KTAO) is another example of a domain specific ontology [1]. KTAO has been developed in the context of the Kidney Precision Medicine Project (KPMP), which sets out to address common forms of kidney disease. They review several ontologies for modelling aspects related to the kidney, such as the Cell Ontology, UBERON, and Human Phenotype Ontology, concluding that gaps exist within these, as kidney specific terms can be inaccurately represented, synonyms may be missing, or taxonomic classification require reorganization. To address these gaps, KTAO sets out to logically represent the relations among gene markers, phenotypes, diseases, cell types, and anatomic entities to support modelling of common forms of kidney disease. By reusing terms from existing Open Biological and Biomedical Ontologies, KTOA deepens the granularity of terms describing kidney structure, function, and disease. The ontology is available on BioPortal and contains over 9900 classes. The main contribution of KTAO is that it supports modelling molecular data, to define novel subtypes of current (and currently insufficient) disease classifications. Due to the wide range of terms in the ontology though, it was found to be overly engineered for the specific needs of AVV registries. They also do not explore possible alignments with other well known standards in the biomedical domain, such as SNOMED-CT, as a potential candidate for capturing aspects relevant to kidney disease.

The FAIRVASC ontology is also domain specific for AAV and is developed with the help of seven domain experts based on seven AAV registries. It makes use of the previous mentioned SNOMED-CT and ORDO ontologies to provide additional semantics where required. As it is built with a strong alignment to existing registries, the FAIRVASC ontology is light weight compared to some of the other reviewed domain specific ontologies, e.g. KTAO, but by taking the ontological approach, future alignments with this ontologies are possible, if required.

2.3. Federated query approaches in biomedical domain

With the growth of big data, federated queries have been suggested as a solution to the problem of integrating and querying large distributed, heterogeneous data sources. Most solutions collect data about the main characteristics of the heterogeneous data sets, and then rely on a global ontology to describe different terms and the relationships between those terms [17]. Within the semantic web, most available triplestores now support federation at their endpoints and provide web access interfaces to enable this, e.g. Fuseki, GraphDB, Stardog. The federated query system then provides a unified access interface to the different data sources. Those data sources expose their endpoint, while keeping full autonomous control over how their data is represented internally. However, a federated query must be expressed using a global schema and then either a) the data is represented in a manner consistent with a global schema, or b) a mechanism must exist for transforming the query into a local query which will work on the internal data source. The latter approach requires query parsing, optimization, execution and result conciliation. Ultimately, a mapping must be created between the global and local query. The main benefit of query conversion is that it avoids redundancy of data, as the original data source is untouched. However, where sensitive data is concerned, it may be necessary to only expose a subset of the full data source and to pseudonymise, in which case the management of the mapping between the common/global query and the local query becomes increasingly complex. In the former approach, the data itself must be mapped onto a representation consistent with the global schema. When converting the underlying data source, it is also possible to expose only the relevant data and this method therefore provides control over what data is exposed at the endpoint and thus can provide security when it comes to sensitive data.

Within the biomedical domain, a number of research efforts have explored different federated querying approaches. Weber [31] explores the use of federated queries for integrating data at a national level over different clinical research data repositories. They developed a framework for addressing challenges related to; functional and temporal equivalence (can registries process the same queries, do they cover same data ranges), data release synchronicity (how to deal with registries updating data at different times), ontological equivalence (is there a shared vocabulary), semantic discernibility (e.g. is a complication date the date it actually occurred, or merely the date it was recorded), system availability (the time a registry is available), population overlap (are patients recorded twice in different registries) and data access restriction (are there specific requirements of a person wishing to query the data). Using this framework they evaluate data from Harvard's four-site Shared Health Research Information Network (SHRINE). One of their key findings is that variability among sites in a small network must be addressed by conforming to a common set of features, date ranges, ontologies, and data access restrictions. However, this variability can be an asset in a large network as it is more likely that researchers find a subset of data in the registries that is most appropriate for their queries. Federated query approaches should therefore take this into account and provide a level of flexibility and extensibility in their methodology so that networks of federated registries can grow over time and perhaps even provide levels of semantics, so that specific queries can be satisfied for subsets of the total number of registries.

González-Beltrán et al. [13] applied an ontology-based solution for querying distributed databases storing data about cancer, where the ontological description drives the creation of the federated query. They make use of a federated Local-As-View approach to data integration by defining mappings from distributed data sources to a global-schema, which is realised by the National Cancer Institute (NCI) Thesaurus ontology describing the cancer domain. The ontology is used to provide unambiguous meaning to the data sources. However, it is not currently used to provide a unified view for querying the data sources. Their approach specifically looks at the conversion of high level domain-based queries, to enable localised querying of a single or multiple resources. They have successfully built and evaluated the performance of a web based query building tool using their approach. One potential weakness of the approach is that there does not appear to be any domain level expertise to validate the query mappings, which could potentially result in wrong mappings where, as often is the case, ambiguous terms are used with an existing taxonomy. Also, as this approach does not set out to address patient level data, it also does not explore any issues related to GDPR, in particular, data privacy and security of data.

Sima et al. [27] propose an ontology-driven linked data integration architecture using SPARQL and leveraging virtual links. A virtual link is an intersection data point between two data stores. The links are required in order to enable federated queries, given that they act as join points between the federated sources. This was tested over three bioinformatics databases containing: evolutionary relationships among genes across species (OMA), curated gene expression data (Bgee) and biological knowledge on proteins (UniProt). The UniProt data set was already available in RDF, but a conversion process took place for both OMA and Bgee, resulting in only a subset of data being converted to RDF. This process also required the development of the GenEx ontology[13] for mapping Bgee genetics data. Ontop [6] was used to implement Ontology Based Data Access (OBDA) over the Bgee relational database (see section 3.5 for more discussion on ODBA). The OMA data was converted using a hybrid approach that combines materialization and a possible RDF graph virtualization for the sake of semantic enrichment and knowledge extraction.[14] The paper provides an evaluation of the performance of a set of 12 typical federated queries and demonstrates how conversion of data into RDF can support federated queries over three heterogeneous data sets. Virtual links also provide a means for supporting query development when dealing with data sets which represent multiple domains of biological knowledge. Performance of queries was within seconds even when dealing with large data sets (UniProt RDF release comprises >55 billion triples). These are open data sets though, and as such the paper does not set out to address issues related to sensitive data, such as patient level data.

In [34], Yu and Weber explore the issues related to federated queries of large stores of patient electronic health records (EHRs) with a focus on balancing accuracy and privacy. They propose a method using HyperLogLog (HLL) probabilistic sketching algorithm which uses a hash function that maps patients to a random number between 0 and 1, and then estimates the number of patients who match a query by keeping track of just the minimum hash value. This approach avoids some of the potential attacks related to hashed patient identifiers, e.g. dictionary attacks and multi-party computation (MPC) and homomorphic encryption techniques, which can be computationally complex. Also, approaches which share data back and forth between federated sites can be unreliable, due to the unreliable nature of federated queries. Yu and Weber are careful though not to endorse a one size fits all approach, as different applications and institutions will have different needs, but by providing analysis of the different presented approaches, they provide a means for assessing the approach best suited for a particular use case in terms of performance.

Collectively, we can see here a range of different approaches to supporting federated queries, and these range from mapping data to mapping queries, or hybrid approaches, as well as research efforts into considerations when dealing with sensitive data (Table 1). In the rest of this paper we explore the approach taken to data integration and federated query support for FAIRVASC, and the reasoning behind the decisions made. The paper sets out to provide a clear methodology for generating a global data schema (ontology), mappings to covert local data into data that adheres to this schema, and development of federated querying which enables data integration while ensuring a sufficient level of privacy and security is maintained for sensitive data. AAV registries have several requirements which make using the data uplift approach, which maps the data into a representation which adheres to a global schema, more beneficial. These will be explained more in the next sections, but is related to the scope of the data in the registries (i.e. no requirement for integration of data from other domains for example), and the sensitivity of the data, meaning that exposure of the full database is not an option when exposing patient data for federated querying.

