German Medical Data Sciences: Bringing Data to Life
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI210052

Ontological Modelling and Execution of Phenotypic Queries in the Leipzig Health Atlas

Alexandr UCITELI[a,1], Christoph BEGER[a,b], Jonas WAGNER[a,c], Alexander KIEL[c], Frank A. MEINEKE[a], Sebastian STÄUBERT[a], Matthias LÖBE[a], René HÄNSEL[a], Judith SCHUSTER[a], Toralf KIRSTEN[c,d], Heinrich HERRE[a]
[a] *Institute for Medical Informatics, Statistics and Epidemiology (IMISE), University of Leipzig, Germany*
[b] *Growth Network CrescNet, University of Leipzig, Germany*
[c] *LIFE Research Centre for Civilization Diseases, University of Leipzig, Germany*
[d] *Faculty Applied Computer and Biological Sciences, University of Applied Sciences Mittweida, Germany*

Abstract. Sharing data is of great importance for research in medical sciences. It is the basis for reproducibility and reuse of already generated outcomes in new projects and in new contexts. FAIR data principles are the basics for sharing data. The Leipzig Health Atlas (LHA) platform follows these principles and provides data, describing metadata, and models that have been implemented in novel software tools and are available as demonstrators. LHA reuses and extends three different major components that have been previously developed by other projects. The SEEK management platform is the foundation providing a repository for archiving, presenting and secure sharing a wide range of publication results, such as published reports, (bio)medical data as well as interactive models and tools. The LHA Data Portal manages study metadata and data allowing to search for data of interest. Finally, PhenoMan is an ontological framework for phenotype modelling. This paper describes the interrelation of these three components. In particular, we use the PhenoMan to, firstly, model and represent phenotypes within the LHA platform. Then, secondly, the ontological phenotype representation can be used to generate search queries that are executed by the LHA Data Portal. The PhenoMan generates the queries in a novel domain specific query language (SDQL), which is specific for data management systems based on CDISC ODM standard, such as the LHA Data Portal. Our approach was successfully applied to represent phenotypes in the Leipzig Health Atlas with the possibility to execute corresponding queries within the LHA Data Portal.

Keywords. Phenotype, Biomedical Ontologies, Information Storage and Retrieval, Selection Criteria, Domain-specific Language, Web Archive, Metadata

1. Introduction

Clinical trials, epidemiological studies and other research projects are typically used to determine medical phenomena. Commonly, these projects produce and use data about

[1] Alexandr Uciteli, IMISE, University of Leipzig, Härtelstraße 16-18, 04107 Leipzig, Germany; E-mail: auciteli@imise.uni-leipzig.de

patients potentially having a disease under investigation, inhabitants of geographical regions of interests, and biosamples taken from them. While these data are often very specific for the intended goal of the trial, study, and research project, they can be of high impact for other research projects, too. Meta-studies, for instance, compare the outcome of different trials and studies, but often use data contained in research papers and corresponding supplemental material available on authors' web pages, etc. Moreover, new trials and studies could refer to data of previous projects, e.g., when they compare their outcome with the outcome of the own study, when they directly merge their data with own data in a scientific analysis, or when they want to reproduce the results by reusing data from such a previous study. Making medical data of patients available for other scientific initiatives and projects fundamentally requires both, a technical infrastructure and organizational principles on how data can be accessed.

The Leipzig Health Atlas (LHA) [1,2] is a technical and organizational platform in the described context. The aim of this platform is twofold. On the one hand, it provides a technical infrastructure for scientists allowing archiving, presenting and secure sharing a wide range of publication results. These results include published data, their corresponding metadata as well as several kinds of models (e.g., statistical, risk, and simulation models) that have been implemented in novel software tools. All these are stored within the LHA according to the FAIR data principles (findable, accessible, interoperable and reusable) [3]. On the other hand, fellow scientists, physicians and interested people can use data, metadata and presented software tools for their own research. While some of these are freely available, the access to others is restricted and necessitates the user to register on the LHA platform. Then, he can upload and manage his research results and grant access permissions to other research groups or individual users or basically request access to data and toolsets. To meet these requirements, the LHA platform reuses three different major components that have been previously developed by other projects. Fundamentally, the SEEK [4] management platform is the overall hosting platform that structures projects, studies and the produced data, models, and software tools according to the ISA (investigation, study, assay) [5] framework. While the SEEK platform only allows to represent research data as data files, the LHA Data Portal [6] (a clone of the LIFE Data Portal [7]) allows to create and execute queries against metadata and research data to find out a) what data elements are available in this data collection and b) sample sizes (also called feasibility queries) meeting specific filter criteria of interest. Finally, the PhenoMan [8] is an ontological framework allowing to model and represent phenotype algorithms.

