Towards a rare disease registry standard: semantic mapping of common data elements between FAIRVASC and the European Joint Programme for Rare Disease

Beyza Yaman[1 [0000−0003−2130−0312]], Kris McGlinn[1 [0000−0002−7023−5169]], Lucy Hederman[1 [0000−0001−6073−4063]], Declan O'Sullivan[1 [0000−0003−1090−3548]] and Mark A. Little[1,2 [0000−0001−7318−375]]

[1] ADAPT Centre, Trinity College Dublin, Dublin, Ireland
[2] Trinity Translational Medicine Institute, Trinity College Dublin, Dublin, Ireland
{beyza.yaman,kris.mcglinn,lucy.hederman,
declan.osullivan,mark.little}@adaptcentre.ie

Abstract. This paper describes the extension of the FAIRVASC rare disease ontology, with Joint Research Council Common Data Elements (CDE), and mapping to the European Joint Programme on Rare Diseases (EJP RD) CDE ontology. We use the rare autoimmune condition ANCA vasculitis as a model disease to illustrate this. Semantic modelling of CDE for Rare Diseases over registry data is important to represent the specific concepts around these conditions. We describe the development of an ontology which facilitates simultaneous uplift of tabular data into a common RDF format from several registries. The ontology allows the data to be integrated across the registries and increases the interoperability and standardisation among datasets, thus enhancing collaboration with external stakeholders. The ontology therefore creates an effective rare disease research environment which enables the disease and its impact on the patient to be investigated in an effective manner across national borders. This paper presents the methodology and road map to implement the CDE ontology for the health domain.

Keywords: ontology · health informatics · rare disease.

1 Introduction

The quantity of digital health data sources has been growing immensely, with resultant increase in global complexity and size. These data sources provide structured or unstructured clinical information which makes data management and analysis a highly important task in the medical domain. With improved management and analysis of health data there is significant potential to discover new solutions for difficult health challenges, particularly in the rare disease space, where datasets are sparse and distributed across multiple countries.

Autoimmune disease is one such area, where the cause of the disease is unknown and management unclear. Immune systems that are overactive are directed against self antigens which harm and damage the body's own tissues (autoimmune diseases)[8]. The immune system may produce antibodies that, instead of defending against infections, attack the body's own tissues in reaction to an unknown stimulus. The goal of autoimmune disease treatment is to reduce the immune system activity; however this treatment can cause the patient to have severe infections and in some cases increases the risk of cancer. Large datasets are required to study these challenges, but these are rarely available for use in one location, necessitating a combination of multiple datasets.

The European Joint Programme on Rare Diseases (EJP RD) unites 130 institutions from 35 countries to create an effective ecosystem between research, care and medical innovation. EJP RD has two major objectives: i) Through the creation, demonstration, and promotion of Europe/world-wide sharing of research and clinical data, materials, processes and expertise to increase the integration, effectiveness, production, and social impact of rare disease research. ii) Implement and further develop an efficient model of financial support for all categories of rare disease research as well as rapid utilisation of research findings for patient benefit [3]. This will enhance the lives of individuals with rare diseases by giving new and improved treatment choices and diagnostic tools. The Common Data Elements ontology is one such tool to increase the research efficiency in the domain.

The PersonAlisation of RelApse risk in autoimmune DISEase (PARADISE) project is a specific example of a programme that studies precise tailoring of immunosuppressive drugs and prediction of disease reactivation. PARADISE is an interdisciplinary project bringing computer scientists, clinicians and health informatics together to solve this problem. It builds upon the FAIRVASC EU project[4] and the AVERT project[5], which have established the foundation for the proposed semantic web technology approach. The use of semantic web based ontologies underpins integration of different data sources in these projects[6]. Ontologies are used to describe a domain with formalised definitions and axioms to infer more meaningful information from the data [4]. The standard definition of the concepts in the domain through an ontology can enable the straightforward integration of the data with the least effort possible. Such definition of concepts and relations through a W3C standard based representation of the ontology also allows for the use of pre-existing tools and applications for health data and maximises interoperability of systems in the domain. This is in line with the European Commission recommendation on interoperability of electronic health record systems across borders such that any other system or application in Europe can comprehend and interpret the information that has been shared[6]. It is also significant to re-use the ontologies created in a specific domain to increase