3. FAIRVASC: an ontology for ANCA associated vasculitis

3.1. *Methodology for ontology development and semantic uplift*

Several methodologies have been developed to support ontology development. These include the older Methontology [12] and On To Knowledge [30] developed during the late 90s and early 00s, to the more recent NeON (an update of Methontology) [29] and Simplified Agile Methodology for Ontology Development (SAMOD) [22]. NeON takes a very broad look at ontology development, satisfying several different approaches within the one methodology. SAMOD takes a more refined approach, inspired by the Test-Driven Development process in software engineering.

FAIRVASC focuses specifically on AAV and does not (yet) require interlinking with other medical data sets and, as it is developed under the direction of domain experts across seven European registries, our methodology borrows from SAMOD guidelines. These highlight the need to keep the number of entities low, with a focus on the most relevant concepts first, the avoidance of unnecessary semantic structure, and the provision of self-explanatory entities. Combined with these guidelines, we apply the existing standard methodology for ontology development, developed by Stanford University.[15] This involves iterative cycles, each consisting of defining the scope (a set of key competency questions, see section 3.2), exploration of existing ontologies and vocabularies for reuse, enumeration of terms, definition of classes, properties and constraints, and finally the creation of instances (see Ref. [20] for a graphical representation of these steps).

As the ontology covers several registries, it was important to include experts from each registry in the process of ontology development, as recommended when integrating with SNOMED-CT (Lee et al. [18]). Three teams were created to support the ontology development (see Fig. 1). The first team of experts (at least one representative from each registry) in vasculitis formed the Query Implementation Team (QIT), tasked with developing competency questions and the human readable form of queries. These are later turned into SPARQL queries for querying the data (number 2 in Fig. 1). Two additional teams were created to support data analysis and harmonisation called the Harmonisation Implementation Team (HIT), and implementation of the mappings and SPARQL queries, called the FAIRVASC Implementation Team (FIT).

The FIT has one member from each registry with IT background who were tasked with implementing the ontology (classifications, and

[13] https://biosoda.github.io/genex/.
[14] https://qfo.github.io/OrthologyOntology.
[15] https://protege.stanford.edu/publications/ontology_development/ontology101.pdf.

Table 1
Federated querying in the biomedical domain.

	Central Ontology/ Ontologies	Query Mapping	Data Mapping	Use of domain experts	Implemen-ted Query Tool	Address Data Sensitivity
González-Beltrán et al. (2012)	x	x			x	
Sima et al. (2019)	x	x	x		x	
Yu and Weber (2019)						x

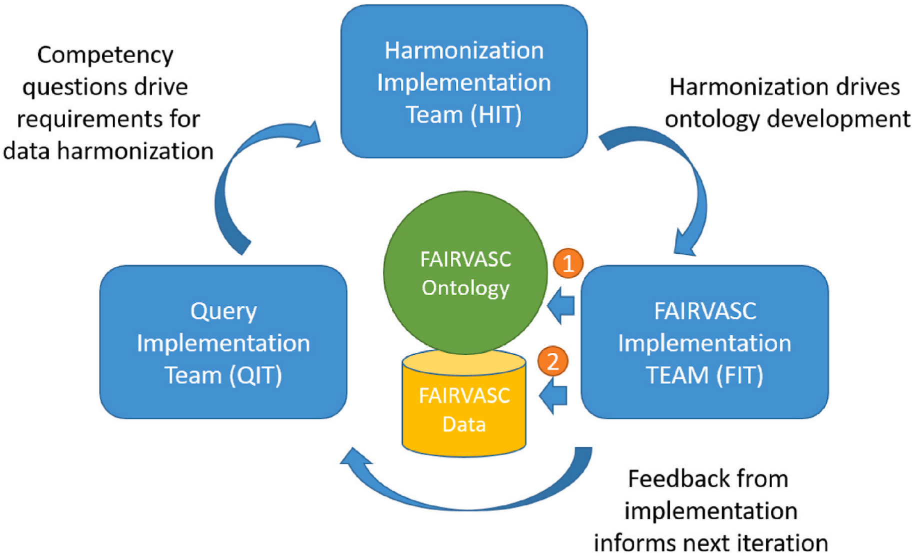

Competency questions drive requirements for data harmonization
Harmonization Implementation Team (HIT)
Harmonization drives ontology development
FAIRVASC Ontology
1
2
Query Implementation Team (QIT)
FAIRVASC Implementation TEAM (FIT)
FAIRVASC Data
Feedback from implementation informs next iteration

Fig. 1. Overview of interaction between FAIRVASC implementation teams.

relations), the mappings to generate RDF and the installation of the registries' local triplestore (section 3.5 and 3.6). In the following sections the different steps of the methodology, described here [20] and above as applied to the FAIRVASC ontology development are discussed in more detail.

3.2. *Determine the scope - development of high level competency questions*

Competency questions are natural language questions outlining and constraining the scope of knowledge to be represented in an ontology. These questions were formalized as specific SPARQL queries (see section 3.6). The questions presented here cover the first iteration of the FAIRVASC ontology development (version 1) and should not be considered an exhaustive list of questions required of the ontology. The QIT team identified two key areas of importance for AAV, one related to exposure and the second related to outcomes. The following variables representing exposures of key importance to the FAIRVASC ontology were identified: age, gender, ANCA auto-antibody subtype,[16] AAV diagnosis,[17] affected organ pattern,[18] induction treatment,[19] maintenance treatment[20] and serum creatinine level at presentation[21] Variables related to important clinical outcomes: Death, End-stage kidney disease, serious complication such as severe infection, cancer and cardiovascular disease. Questions took the form of "What number of patients died, stratified by age, gender, main diagnosis, and ANCA specificity?". Once these questions were defined, the next step was the analysis of the data registries to identify what data within the registries can be used to satisfy them.

Table 2
Sample of the Irish RKD data dictionary.

Variable Name	Form Name	Field Type	Field Label	Choices
patient_id	patient	text	Patient Id	
gender	baseline _characteristics _common	dropdown	Gender	1, Male — 2, Female — 3, Undetermined

3.3. *Data analysis*

The data analysis phase consists of two parts, the analysis of the registries and then a harmonisation process across the registries.

[16] AAV is associated with a type of auto-antibody that can be identified using techniques such as enzyme-linked immunosorbent assay (ELISA) and immunofluorescence (IF). These ANCA's can be of different class depending on the type of antigen target they bind to, such as myeloperoxidase (MPO) and proteinase 3 (PR3).

[17] AAV is an umbrella term for the diagnoses; Granulomatosis with Polyangiitis (GPA), Microscopic Polyangiitis (MPA), Eosinophilic Granulomatosis with Polyangiitis (EGPA) and Unclassified ANCA associated vasculitis.

[18] Many organs can be impacted by AAV.

[19] Induction treatment is a treatment to initiate disease remission, i.e. an absence of signs or symptoms of active disease.

[20] Maintenance treatment is treatment to maintain disease remission.

[21] Serum creatinine is a blood test and clinically used as a marker of kidney function. Presentation creatinine level is a patients serum creatinine level at AAV diagnosis.

3.3.1. Analysis of the data registries to inform ontology development

The seven data registries (Ireland, UK, France, Germany, Sweden, Czech Republic, and Poland) provided both a data dictionary and a set of sample data, i.e. simulated data generated for testing purposes. The data dictionaries are spreadsheets or CSV files which list semantics about different terms in the registry. These are generated depending on the particular implementation of the registry at the pilot. For example, in the case of the Irish RKD registry, the data dictionary was generated by a REDCap database.[22]

REDCap is a commonly used solution which provides a database server (built on MySQL). The "schema" file can be exported as CSV. The Irish registry, which uses REDCap, focuses on the structure of forms for data entry, i.e. forms related to patient data, encounters, baseline characteristics, etc. and so the generated schema reflects this. Table 2 shows a snippet of the schema, with two terms "patient_id" and "gender". The columns given here are the name of the term "Variable Name", the related form "Form Name", patient and baseline characteristics common, the "Field Type", text or dropdown, the field label (in the form) and, in the case of a dropdown, the list of possible values. The other data registries also generated data dictionaries, which varied in the types of column names reflecting their own internal database structure. For example, in the case of the German registry (GeVas) the dictionary had several separate CSV files each for different areas such as demographics, disease diagnosis, etc. whereas the output for RKD was a single CSV for all data.