In this paper, we focus on the interrelation between these major components of the LHA platform, i.e., joining the SEEK platform with the LHA Data Portal and the PhenoMan (Fig. 1 a). The aim is to model and automatically generate queries using a novel ontology-based framework. The queries are represented on the web portal (LHA) and executed by a data repository (LHA Data Portal). Such queries support the identification and classification of individuals (persons or organisms) whose properties meet specific phenotype classes (e.g., 'select men aged 40-60 with myocardial infarction'). We therefore call such queries phenotypic queries throughout this paper. The queries can be useful, e.g., for feasibility studies, to define study cohorts [9] or to provide the best available care for each patient based on stratification into phenotype subclasses [10].

2. State of the Art

Phenotypic queries are comparable with eligibility criteria queries and can be used for the same purpose. Therefore, we focus on approaches for ontological modelling and expression of eligibility criteria in a domain-specific language (DSL).

Different approaches for selecting eligible study participants using a DSL are described in the literature. The medical query language SNAggletooth Query Language (SNAQL) was proposed by Bucur et al. [9] to formalize and execute medical query statements. The main purpose of SNAQL is to define queries in a language close to natural language, while keeping the implementation effort to a minimum. The SNAQL syntax enables the definition of medical concepts, the specification of contextual constraints as well as using unary and binary operators (e.g., and, or, not). Bucur et al. applied SNAQL to a software application to create and select patient cohorts by defining filters specifying characteristics of the desired population.

Zhang et al. [11] compares three kinds of patient query paradigms (called A, B, and C) used in clinical trial recruitment. In paradigm A, a clinical engineer or database administrator translates clinical researchers' natural language criteria into database-specific query language (e.g., SQL). In paradigm B, this human "query translator" is replaced by a formal model, such as an ontology. The user then builds filter criteria using the predefined concepts of this model (ontology). The criteria are then translated to SQL by the underlying mapping between the formal model and the individual database. In paradigm C, recruitment criteria are rule-based captured, which are then translated into a query-oriented DSL syntax. The DSL is then automatically transformed to SQL. The authors focus on last paradigm and use LINQ (language-integrated query) as DSL. LINQ supports nearly 40 operators, including "select", "from", "in", "where" and "order by".

Our approach combines both paradigms, B and C. We use an ontology to support query formulation by the user and translation into a final query language using a DSL.

A number of groups are developing knowledge representation formalisms for eligibility criteria [12]. Zhang et al. [13] present an approach to make eligibility criteria computable using the ontology-based data access framework Ontop [14]. The Ontology for Computable Eligibility Criteria was created to represent the eligibility criteria of Hepatitis C trials. SPARQL queries are utilized to query relational databases as virtual RDF graphs. In [15], the Eligibility Rule Grammar and Ontology (ERGO) is used to annotate free-text criteria and to transform them into a computable form. Chondrogiannis et al. [16] proposed a CDISC-compliant schema to organize criteria along with a patient-centric model for their formal expression. The Eligibility Criteria Ontology (EC-O) was developed to cover a wide range of parameters allowing the specification of eligibility criteria in clinical trials. The eligibility criteria are formally represented in XML or SPARQL. The XML schema as well as the structure of the SPARQL queries are based on the EC-O.

3. Concept

3.1. Study Data Query Language (SDQL)

Following the predefined structure of SEEK, the Leipzig Health Atlas gives an overview about research projects, studies (incl. trials), datasets and models. Study data and metadata reside in files in different formats. Some of them are human-readable and even

browsable within the LHA web page. For metadata in the CDISC ODM (Operational Data Model [17]) format, we offer a more detailed approach with the LHA Data Portal. The LHA Data Portal is based upon the LIFE Data Portal (LDP) [7,18], a software infrastructure that has been developed at the Leipzig Research Center for Civilization Diseases (LIFE). This infrastructure was adopted and reused for the LHA platform. In particular, metadata describing publication data has been imported into the LHA Data Portal. In this way, study metadata are browsable and findable for an interested user using the search interface of the Data Portal. Moreover, clinical data being available in CDISC ODM format were imported into LHA Data Portal. This makes it possible to execute feasibility queries directly on clinical data. This feature has been implemented to query the current state of ongoing studies and is now utilized to enable a query mechanism for the Leipzig Health Atlas. The Data Portal utilizes a novel DSL, the Study Data Query Language (SDQL) [19], to specify eligibility criteria. By describing the DSL using a context-free grammar, it can be ensured that inclusion and exclusion criteria have been specified syntactically correct. Figure 1 b) shows the SDQL grammar in Backus-Naur form. The terms of the language are strongly oriented towards ODM, in order to use SDQL also outside the LIFE Research Centre and LHA.