[3] https://www.ejprarediseases.org/what-is-ejprd/project-structure/
[4] https://fairvasc.eu/
[5] https://www.tcd.ie/medicine/thkc/avert/
[6] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32008H0594&from=EN

the interoperability and reduce the engineering time and effort. Interoperability is important in the health domain because while clinicians can make the distinction of the same illness even if they have been described in a different way, machines cannot understand that disinction. Thus, semantic interoperability establishes a shared understanding that allows computers to communicate reliably. The capacity of multiple ontologies to map diverse concepts to shared semantics, or meaning, is essential for machine communication. Communicating data in an effective way is a challenging task without semantic interoperability across diverse healthcare IT systems. In our case we are using declarative mapping as a way of specifying how data can be transformed from a data model according to one ontology to a data model according to a different ontology.

Use of standard based approaches also has a huge impact on the understanding of the data as well as concluding meaningful solutions. Another proposal of relevance is the ISO/IEC 11179 Metadata Registry[7] standard, which is a global ISO standard for expressing metadata in a metadata registry for an organisation, which is also used for the health registry domain. It outlines the process of standardising and registering metadata in order to make data more comprehensible and shared. The Common Data Elements (CDE) ontology, developed by the EU Joint Research Centre and implemented by The European Joint Programme on Rare Diseases[8](EJP RD) is another approach to standardise the modelling of a rare disease.

The primary focus of this paper is to map the CDE onto the FAIRVASC ontology using the ontology developed by the EJP RD, thereby linking FAIRVASC and EJP RD. The development of the mapping demonstrates the utility of taking an ontology based approach to support interoperability in the health domain. The CDE extension of the FAIRVASC ontology, which represents all of the EJP RD Common Data Elements, is therefore the study's contribution. Because the CDE ontology is broad and generic, we worked from the CDE elements pertaining to our project first, and created the mappings to the generic CDE ontology. Creating these mappings enables rare disease registries interoperability to inform clinical care and increase the understanding among registries and clinicians. Thus, in this project, FAIRVASC ontology is adopted and extended to build on the PARADISE project.

The paper is structured as follows: Section 2 presents the FAIRVASC Project and EJP RD CDEs upon which this work is built. Section 3 introduces the PARADISE project. Section 4 discusses the newly created attributes for FAIRVASC ontology, needed to comply with the EJP RD proposal, and the mappings between FAIRVASC and CDE ontologies. Finally, Section 6 presents the conclusions and future work.

[7] https://www.iso.org/standard/60341.html
[8] https://www.ejprarediseases.org/

2 FAIRVASC Project

This section presents the concepts which PARADISE project is depending on. Section 2.1 presents the elements that underpin the relationship between FAIRVASC, PARADISE and EJP RD. Section 2.2 introduces the FAIRVASC ontology. Section 2.3 provides a brief introduction to the Common Data Elements ontology proposed by EJP RD.

2.1 Relationship between FAIRVASC, PARADISE and EJP RD

The Rare Kidney Disease Registry and Biobank[9] (RKD) is one of the seven FAIRVASC registries and is also the source registry for the PARADISE project. It was founded in 2012 to conduct research on rare kidney disease in Ireland.

The RKD Registry data records data on most patients with ANCA vasculitis in Ireland. Clinical study cases are distributed and tracked over multiple hospitals, research institutes and clinics. Due to the rare nature of the disease, the low patient numbers require all the research institutes and clinics to be involved in the study. The RKD Registry data is collected from the patients manually through hospital registration and a mobile patient application called patientMpower[10]. Using both infrastructures enables longitudinal tracking of the patient's condition. The data is stored in the REDCap web platform[11] which provides a secure platform for managing and maintaining online databases. REDCap provides automated export procedures as well as common statistical packages, ad-hoc reporting tools.

As one of the rare disease registries embedded within the European Reference Network for rare immune disorders, ERN-RITA[12], the RKD registry seeks to be fully interoperable with the registry structure envisaged by EJP RD through its engagement with FAIRVASC. When designing the PARADISE project, the FAIRVASC ontology was adapted and mapped onto the CDE of the EJP RD ontology. Figure 2 presents the relation between FAIRVASC and EJP RD. Registry datasets are uplifted using R2RML mappings and FAIRVASC ontology. Uplifted data is stored in a triplestore which is then later queried via a query interface. The end user can pose queries using FAIRVASC and CDE ontologies using created mappings. PARADISE project employs the extended FAIRVASC ontology for its use case.