The first step of analysis was therefore to become familiar with the terms available, and how they align with the competency questions. The sample (simulated) patient data, for testing purposes, was also provided as CSV files, where the column names correspond to terms in the data dictionaries.

3.3.2. Data harmonisation: enumerating terms, class and property definitions

Data dictionaries did not always align perfectly with the simulated data provided, so a good understanding of the column names in the simulated data, and the types of values that were being stored in corresponding cells, was required. From the initial analysis a set of harmonised terms were identified by the core HIT team. Several fundamental terms were easy to harmonise, such as year of birth, gender, and main diagnosis. Patient ID required an agreement between the registries on creation of a dedicated FAIRVASC patient ID, providing some harmonisation in the representation of Unified Resource Identifiers (URIs). These were then presented to the full HIT team who signed off on these terms. For more complex areas of the registries, e.g. disease complications, specific sessions were devoted to discussing the harmonisation of terms across registries. Table 3 shows the outcome of these discussions for a subset of values related to complications.

The table gives several columns: the first column "Generic Normalised Term" is the term which covers the largest number of registries, e.g. the SNOMED-CT code for "Respiratory tract infection 275498002" can be satisfied by 5 registries (second column value) and subsumes "Pneumonia", "Respiratory system and lungs", "Lower Respiratory Infection", "Upper respiratory tract infection", "pneumocystis jirovecii infection" etc. across the registries.[23] If any of these are given as an infectious complication, the value will generate a corresponding RDF triple for "Respiratory tract infection". The table indicates that a row is part of a generic term by the two colour coding. The green cell colour indicates that a particular values is present in a registry.

The third column "Specific Normalised Term" gives a more specific mapping, which can be satisfied by registries. Again, these were mapped to SNOMED-CT codes. We also provide a mapping for all registries which indicates simply whether an infectious complication occurs, as some registries provide no granularity on the nature of infection. This flexible approach then allow for queries which can cover a range of registries, with differing levels of granularity, mapped to SNOMED-CT. This process was applied to a range of different types of complications, including also malignancies, thromboembolic and cardiovascular disorders. Mappings were also made for recording blood test results, organ pattern, diagnosis, death, end stage kidney disease and Birmingham Vasculitis Activity Score (BVAS).[24] The next section introduces the ontology developed as a result of this process.

3.4. The FAIRVASC ontology

After the process of harmonisation was complete, the outcomes were documented in a specification document (available here[25]) and integrated into the FAIRVASC OWL ontology. The ontology was developed using Protégé,[26] a tool for ontology development; see Fig. 2 for a snapshot of the FAIRVASC class structure in Protégé. Each class (those in bold are FAIRVASC classes, the others belong to the BVAS ontology), its relationships and data properties were defined using this tool. The ontology has 9 top level classes (ignoring BVAS), 13 overall, 9 object properties, and 24 data properties. Top level classes are; Patient, Patient Overview, Diagnosis, Clinical Outcomes, Encounter, Clinical Test and Organ Pattern. A definition of each of these and how they are related to each can be seen in Table 4.

The key subject in FAIRVASC for all data registry records is the **Patient**. Patient level data has been limited to their gender and year of birth, to avoid data which could potentially be used to identify them. Each patient has a relation to their **Patient Overview**, their **Encounters** and **Complications**. The **Patient Overview** gives details about the patient's disease characteristics over the course of the disease, but has no temporal aspect, i.e. no timestamp recording when these details were taken, whereas **Encounters** and **Complications** can include values taken over the course of the patient's disease and therefore include a timestamp. Therefore a patient may have multiple encounters and multiple recorded complications, whereas they would have a single set of values for organ patterns classes as part of **Patient overview**, which have accumulated over the course of their disease. Table 5 gives an example of data properties for the class **Diagnosis**, which is related to a **Patient** via **Patient Overview** using the object property *hasDiagnosis*. It should be noted that wherever a data type is a URI, this indicates that there has been a direct mapping with another available ontology (e.g. ORDO or SNOMEDT-CT). These are also available along with the object properties, etc. in the online documentation for the ontology.

3.5. Semantic uplift (generating RDF instances)

Once an ontology was created, RDF instances were generated from the data in each registry, published, and then made available for querying. This required the conversion of the non-RDF tabular data into Linked Data based upon semantic-web technologies. This is called semantic uplift [10]. The mapping of data can take one of two approaches. The first is a direct mapping approach which reflects the structure of the tabular data, is implemented by an algorithm and requires no mapping

[22] https://www.project-redcap.org/.

[23] It should be noted, a translation also took place where cell values were given in a foreign language, e.g. Polish.

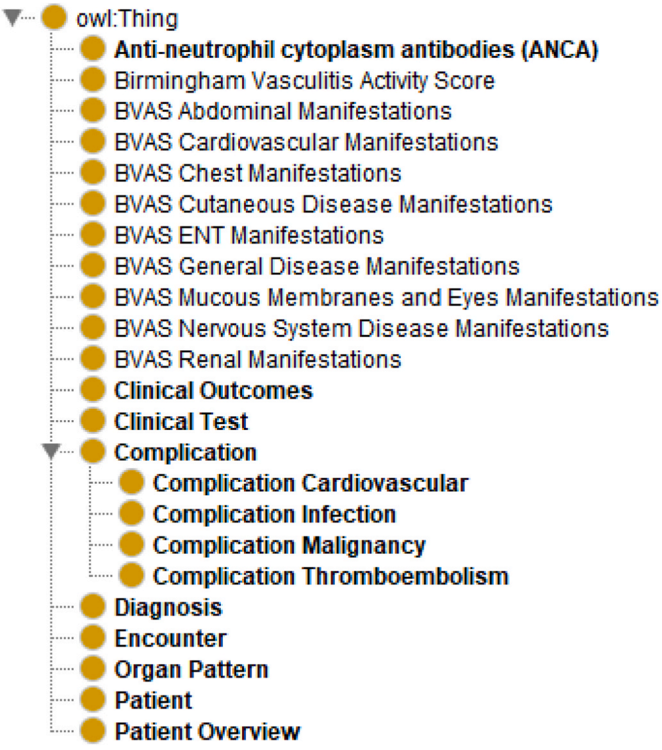
[24] The FAIRVASC ontology imports the Birmingham Vasculitis Activity Score (BVAS) classes from a related ontology developed in parallel with FAIRVASC. This is not covered in this paper but was developed to standardise the representation of BVAS scoring across the registries as RDF.

[25] https://fairvasc.eu/2021/09/03/fairvasc-registry-legal-metadata-profiles-have-been-captured/.

[26] https://protege.stanford.edu/.

Table 3
Normalisation of terms across registries for complications which are infections. Green cells are those which are present in a registry.

Generic Normalised Term	No. of registries	Specific Normalised Term	No. of registries	RKD	GeVas	Czech	Skane	PolVas	UKiVas	GVEF
Respiratory tract infection 275498002	5	Respiratory tract infection 275498002	1			Respiratory system and lungs				
Respiratory tract infection 275498002		Pneumonia 233604007	3	Pneumonia (233604007)	Pneumonia		Pneumonia	NA		NA
Respiratory tract infection 275498002		Lower respiratory tract infection 50417007	1						AE LowerRespInfect	
Respiratory tract infection 275498002		Upper respiratory infection 54150009	2	Other	Upper respiratory tract infection	Other	Other >type of inf other	NA	AE UpperRespInfect	NA
Respiratory tract infection 275498002		Pneumocystosis pneumonia 415125002	1	Other	NA	Other	Pneumocystis jirovecii infection	NA	AE Infection	NA
Urinary tract infectious disease 68566005	5	Urinary tract infectious disease 68566005	2	Other	Urinary tract infection	Kidney and urinary tract	Other >type of inf other	NA	AE Infection	NA
Urinary tract infectious disease 68566005		Pyelonephritis 45816000	3	Pyelonephritis (4581600)	Pyelonephritis		Pyelonephritis	NA	AE Infection	NA



- owl:Thing
  - Anti-neutrophil cytoplasm antibodies (ANCA)
  - Birmingham Vasculitis Activity Score
  - BVAS Abdominal Manifestations
  - BVAS Cardiovascular Manifestations
  - BVAS Chest Manifestations
  - BVAS Cutaneous Disease Manifestations
  - BVAS ENT Manifestations
  - BVAS General Disease Manifestations
  - BVAS Mucous Membranes and Eyes Manifestations
  - BVAS Nervous System Disease Manifestations
  - BVAS Renal Manifestations
  - Clinical Outcomes
  - Clinical Test
  - Complication
    - Complication Cardiovascular
    - Complication Infection
    - Complication Malignancy
    - Complication Thromboembolism
  - Diagnosis
  - Encounter
  - Organ Pattern
  - Patient
  - Patient Overview

Fig. 2. Overview of FAIRVASC ontology classes taken from Protégé.

file to be created, but is then only meaningful to those who understand what is contained in the initial tabular data schema. The second is a declarative mapping approach allowing one to relate structure of the tabular data to vocabularies.