Individual expressions of the language are explained in the course of the paper using some examples. More details about and an evaluation of the Data Portal as well as the developed SDQL will be part of a future paper.

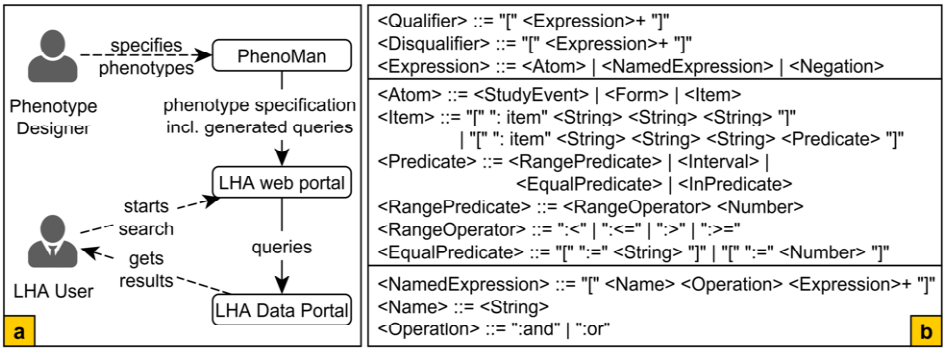

Figure 1. a) System components b) SDQL in Backus-Naur form (selected fragments)

3.2. Ontological Modelling of Phenotypes

We developed an ontology-based phenotyping framework, the Phenotype Manager (PhenoMan), to model and execute phenotype algorithms [8]. The underlying ontological model, the Core Ontology of Phenotypes (COP), distinguishes between single, combined and derived phenotype classes. We define single phenotypes as single properties of an organism (e.g., weight or height). A combined phenotype is a combination of single phenotypes (e.g., a combination of weight and height), whereas a derived phenotype is an additional property (e.g., BMI) derived from the corresponding phenotypes by mathematical calculations (e.g., BMI = weight[kg] / height[m]^2). Additionally, we differentiate between restricted and non-restricted phenotype classes. For instance, the phenotype class ‘Gender’ is associated with all genders of all living beings (instances) and, thus, the class is non-restricted, i.e., the class contains instances as many as living

beings exist. Similarly, the phenotype class 'Female' contains genders of all available females, and thus, is restricted to only this set.

Let us consider an example. The World Health Organization (WHO) recommends the following cut-off points for waist circumference as an indicator for risk of metabolic complications [20]:

- > 94 cm (male); > 80 cm (female): Increased risk,
- > 102 cm (male); > 88 cm (female): Substantially increased risk.

We use the Phenotype Algorithm Specification Ontologies (PASO) [8] to model specific phenotypes (algorithms). In PASO, the individual properties (e.g., gender and waist circumference) are specified as Abstract Single Phenotype (ASiP) classes (Fig. 2: A). A Restricted Single Phenotype (RSiP) class is used to model a value range restriction of an individual property (e.g., waist circumference > 102 cm). The value range is represented as an anonymous equivalent class based on the corresponding property restriction (Fig. 2: A1, A2). Boolean connections (e.g., gender = female AND waist circumference > 88 cm OR gender = male AND waist circumference > 102 cm) are expressed using Restricted Combined Phenotype (RCoP) classes and general class axioms (Fig. 2: B, B1).

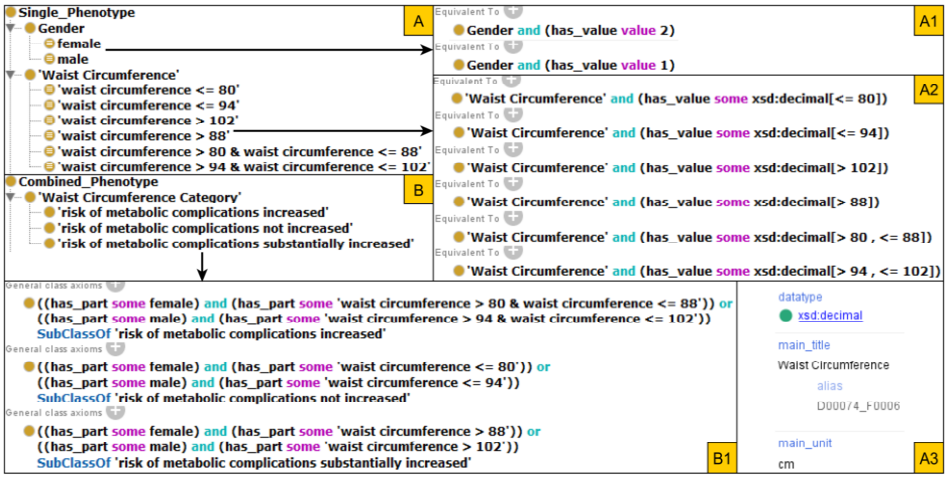

Figure 2. Phenotype Algorithm Specification Ontology (PASO) for waist circumference (Protégé screenshots)

The detailed structure of the ontology and our definition of the phenotype notion is presented in [8], whereas the current paper focuses on generating and executing SDQL queries based on the ontological phenotype specification.