2.2 FAIRVASC Ontology

FAIRVASC is a research project of the European Vasculitis Society and RITA European Reference Network, bringing together computer scientists, clinicians and patient organisations. It comprises ten partners across Europe that represent

[9] https://www.tcd.ie/medicine/thkc/research/rare.php
[10] https://info.patientmpower.com/
[11] https://projectredcap.org/software/
[12] https://ern-rita.org/

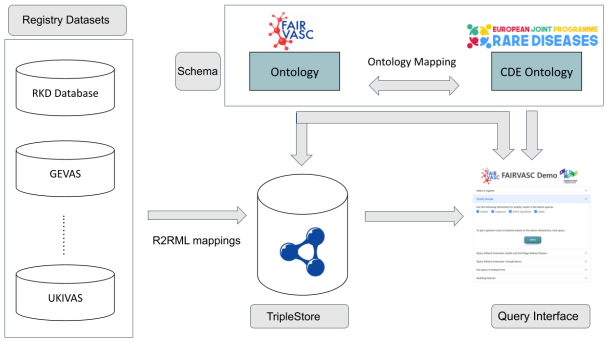

Fig. 1. Relation between FAIRVASC and CDE ontologies

all aspects of care of patients with the rare disease ANCA vasculitis. Seven national registries are partners in the project, namely, Ireland's RKD registry, the UK's UKIVAS registry, the French Vasculitis Study Group registry, the Czech Vasculitis Registry, the Polish Vasculitis Registry POLVAS, the GEVAS German/Austrian/Swiss registry and Sweden's Skåne Vasculitis Inception Cohort.

The FAIRVASC ontology was created to manage data related to ANCA vasculitis, and is based on the harmonisation of terms in the seven registries. Each registry provides feedback on proposed harmonised clinical terms. Once agreed, these harmonised terms are formally integrated into the FAIRVASC ontology. Protege[7], an ontology building tool, was used to create the ontology[13] (Figure 1) for a snapshot of the FAIRVASC class hierarchy in Protege. This tool was used to define each class, its connections, and data characteristics (those in bold are FAIRVASC classes, the rest are from the Birmingham Vasculitis Activity Score[14] ontology which was developed to standardise the representation of Birmingham Vasculitis Activity Scoring across the registries as RDF). There are 9 top-level classes, 13 classes total, 9 object attributes, and 24 data properties in the ontology. Patient, Patient Overview, Diagnosis, Clinical Outcomes, Encounter, Clinical Test, and Organ Pattern are the top-level classes.

Rather than creating a new ontology, existing ontologies (or a part of an existing ontology) are employed in the project. NCIT, SNOMED-CT, Medical Dictionary for Regulatory Activities Terminology (MedDRA) and the Orphanet nomenclature (ORPHAcode) ontologies are explored. However, it was seen that some of the ontologies had highly restrictive licensing (e.g. SNOMED-CT), thus mappings were created between the less restrictive ontologies (NCIT and Orphanet) in order to add rich semantics to the data.

[13] http://ontologies.adaptcentre.ie/fairvasc/
[14] http://ontologies.adaptcentre.ie/bvas/