Declarative mappings using, for example, the World Wide Web Consortium (W3C) standard for declarative mappings, called R2RML,[27] allow one to declare how data from the non-RDF resources should be transformed into RDF, while relying on the underlying relational database technology (usually SQL) to manipulate the data. R2RML is a language for defining mappings between tabular data and RDF graphs. Mappings can include target vocabularies, so that, for example, columns in a table can be assigned specific definitions given by existing vocabularies on the web. This is a powerful tool to enable the use of common vocabularies to describe data sets, and for bringing semantics to tabular data through conversion to RDF.

When working with a database, it is possible to implement declarative mappings in two ways. The first is to use Ontology Based Data Access (OBDA) [28] and the second is to make use of a processor to convert the data into RDF and upload it into an RDF triplestore [10]. The former approach supports queries over the conceptual ontology and then, based on the declarative mappings, converts these into an appropriate query for the underlying database. Open source tools such as Ontop [6] are available for OBDA. This approach though does not work with tabular data in the form of CSV files which were required for FAIRVAS registries. Therefore, the first approach was taken and an R2RML processor was employed to convert the tabular CSV data into RDF.

The R2RML mappings were developed by the FAIRVASC Implementation Team (FIT) in an incremental fashion following the harmonisation efforts of the HIT. As each new aspect of the registries were added to the ontology, a mapping was developed and that mapping was then tested over simulated data. The mappings are not publicly available, but a snippet of a mapping for the Irish RKD registry can be seen in Listing 1.1. The mapping shows the required prefixes (i.e. name spaces of the different vocabularies used in the mapping process), it gives some pre-processing of the data (using SQL queries) and then the mapping of the selected values from the CSV data to subject and predicates.

[27] https://www.w3.org/TR/r2rml/.

Listing 1.1: R2RML for converting patient data in RKD

```
@prefix rr: <http://www.w3.org/ns/r2rml#> .
@prefix fvc: <http://ontologies.adaptcentre.ie/fairvasc#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<#patient>
        rr:logicalTable [ rr:sqlQuery """SELECT
                RECORD_ID, YEAR_OF_BIRTH, GENDER,
                FORMATDATETIME(PARSEDATETIME(DATE_OF_VISIT,'dd/MM/yyyy'),
                'yyyy-MM-dd')||'T00:00:00' AS ENCOUNTER_DATE,
                CASE WHEN GENDER = 1 THEN 'Male'
                        WHEN GENDER = 2 THEN 'Female'
                        WHEN GENDER = 3 THEN 'Undetermined'
                END AS GENDER_
        FROM RKD""" ];

        rr:subjectMap [
                rr:template "http://data.fairvasc.ie/
                resource/rkd/patient/{RECORD_ID}" ;
                rr:class fvc:Patient;
        ] ;

        rr:predicateObjectMap [
                rr:predicate fvc:patientID ;
                rr:objectMap [
                        rr:column "RECORD_ID" ;
                        rr:datatype xsd:string;] ;
        ] ;

        rr:predicateObjectMap [
                rr:predicate fvc:yearOfBirth ;
                rr:objectMap [
                        rr:column "YEAR_OF_BIRTH" ;
                        rr:datatype xsd:gYear;] ;
        ] ;

        rr:predicateObjectMap [
                rr:predicate fvc:gender ;
                rr:objectMap [
                        rr:column "GENDER_" ;
                        rr:datatype xsd:string;] ;
        ] ;

        rr:predicateObjectMap [
                rr:predicate fvc:hasPatientOverview;
                rr:objectMap [
                rr:parentTriplesMap <#patient_overview> ;
                rr:joinCondition [
                        rr:child "RECORD_ID" ;
                        rr:parent "RECORD_ID" ;] ;
                ] ;
        ];
```

3.6. Publishing the RDF data and enabling federated queries over a prototype interface

Once the RDF data is generated, it must next be uploaded into a triplestore which supports SPARQL federated queries. There are several triplestores available such as Apache Jena Fuseki, Oracle Spatial and Graph, Strabon, Stardog, GraphDB and Parliament.[28] There are various pros and cons with these different implementations, from performance, to different features supported, e.g. parallel queries, to geospatial support [7,9,19]. As the amount of patient data for AAV is not excessive (a total of 6000 patients across the registries) Apache Jena Fuseki, which is a free and open source Java framework for building Semantic Web and Linked Data, was found to be more than adequate to support queries over the current tranche of patient data based on the first FAIRVASC ontology iteration. It is also easy to install and provides a web client for testing queries.

A federated query approach allows data to sit on each registry's local network, which exposes an endpoint of a locally run triplestore. This means no patient data is shared without the explicit authority given from the registry. The patient data uplifted to the local triplestore is limited in scope to avoid patient identification. Nonetheless additional security was put in place in the form of basic authentication using shiro[29] which comes with, and can be configured on, each Fuseki triplestore. This allows queries on the endpoint only when a username and password are provided.

The interface itself was implemented using Django, NGINX, HTML5, JavaScript and Bootstrap[30], and used Django's user authentication system, which handles user accounts, groups, permissions and cookie-based user sessions, see Fig. 3.

The interface currently supports basic federated querying of simulated data on the registries, which includes queries related to counts of patients stratified by age, gender, diagnosis, ANCA Specificity, death, end stage kidney disease, complications, etc. which were all identified early on in the set of experimental questions presented in section 3.2. A federated query is sent from a central server to each registry triplestore. Results are aggregated, so only counts are returned. It should be noted that results below the count of 5 are ignored, so as to ensure anonymity of patient data. The interface also supports querying of provenance data related to uplift and queries made (see next section). Results are returned in tabular form, and can be downloaded as CSV. Fig. 4 shows the interface, Fig. 5 shows a sample of data returned, and Listing 1.2 gives the SPARQL query to return these data. These results are taken from the Irish registry.

3.6.1. R2RML and SPARQL updates to record provenance data

Due to the sensitive nature of the data within the registries, and also to support analysis of changes to the data over time, an auditing service has been developed to record information about the uplift process as well as on the querying of data. This service generates RDF using the Provenance Ontology (PROV-O)[31] ontology. The service enables a secure audit trail of research activities to be maintained to both audit the

Listing 1.2: Federated SPARQL Query for returning patients stratified by gender, diagnosis, ANCA specificity and death

```
SELECT (SUM(?patientCount) as ?totalPatientCount) ?registry ?gender
?mainDiagnosis ?ancaSpecificity ?hasDied
WHERE
{
 {
 }
  UNION
  {
  SERVICE <triplestore_1>
  {
   SELECT (COUNT(DISTINCT ?open_fairvasc_tcd_query_1_end) as ?patientCount)
   ?registry ?gender ?mainDiagnosis ?ancaSpecificity ?hasDied
    {
    ?open_fairvasc_tcd_query_1_end a fvc:Patient.
    ?open_fairvasc_tcd_query_1_end fvc:hasPatientOverview ?patientOverview.
    ?open_fairvasc_tcd_query_1_end fvc:hasClinicalOutcomes ?o.
    ?open_fairvasc_tcd_query_1_end fvc:gender ?gender.
    ?patientOverview fvc:hasDiagnosis ?dia.
    ?dia fvc:mainDiagnosis ?mainDiagnosis.
    ?patientOverview fvc:hasANCA ?anca.
    ?anca fvc:ancaSpec ?ancaSpecificity.
    ?o fvc:death ?hasDied.BIND('RKD' AS ?registry)
   } group by ?registry ?gender ?mainDiagnosis ?ancaSpecificity ?hasDied
   having (COUNT(DISTINCT ?open_fairvasc_tcd_query_1_end) > 4)
   }
  }
 } group by ?registry ?gender ?mainDiagnosis ?ancaSpecificity ?hasDied
```

[28] https://jena.apache.org/documentation/fuseki2/, www.strabon.di.uoa.gr, www.stardog.com, www.ontotext.com/products/graphdb, parliament.semwebcentral.org.

[29] https://shiro.apache.org/.

[30] https://www.djangoproject.com/, https://www.nginx.com/, https://dev.w3.org/html5/spec-LC/, https://www.javascript.com/, https://getbootstrap.com/.

[31] https://www.w3.org/TR/prov-o/.

Table 4
Top level classes in FAIRVASC ontology.

Class	Definition	Object Properties	Data properties
Patient	This class covers the concept of a Patient in the data registries. Each patient will have data properties gender, data of birth and also have object properties for ANCA subtype, ANCA-associated vasculitis (AAV) diagnosis, and presentation creatinine.	hasPatientOverview, hasEncounter, hasClinicalOutcomes	patientID, yearOfBirth, gender
Patient Overview	This class represents an overview of a Patient's key AAV disease characteristics. This consists of the patient's current vasculitis diagnosis, ANCA subtype and which organs have been affected by vasculitis over the course of their disease.	hasDiagnosis, hasANCA, hasOrganPattern	
Encounter	This class covers "encounters" the patient has had with medical professionals which have been documented in the local vasculitis registry. Types of encounters include; BVAS Version 3 (see BVAS class below), Tests (e.g. blood tests such as serial serum creatinine levels), and a record of complications over the course of their disease etc.	hasBVAS, hasClinicalTest, hasComplication	dateOfEncounter
Clinical Outcomes	Clinical outcomes represent the class of measurable changes in health, function or quality of life that result from AAV or its treatment. Important outcomes include end-stage kidney disease (ESKD) and death (and cause of death). The last recorded contact of the patient with the registry is an important associated variable to collect to inform length of follow-up and survival analysis.	hasComplication	ESKD, dateOfESKD, death, dateOfDeath, causeOfDeath, lastRecordedContact

uplift process and also the query process.

The addition of provenance data recording uplift depends on additional R2RML mapping included with the R2RML for registry uplift, generating RDF provenance data according to the schema presented in Table 6. This makes use of three main classes within PROV-O: Activity, Entity and Agent. Provenance data is added to the triplestore each time an uplift is run over the registry data. This records the time the prov:Activity ProvenanceActivityUplift takes place (prov:startedAtTime and prov:endedAtTime), it records who undertook the activity (prov:Associated With prov:Entity), what the activity used (prov:used prov:Entity) i.e. a CSV file and the R2RML file, and finally what the activity generated (prov:generated prov:Entity), i.e. the resulting graph. For each prov:Entity a title of the Entity is recorded (dcterms:title), for example "RKD CSV File", "RKD R2RML File" or "RKD RDF graph". For the CSV and R2RML provenance entities, there is also a responsible person attributed to the file (prov:wasAttributedTo prov:Agent), i.e. the person who created the file. Each person is recorded in the graph with two details, their foaf:name and foaf:mbox, e.g. "Guy Incognito" and "guy.incognito@underhill.com". In the case of the graph, the inverse property to prov:generated, prov:wasGeneratedBy prov:Activity is used to associate it to the ProvenanceActivityUplift activity. Listing 1.3 gives a SPARQL query for querying the provenance data on uplift.

Table 5
Properties of class Diagnosis in FAIRVASC ontology.

Data Property	Definition	Data type
dateOfDiagnosis	The date a diagnosis took place	xsd:dateTime
mainDiagnosis	A diagnosis is one of four possible values, and these are currently mapped directly to Orpha codes; ANCA vasculitis unclassified - ORPHA:156 152 — Eosinophilic granulomatosis with polyangiitis (Churg Strauss) - ORPHA:183 — Granulomatosis with polyangiitis (Wegener) - Orpha:900 — Microscopic polyangiitis (including renal limited vasculitis) - ORPHA:727.	xsd:anyURI