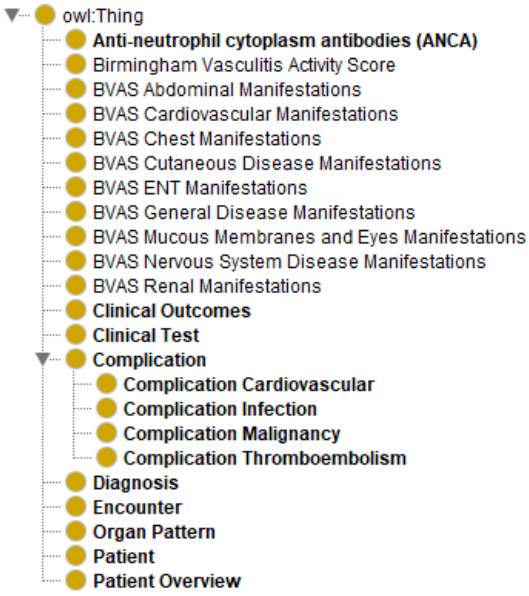

Fig. 2. Overview of FAIRVASC ontology classes taken from Protégé

2.3 Common Data Elements for Rare Diseases

The European Joint Programme on Rare Diseases (EJP RD) aims to enhance treatment for individuals with rare diseases as part of wider European Union-wide initiatives to coordinate actions to address data fragmentation concerns in European medical registries. To stimulate and facilitate research in the rare illnesses sector, EJP RD has designed a virtual platform architecture, which is a service-oriented eco-system of interconnected online services. They identified five key services: authentication and authorization, rare disease data discovery and elaboration, data request and access, dataset enhancement (e.g. pseudonymization), and services for making resources FAIR for federated use, such as catalogues of ontological models for rare disease data and metadata[15]. EJP RD emphasises the use of standard data representations such as ontologies and their reuse, as well as use case driven design. The work in this article is well-suited to the EJP's general structure.

The EU Joint Research Centre initially created Common Data Elements for Rare Diseases[1] which comprises 16 data components that must be reported by each European rare illness registry as they are considered crucial for future study. The elements are grouped around 8 main subjects namely pseudonymization of the patient, a patient's personal information, status, diagnosis, disease history, care pathway, disability as well as information on consent for research purposes. A Common Data Elements ontology[5] was created to model the components

[15] https://www.ejprarediseases.org/

semantically[16]. The authors proposed a generic semantic data model of the set of common data elements for rare disease registration. CDE ontology is based on Semantic Science Integrated Ontology (SIO)[2] as the core framework for representing the entities and their relationships. The ontology also integrates the Orphanet Rare Disease Ontology[9], Human Phenotype Ontology[**?**] and National Cancer Institute Thesaurus[3] to describe the diagnoses. The authors also provide a list of templates to convert the CDE data CSV files to RDF in a semi-automatic way, as well as a SHeX model to enable the validation of the converted files.

3 PARADISE Project

The PARADISE project stems from data generated in the Irish RKD registry, which is also a FAIRVASC registry. It addresses the greatest challenge in autoimmune disease, which is to devise personalised strategies for precise tailoring of immunosuppressive drugs to prevent disease relapse or "flare". PARADISE specifically focuses on ANCA vasculitis which is an archetypal heterogeneous relapsing and remitting chronic autoimmune disease. Due to the lack of strong prognostic techniques, clinicians treating ANCA vasculitis use conventional dosage regimens that give time-based immunosuppressive drugs dose modifications but presume no disease progression and provide limited customised accounting. Deviations from these regimens are based on clinician intuition and experience, as well as recognized biomarkers (e.g., urinalysis/autoantibody levels) and an estimate of past immunosuppressive drugs exposure. Future flare risk predictions are subjective, inconsistent, and frequently inaccurate.

The PARADISE project is being led by the SFI ADAPT centre and Trinity School of Medicine. The goal is to investigate a novel solution for individualised flare risk prediction in autoimmune disease and prevention of over-treatment with immunosuppressive drugs. The PARADISE consortium aims at solving this problem by combining semantic web technologies, clinical expertise, targeted biomarker analysis, and patient-sourced health data by integrating readily implementable data streams in the physician workflow and patient self-management tools.

4 FAIRVASC-CDE Mappings

In order to develop the mapping between the FAIRVASC and CDE ontologies, we adopted a consultation process bringing together health informaticians, clinicians and computer scientists deeply knowledgeable of the main concepts in the rare kidney disease domain. A three step methodology was followed to extend this ontology: i) EJP-RD CDE ontology was compared with the FAIRVASC ontology and missing data elements in FAIRVASC ontology were detected. ii) Missing CDE related to the existing concepts were created for the FAIRVASC ontology

[16] https://github.com/ejp-rd-vp/CDE-semantic-model

iii) The mappings between the ontologies were created. Moreover, following the creation of ontology concepts, Relational Database to RDF Mapping Language (R2RML) scripts were written to uplift the registry data. As the FAIRVASC ontology is aligned with the RKD database, the database has already had most of the relevant concepts in the data. During design of the FAIRVASC ontology, the EJP RD concepts were constantly reviewed and, where possible, we adopted the concepts suggested by the CDE ontology in order to increase the potential for interoperability of RKD datasets and therefore PARADISE project. The review of the ontologies was undertaken manually.