Listing 1.3: SPARQL Query for data on uplift activity

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT ?activity ?agentResponsible ?agentName ?usedEntity ?usedTitle
?generatedEntity ?generatedTitle
WHERE
{
    ?activity a fvc:ProvenanceActivityUplift;
            prov:used ?usedEntity;
            prov:generated ?generatedEntity;
                          prov:wasAssociatedWith ?agentResponsible.
          ?usedEntity dcterms:title ?usedTitle.
          ?generatedEntity dcterms:title ?generatedTitle.
          ?agentResponsible foaf:name ?agentName.
}
```

The recording of queries is developed in the context of the requirement to keep the data in the triplestore secure. Each triplestore has an endpoint exposed to enable queries (with some authentication, i.e.

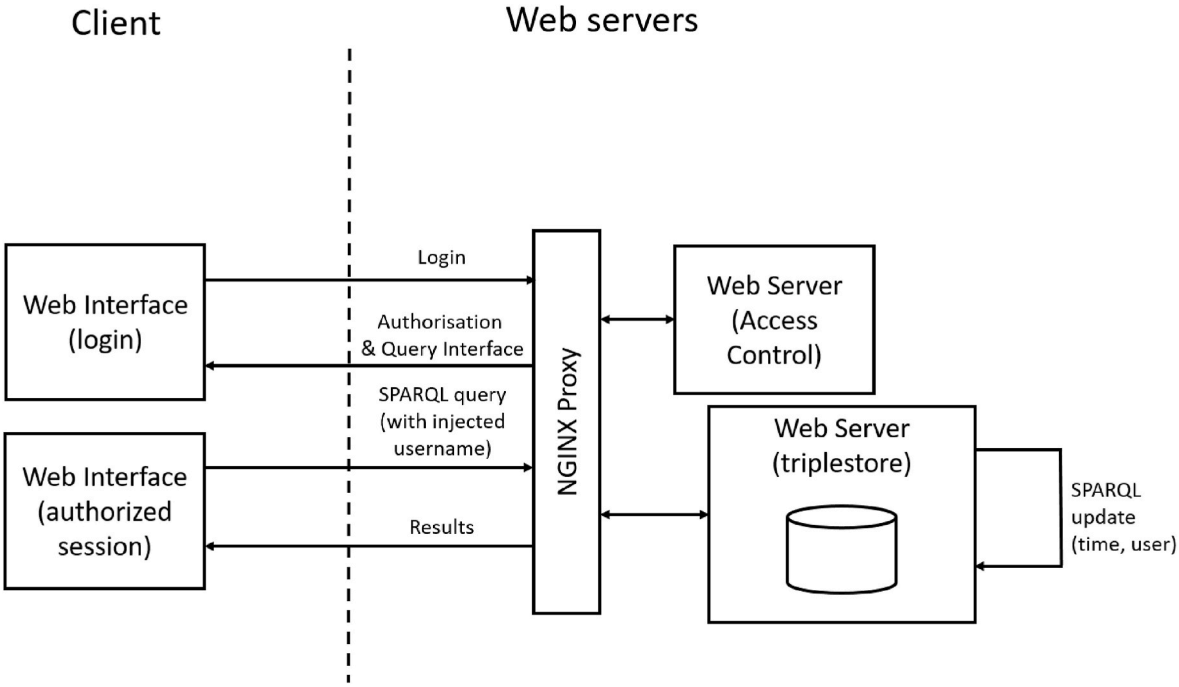
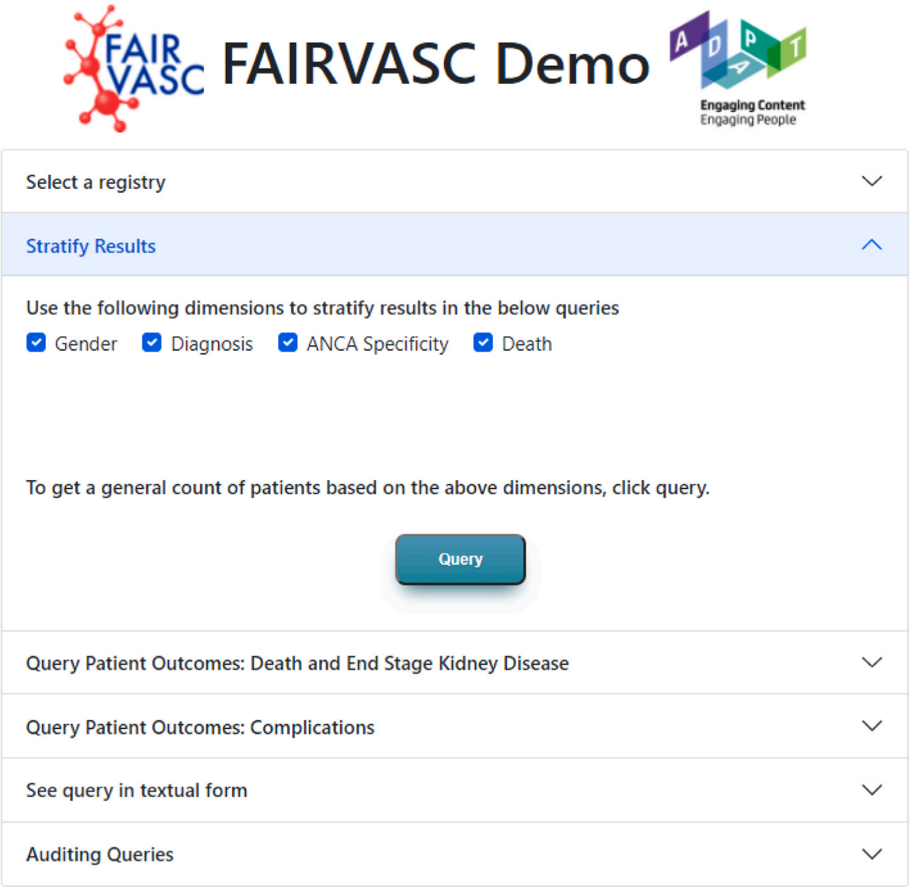

Client
Web servers
Web Interface (login)
Web Interface (authorized session)
Login
Authorisation & Query Interface
SPARQL query (with injected username)
Results
NGINX Proxy
Web Server (Access Control)
Web Server (triplestore)
SPARQL update (time, user)

Fig. 3. Query interface architecture.



FAIRVASC Demo

Select a registry

Stratify Results

Use the following dimensions to stratify results in the below queries

☑ Gender ☑ Diagnosis ☑ ANCA Specificity ☑ Death

To get a general count of patients based on the above dimensions, click query.

Query

Query Patient Outcomes: Death and End Stage Kidney Disease

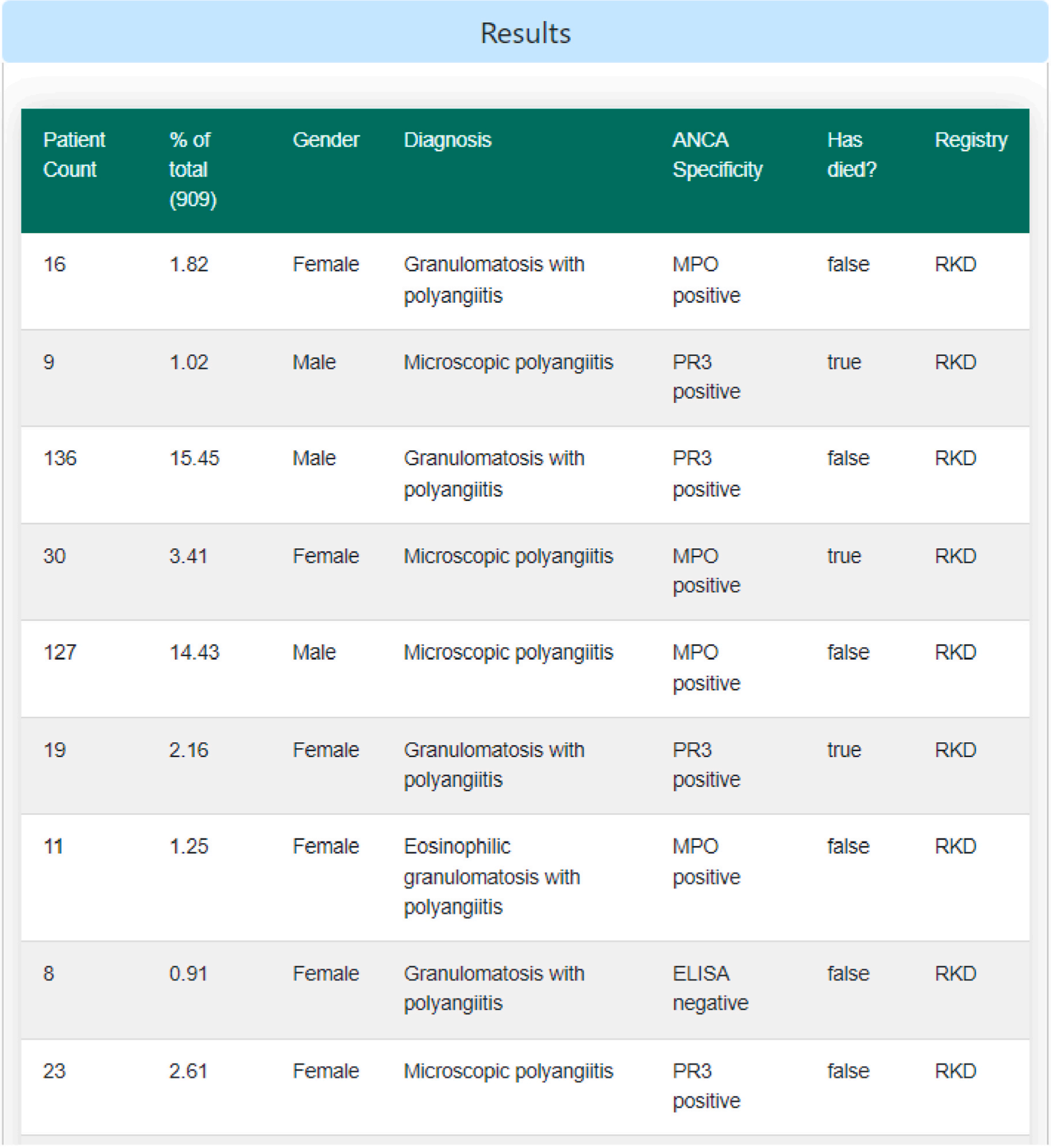
Query Patient Outcomes: Complications

See query in textual form

Auditing Queries

Fig. 4. The FAIRVASC query interface to select patients and stratify by gender, diagnosis, ANCA specificity and death.

username and password), but no updates to the graph are supported by this endpoint. Updates to the graph can only be run locally. To record when and who makes a query, the server log (generated by each triplestore) is used to generate a SPARQL update. So, in practice, when a query is made to the triplestore it is logged in the local log file, the log is read and a SPARQL update is generated (using some Java code) which adds provenance data to the graph. It should be noted that due to the large number of combinations of dimensions possible for queries over



Results

Patient Count	% of total (909)	Gender	Diagnosis	ANCA Specificity	Has died?	Registry
16	1.82	Female	Granulomatosis with polyangiitis	MPO positive	false	RKD
9	1.02	Male	Microscopic polyangiitis	PR3 positive	true	RKD
136	15.45	Male	Granulomatosis with polyangiitis	PR3 positive	false	RKD
30	3.41	Female	Microscopic polyangiitis	MPO positive	true	RKD
127	14.43	Male	Microscopic polyangiitis	MPO positive	false	RKD
19	2.16	Female	Granulomatosis with polyangiitis	PR3 positive	true	RKD
11	1.25	Female	Eosinophilic granulomatosis with polyangiitis	MPO positive	false	RKD
8	0.91	Female	Granulomatosis with polyangiitis	ELISA negative	false	RKD
23	2.61	Female	Microscopic polyangiitis	PR3 positive	false	RKD

Fig. 5. Results returned from the query in Fig. 4.

Table 6
FAIRVASC PROV-O classes.

Class	Definition	Object Properties	Data properties
Activity/ ProvenenaceActivityUplift	This class represents provenance data related to the uplift process. ProvenanceUplift is a subclass of PROV-O Activity.	prov:used, prov:generated, prov: wasAssociatedWith	prov:startedAtTime, prov:endedAtTime
ProvenanceActivityQuery	This class represents provenance data related to the querying of a registries RDF graph. ProvenanceActivityQuery is a subclass of PROV-O Activity.	prov:wasAssociatedWith, prov: used	prov:startedAtTime, prov:endedAtTime
prov:Entity	See PROV-O	prov:wasAttributedTo, prov: wasGeneratedBy	dcterms:title
prov:Agent, prov:Person	See PROV-O		foaf:name, foaf:mbox

the FAIRVASC data (e.g. including gender, yearOfBirth, alongside BVAS scores) we categorised queries into the generic query types, for example related to death, or other specific outcomes, or complications, etc. This process of categorising query types is still under way, but it is likely to correspond closely to the experimental questions currently under development by the QIT team.

4. Challenges and recommendations

4.1. Challenges

The challenges encountered during development of the FAIRVASC ontology have been related to the process of harmonisation of terms across the registries, the development of formal queries, and the implementation of the mappings with the registries. With respect to harmonisation, the following challenges were identified:

- Difficulty creating precise classifications: due to the nature of the human body, and range of diseases etc. it is often difficult to come to a shared definition for a term. This is already apparent with any analysis of existing classifications, where it can be often difficult to see whether a class is subsumed by one or another class, e.g. “Pneumocystis infection” as a sub-class of lung infection in the example in section 2.2. This is equally true when creating new classifications.
- Variance between registries: Due to the large differences between how the registries structure and record data, there must be trade offs when doing harmonisation, i.e. between creating generic classifications which cover all or most registries, but have limited use, and more specific classifications which cover only a subset of the registries but which may be more meaningful to researchers.

With respect to the development of queries:

- Aligning clinical knowledge with query capabilities: Local registry experts have the domain expertise to understand the types of queries which are relevant. It is important to communicate both the capabilities and limitations of a federated query approach.

With respect to implementation of mappings:

- Working with new technologies: As the local data schema are likely to change over time, it is necessary to have the local IT teams develop and maintain the mappings. Technologies such as Linked Data, the semantic web, RDF and R2RML are not technologies most IT teams will have used in the past.

4.2. Recommendations and future work

To address the challenges of harmonisation, it was essential to keep each registry informed about decisions being made and to have active input from them. Ultimately, it is difficult (if not impossible) to understand the specifics of a registry term definition without having experienced experts from that registry involved to explain. This is especially true when dealing with different languages. The local experts are also in a position to agree to any decisions made, for example, where a specific term is chosen in the ontology and their own term is mapped to that, and there may be slight differences in the semantics. Regular meetings with expert representatives from all registries are vital when undertaking harmonisation. These experts also have a key role in developing and agreeing to definitions for terms, so that the semantics of a term can be scrutinised, and decisions regarding its use clarified in the ontology itself.

For query development, it is again essential to have regular meetings and discussion sessions with the clinical experts. Being able to clearly demonstrate the types of queries that are supported, and have the clinical experts discuss what data they want from their queries not only drives ontology development and harmonisation, but also informs the development of the front end query interface. Here, it is also recommended to take an iterative development approach, starting small with implemented examples, and then building. This reduces the possibility of misunderstandings during development.

With respect to implementation of mappings, it is necessary to train IT staff with the necessary skills. This may take the form of formal training sessions, but it is also expected that these teams will require additional support while the mappings are initially developed, so that they can become comfortable with the technologies involved. Again, meeting regularly is important during this process.

Future work will see a second version of the FAIRVASC ontology developed which brings together a wider range of concepts of relevance to researchers of AAV. These include induction therapy, blood tests and medications. In addition, integrating additional methods, such as homomorphic encryption, to bring new layers of security to the federated query approach to ensure data privacy when revealing more fine grained information, are being explored.

5. Conclusion

This paper has addressed two key challenges to support integration of and querying over sensitive patient data related to the rare disease AAV. These are a) the need to harmonise data at a semantic level, so that queries over multiple registries return data which is comparable and b) to do so while also preserving the privacy of registry patient data, for example by returning only aggregated data. The first challenge has been met through the development of the FAIRVASC ontology and R2RML mappings using an established methodology and working with a team of domain and IT experts from each of seven European AAV data registries. These teams worked together to develop the competency questions, harmonise terms across registries, implement the ontology, create and adapt mappings for each unique registry, and then develop SPARQL queries which are implemented on a test query interface, as part of an interactive design, development and implementation life cycle. Using this approach data on over 6000 patients is now available to query.

The second challenge has been met through a combination of pseudonymization of data (during the ontology development and data uplift process) and through the development of an infrastructure which supports federated aggregate queries at each site, using the FAIRVASC query interface. Currently, a collection of specific research driven queries related to patient diagnosis and outcomes is possible, supporting also stratification by age, gender, diagnosis and ANCA specificity. The successful implementation of these queries demonstrates the viability of this approach; the infrastructure works on the basis that registries manage their data locally and, knowing the structured queries, only expose data which cannot be used to identify patients, while giving researchers and other stakeholders the capability of running aggregated queries to analyse data related to AAV from across the seven European registries.

The paper also provides a review of approaches in the literature to data integration and federated querying and shows how the approach taken by FAIRVASC when addressing sensitive patient data related to rare disease is novel. The integration of existing well known biomedical standards will also further support interoperability beyond the scope of this work. The approach has demonstrated itself to be both flexible, i.e. able to handle the range of terms across the existing registries, and extensible, i.e. able to integrate new terms as required and the proven methodology can be applied to integrate an ever growing selection of registries, and the novel use of existing technologies contributes to the growing research field on federation of medical registries.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank Nathan Lea for his work with respect to security and privacy aspects in FAIRVASC, and also all the HIT, FIT and QIT members who contributed towards this work. The work is funded by grant EJPRD19-200, the Meath Foundation 208591, Vifor Pharma Group and also the Health Research Board/Irish Nephrology Society (MRCG-2016-12).

References

[1] Modelling kidney disease using ontology: insights from the kidney precision medicine project, Nat. Rev. Nephrol. 16 (2020) 686, https://doi.org/10.1038/S41581-020-00335-W. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8012202/.

[2] M. Azarm-Daigle, C. Kuziemsky, L. Peyton, A review of cross organizational healthcare data sharing, Procedia Comput. Sci. 63 (Icth) (2015) 425–432, https://doi.org/10.1016/j.procs.2015.08.363.

[3] I.M. Bajema, J.A. Bruijn, A. Casian, M.C. Cid, E. Csernok, E. van Daalen, L. Harper, T. Hauser, M.A. Little, R.A. Luqmani, A. Mahr, C. Ponte, A. Salama, M. Segelmark, K. Suzuki, J. Sznajd, Y.K.O. Teng, A. Vaglio, K. Westman, D. Jayne, The european vasculitis society 2016 meeting report, Kidney Int. Rep. 2 (9 2017) 1018–1031, https://doi.org/10.1016/J.EKIR.2017.09.008. https://europepmc.org/articles/PMC5733672.

[4] S. Blumenthal, Improving interoperability between registries and ehrs, 2018, AMIA Summits Transl. Sci. Proc. 20 (2018). pmc/articles/PMC5961768//pmc/articles/PMC5961768/?report=abstract, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5961768/.

[5] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucleic Acids Res. 32 (1 2004) D267, https://doi.org/10.1093/NAR/GKH061. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC308795/.

[6] D. Calvanese, B. Cogrel, S. Komla-Ebri, R. Kontchakov, D. Lanti, M. Rezk, M. Rodriguez-Muro, G. Xiao, Ontop: answering sparql queries over relational databases, Semant. Web 8 (1 2017) 471–487, https://doi.org/10.3233/SW-160217.

[7] O. Can, E. Sezer, O. Bursa, M.O. Unalir, Comparing relational and ontological triple stores in healthcare domain, Entropy 19 (1) (2017), https://doi.org/10.3390/e19010030. https://www.mdpi.com/1099-4300/19/1/30.