Table 1. FAIRVASC-CDE Mappings

FAIRVASC	CDE-semantic model	Type of Relationship
	1. Pseudonym	
fvc:patientID	cde:identifier / sio:SIO_000300	owl:equivalentProperty
	2. Personal information	
fvc:yearOfBirth (needs transformation)	cde:birthdate_output / SIO:SIO_000300	owl:equivalentProperty
fvc:gender	cde:sex_output / sio:SIO_000300	owl:equivalentProperty
	3. Patient Status	
fvc:death (needs transformation)	cde:status_output / sio:SIO_000300	rdfs:subPropertyOf
fvc:dateOfDeath	cde:death_information_output / sio:SIO_000300	owl:equivalentProperty
	4. Care pathway	
fvc:dateOfCarePathwayStart	cde:carepathway_startdate / sio:SIO_000300	owl:equivalentProperty
	5. Disease history	
fvc:dateOfDiagnosis	cde:diagnosis_startdate / sio:SIO_000300	owl:equivalentProperty
fvc:dateAtOnset	cde:symptom_onset_output / sio:SIO_000300	owl:equivalentProperty
	6 Diagnosis	
fvc:mainDiagnosis	cde:diagnosis_attribute / rdf:type	owl:equivalentProperty
	7. Research	
fvc:BiobankLink	cde:biobank_identifier / sio:SIO_000300	owl:equivalentProperty
fvc:futureContactConsent	cde:consent_output / sio:SIO_000300	rdfs:subPropertyOf - (inverse relation)
fvc:dataUseInRegistryConsent	cde:consent_output / sio:SIO_000300	rdfs:subPropertyOf - (inverse relation)
fvc:biobankConsent	cde:consent_output / sio:SIO_000300	rdfs:subPropertyOf - (inverse relation)

A set of mappings (Table 1) was then developed to ensure the interoperability of the datasets within the consortium and with CDE ontology based datasets. Thirteen of the 16 common elements of the CDE ontology align well with the FAIRVASC ontology. The other three elements are not recorded in the FAIRVASC ontology: “Undiagnosed case”, “Genetic case” and “Classification of functioning, disability” attributes. The undiagnosed case does not exist in the database because FAIRVASC data is specialised on ANCA vasculitis, so any other diagnosis is not recorded in the database. ANCA vasculitis is not a genetic disease, thus that option also does not exist. Similarly, the disability option is

not recorded in the FARVASC ontology. This difference is not surprising as the CDE ontology is targeted at representing diseases in a generic manner, whereas the FAIRVASC ontology has been created to support concepts specific to ANCA vasculitis.

In particular, there are two types of mappings created for the interoperability of the datasets. One of them is property equivalence (`owl:equivalentProperty`) which describes that two properties have same"values". Other one is subproperty declaration (`rdfs:subPropertyOf`) meaning property extension of the property (e.g. death) are also members of the property extension of the high level property (e.g. status output)[17]. On the other hand, `fvc:death` requires transformation because the death information is kept as a boolean value while CDE ontology stores this piece of information as a string. Tranformation will allow the boolean data type to be converted to string before the relation mapping. Other important point on the table is that the last three `rdfs:subpropertyOf` relations are described as inverse relations which means CDE semantic model concept is a subproperty of FAIRVASC concept due to its broadness.

Listing 1.1 presents the R2RML script for the `Diagnosis` concept. FAIRVASC ontology had already have the `Diagnosis` concept, thus, missing CDE components were added to the script to be able to produce these triples as well. FAIRVASC ontology and R2RML scripts are available in an open repository[18].

Listing 1.1. Example R2RML Mapping for Schema Uplifting

```
<#diagnosis>
        rr:subjectMap [
                rr:template "http://data.fairvasc.ie/resource/
                rkd/diagnosis/{RECORD_ID}{DIAGNOSIS_DATE}" ;
                rr:class fvc:Diagnosis;
        ] ;
        rr:predicateObjectMap [
                rr:predicate fvc:mainDiagnosis ;
                rr:objectMap [
                        rr:column "DIAGNOSIS" ;
                        rr:datatype xsd:string;
                ] ;
        ] ;
        rr:predicateObjectMap [
                rr:predicate fvc:hasDateAtOnset ;
                rr:objectMap [
                        rr:column "DATE_OF_SYMPTONS" ;
                        rr:datatype xsd:dateTime;
                ] ;
        ] .
```

The mappings are evaluated in 2 ways: i) Semantic validation is approved by the group of 10 people at the end of the discussions. ii) Technical validation is conducted by producing the triples from the mappings which has been around 50000 triples.