[8] N.I. Cole, H. Liyanage, R.J. Suckling, P.A. Swift, H. Gallagher, R. Byford, J. Williams, S. Kumar, S.D. Lusignan, An ontological approach to identifying cases of chronic kidney disease from routine primary care data: a cross-sectional study, BMC Nephrol. 19 (4) (2018), https://doi.org/10.1186/S12882-018-0882-9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5894169/.

[9] F. Conrads, J. Lehmann, M. Saleem, M. Morsey, A.C. Ngonga Ngomo, Iguana: a generic framework for benchmarking the read-write performance of triple stores, in: C. d'Amato, M. Fernandez, V. Tamma, F. Lecue, P. Cudré-Mauroux, J. Sequeda, C. Lange, J. Heflin (Eds.), The Semantic Web – ISWC 2017, Springer International Publishing, Cham, 2017, pp. 48–65.

[10] C. Debruyne, K. McGlinn, L. McNerney, D. O'Sullivan, A lightweight approach to explore, enrich and use data with a geospatial dimension with semantic web technologies, in: Proceedings of the Fourth International ACM Workshop on Managing and Mining Enriched Geo-Spatial Data, GeoRich '17, ACM, New York, NY, USA, 2017, pp. 1:1–1:6.

[11] V. Dimitrieski, G. Petrović, A. Kovačević, I. Luković, H. Fujita, A Survey on Ontologies and Ontology Alignment Approaches in Healthcare, vol. 9799, Springer Verlag, 2016, pp. 373–385, https://doi.org/10.1007/978-3-319-42007-3_32.

[12] M. Ferndndez, A.G.P. rez, N. Juristo, Methontology: from Ontological Art towards Ontological Engineering, 1997. www.aaai.org.

[13] A. González-Beltrán, B. Tagger, A. Finkelstein, Federated ontology-based queries over cancer data, BMC Bioinf. 13 (1 13) (2011) 1–24, https://doi.org/10.1186/1471-2105-13-S1-S9 (1 2012), https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-S1-S9.

[14] N. Gupta, B. Gupta, Big data interoperability in e-health systems, in: Proceedings of the 9th International Conference on Cloud Computing, Data Science and Engineering, Confluence 2019, 2019, pp. 217–222, https://doi.org/10.1109/CONFLUENCE.2019.8776621.

[15] HIQA, Guidance on Classification and Terminology Standards for Ireland, 2013, pp. 1–47. December, https://www.hiqa.ie/sites/default/files/2017-07/Guidance-on-terminology-standards-for-Ireland.pdf.

[16] D.M.E. Hoque, V. Kumari, M. Hoque, R. Ruseckaite, L. Romero, S.M. Evans, Impact of clinical registries on quality of patient care and clinical outcomes: a systematic review, PLoS One 12 (9 2017), e0183667, https://doi.org/10.1371/JOURNAL.PONE.0183667. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0183667.

[17] Janev, V., Graux, D., Jabeen, H., Sallinger, E.: Knowledge Graphs and Big Data Processing State-of-the-Art . https://doi.org/10.1007/978-3-030-53199-7, http://www.springer.com/series/7409.

[18] D. Lee, R. Cornet, F. Lau, N. de Keizer, A survey of snomed ct implementations, J. Biomed. Inf. 46 (1) (2013) 87–96, https://doi.org/10.1016/j.jbi.2012.09.006. https://www.sciencedirect.com/science/article/pii/S1532046412001530.

[19] K. McGlinn, R. Brennan, C. Debruyne, A. Meehan, L. McNerney, E. Clinton, P. Kelly, D. O'Sullivan, Publishing authoritative geospatial data to support interlinking of building information models, Autom. ConStruct. 124 (2021), 103534, https://doi.org/10.1016/j.autcon.2020.103534. https://www.sciencedirect.com/science/article/pii/S0926580520311146.

[20] K. McGlinn, P. Hussey, LNCS, An Analysis of Demographic Data in Irish Healthcare Domain to Support Semantic Uplift, 12140, Springer, 2020, pp. 456–467, https://doi.org/10.1007/978-3-030-50423-6_34, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7303677/.

[21] A.J. Mohammad, An update on the epidemiology of anca-associated vasculitis, Rheumatology (Oxford, England) 59 (2020) iii42–iii50, https://doi.org/10.1093/RHEUMATOLOGY/KEAA089, 5, https://pubmed.ncbi.nlm.nih.gov/32348522/.

[22] S. Peroni, A simplified agile methodology for ontology development, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 10161 LNCS, 2017, pp. 55–69, https://doi.org/10.1007/978-3-319-54627-8_5.

[23] K.H. Pine, The qualculative dimension of healthcare data interoperability, Health Inf. J. 25 (3) (2019) 536–548, https://doi.org/10.1177/1460458219833095.

[24] M. Rastegar-Mojarad, S. Sohn, L. Wang, F. Shen, T.C. Bleeker, W.A. Cliby, H. Liu, Need of informatics in designing interoperable clinical registries, Int. J. Med. Inf. 108 (2017) 78–84, https://doi.org/10.1016/j.ijmedinf.2017.10.004. https://www.sciencedirect.com/science/article/pii/S138650561730360X.

[25] M. Reisman, EHRs: the challenge of making electronic data useable and interoperable, P T 42 (9) (2017) 572–575.

[26] J. Robson, H. Doll, R. Suppiah, O. Flossmann, L. Harper, P. Höglund, D. Jayne, A. Mahr, K. Westman, R. Luqmani, Damage in the anca-associated vasculitides: long-term data from the european vasculitis study group (euvas) therapeutic trials, Ann. Rheum. Dis. 74 (2015) 177–184, https://doi.org/10.1136/ANNRHEUMDIS-2013-203927. https://pubmed.ncbi.nlm.nih.gov/24243925/.

[27] A.C. Sima, T.M. de Farias, E. Zbinden, M. Anisimova, M. Gil, H. Stockinger, K. Stockinger, M. Robinson-Rechavi, C. Dessimoz, Enabling semantic queries across federated bioinformatics databases, Database (2019), https://doi.org/10.1093/DATABASE/BAZ106 (1 2019), https://academic.oup.com/database/article/doi/10.1093/database/baz106/5614223.

[28] J.Y. Sleeman, J.A. Hammad, A review of accessing big data with significant ontologies, Knowl. Eng. Data Sci. 3 (2) (2020) 67–76, https://doi.org/10.17977/um018v3i22020p67-76. http://journal2.um.ac.id/index.php/keds/article/view/15795.

[29] M.C. Suárez-Figueroa, A. Gómez-Pérez, M. Fernández-López, The NeOn Methodology for Ontology Engineering, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 9–34, https://doi.org/10.1007/978-3-642-24794-1_2.

[30] Y. Sure, S. Staab, R. Studer, On-To-Knowledge Methodology (OTKM), Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, pp. 117–132, https://doi.org/10.1007/978-3-540-24750-0_6.

[31] G.M. Weber, Federated queries of clinical data repositories: scaling to a national network, J. Biomed. Inf. 55 (6 2015) 231–236, https://doi.org/10.1016/J.JBI.2015.04.012.

[32] N. Yadagiri, P. Ramesh, Semantic web and the libraries: an overview, Int. J. Libr. Sci.^{TM} 7 (1) (2013) 80–94. www.ceserp.com/cp-jour. http://www.ceser.in/ceserp/index.php/ijls/article/view/2989.

[33] K. Yin, F. Jeffrey, D. Rebecca, Z. Lina, Disease specific ontology of adverse events: ontology extension and adaptation for chronic kidney disease, Comput. Biol. Med. 101 (2018) 210–217, https://doi.org/10.1016/J.COMPBIOMED.2018.08.024, 10, https://pubmed.ncbi.nlm.nih.gov/30195820/.

[34] Y.W. Yu, G.M. Weber, Federated queries of clinical data repositories: balancing accuracy and privacy, bioRxiv (11 2019), 841072, https://doi.org/10.1101/841072. https://www.biorxiv.org/content/10.1101/841072v1.abstract.

[35] L. Zemmouchi-Ghomari, A.R. Ghomari, Ontology versus terminology, from the perspective of ontologists, Int. J. Web Sci. 1 (2012) 315–331.