[17] https://www.w3.org/TR/owl-ref/#subPropertyOf-def
[18] https://opengogs.adaptcentre.ie/yamanbey/PARADISE

5 Lessons Learned

We conducted collaborative meetings with clinicians and computer scientists. Even if pandemic has been a non-pleasant experience for everyone, it paved the way for regular and timely online meetings and discussions. Communicating computer science terms with researchers who have medicine background could be challenging, however, due to their experience with Semantic Web since the beginning of the FAIRVASC project it was an straightforward task. Analysis of the RKD data was conducted manually and this is a time-consuming process but it is a necessary step to create R2RML mappings and uplift data. In our use case, the adoption of semantic technologies and creating mappings between FAIRVASC and CDE ontologies for rare diseases has shown the following: i) EJP RD ontology has been implemented in a generic way to be able model all the rare diseases, thus, some concepts does not apply to our registry (e.g. genetic disease concept) ii) interdisciplinary nature of the project shows that overlong problems in health domain could be solved via mixed-proficient interest group collaborations iii) created mappings increases the understanding between registries from various countries by specifically declaring equality and hierarchy among concepts.

6 Conclusions and Future Work

The health domain data is increasing as well as its complexity and volume. Thus, it is highly important to keep the semantic interoperability as rich as possible to reduce the semantic heterogenity and increase the understanding among registries. There is a high demand on ontologies from the health domain and put the data in context. We have described the development of a model which can be used to uplift tabular data into a common RDF format. Linkage of the created ontology will help the data to be integrated and increase the interoperability among datasets including collaboration with external groups and projects. Other researchers could follow the proposed steps and create the mappings for their systems to enrich their data. On the other hand, users who have experience with the mapped ontology could benefit from these mappings and query data according to their knowledge. This will provide more complex SPARQL query feature to the system without losing the practicality.

Acknowledgements

This work is funded by grant EJPRD19-200, the Meath Foundation 208591 and also the Health Research Board / Irish Nephrology Society (MRCG-2016-12). Research also supported by ADAPT SFI Research Centre (grant number 13/RC/2106_P2).

References

1. Set of Common Data Elements , EU RD Platform . https://eu-rd-platform.jrc.ec.europa.eu/set-of-common-dataelements_en. [Online; accessed 24.02.2022].
2. M. Dumontier, C. J. Baker, J. Baran, A. Callahan, L. Chepelev, J. Cruz-Toledo, N. R. Del Rio, G. Duck, L. I. Furlong, N. Keath, et al. The semanticscience integrated ontology (sio) for biomedical research and knowledge discovery. *Journal of biomedical semantics*, 5(1):1–11, 2014.
3. J. Golbeck, G. Fragoso, F. Hartel, J. Hendler, J. Oberthaler, and B. Parsia. The national cancer institute's thesaurus and ontology. *Journal of Web Semantics First Look 1_1_4*, 2003.
4. R. Hoehndorf, P. N. Schofield, and G. V. Gkoutos. The role of ontologies in biological and biomedical research: a functional perspective. *Briefings in bioinformatics*, 16(6):1069–1080, 2015.
5. R. Kaliyaperumal, M. D. Wilkinson, P. A. Moreno, N. Benis, R. Cornet, B. dos Santos Vieira, M. Dumontier, C. H. Bernabe, A. Jacobsen, C. M. Le Cornec, et al. Semantic modelling of common data elements for rare disease registries, and a prototype workflow for their deployment over registry data. *medRxiv*, 2021.
6. K. McGlinn and P. Hussey. An analysis of demographic data in irish healthcare domain to support semantic uplift. In *International Conference on Computational Science*, pages 456–467. Springer, 2020.
7. N. F. Noy, M. Crubézy, R. W. Fergerson, H. Knublauch, S. W. Tu, J. Vendetti, and M. A. Musen. Protégé-2000: an open-source ontology-development and knowledge-acquisition environment. In *AMIA... Annual Symposium proceedings. AMIA Symposium*, pages 953–953, 2003.
8. P. Travers, M. Walport, M. J. Shlomchik, and M. Janeway. *Immunobiology: the immune system in health and disease*. Churchill Livingstone, 1997.
9. D. Vasant, L. Chanas, J. Malone, M. Hanauer, A. Olry, S. Jupp, P. N. Robinson, H. Parkinson, and A. Rath. Ordo: an ontology connecting rare disease, epidemiology and genetic data. In *Proceedings of ISMB*, volume 30. researchgate. net, 2014.