4. Implementation

4.1. Generation of SDQL Queries by PhenoMan

The PhenoMan Core API is implemented (in the SMITH project [21]) in Java using the OWL API [22]. We developed an extension of PhenoMan to generate phenotype representations within the LHA platform including SDQL queries for the LHA Data Portal. This section sketches the main functionality of the query generator.

The PhenoMan interprets the underlying ontology (PASO) and creates different kinds of queries for specific phenotype classes using the query generator. The simplest queries are generated for ASiP classes (e.g., Gender and Waist Circumference). Such queries are only intended to request the number of study participants having a value (a database entry) of the corresponding property. The OID of the required ODM item (property) must be specified as an annotation of the corresponding class in the ontology (e.g., alias = D00074_F0006 for Waist Circumference, Fig. 2: A3). The resulting query is [[:item "D00074" "D00074_DEFAULT" "D00074_F0006"]], where "D00074" is the form name, "D00074_DEFAULT" is the item group name and "D00074_F0006" is the item name (Fig. 1 b: <Item>).

The value range queries are generated from RSiP classes and their property restrictions (Fig. 2: A, A1, A2). The query [[:item "D00074" "D00074_DEFAULT" "D00074_F0006" [:> 102]]] (Fig. 1 b: <Item> with <Predicate>), for example, returns the number of study participants with waist circumference greater than 102 cm.

Finally, the queries for Boolean expressions are created based on general class axioms describing the RCoP classes (Fig. 2: B, B1). The Boolean queries combine the individual value range queries by suitable operators (:or, :and). According to the SDQL specification, Boolean queries (disjunctions and conjunctions) must have a name followed by an operator and the corresponding operands (Fig. 1 b: <NamedExpression>). The PhenoMan names the Boolean queries automatically ("AND_1", "AND_2", "OR_1", "OR_2", etc.). The complete query to request the number of study participants with substantially increased risk of metabolic complications is:

```
[["OR_1" :or
   ["AND_1" :and
      [:item "D00153" "D00153_DEFAULT" "D00153_GENDER" [:= 2]]
      [:item "D00074" "D00074_DEFAULT" "D00074_F0006" [:> 88]]]
   ["AND_2" :and
      [:item "D00153" "D00153_DEFAULT" "D00153_GENDER" [:= 1]]
      [:item   "D00074"   "D00074_DEFAULT"   "D00074_F0006"   [:>
102]]]]]
```

Ontological phenotype specifications including the generated SDQL queries are transferred to the LHA using a REST interface (see LHA extension). The presented example is available under: https://www.health-atlas.de/phenotype_algorithms/BMI_Waist_Hip. The links to the LHA Data Portal contain encoded queries as a GET parameter. When the user is logged in, the Data Portal returns the corresponding results.

4.2. LHA Extension of the SEEK Platform

We extended the SEEK platform by adding new content types for phenotype algorithms, phenotype groups and phenotypes with respective property fields (e.g., title, description, unit, formula, query). A phenotype algorithm corresponds to a complete PASO, whereas the content type 'phenotype' represents the individual phenotype classes. Phenotype groups structure the phenotype algorithms, such that closely related phenotype classes share the same group. Entries of the described content types can be technically created, updated and deleted using the REST interface of SEEK, which is based on the OpenAPI standard [23]. A user with administrative privileges is allowed to perform respective post, patch and delete requests with the phenotypic properties as JSON data. After creation of

a phenotype algorithm with its components, the properties of the algorithm are browsable and referenceable within the LHA platform. The queries are integrated in the links to the LHA Data Portal and can be executed by clicking the links.

5. Lessons Learned (Discussion)

Our objective was to combine three existing systems: the PhenoMan (to model phenotypes and to generate queries), the LHA platform (to represent the phenotypes including queries as links to the Data Portal) and the LHA Data Portal (to execute the generated queries). All requirements for the overall solution were fulfilled. We succeeded in integrating structured phenotype specifications in the LHA with the possibility to query underlying data. However, our solution has some limitations. The use of a SEEK-based platform including the phenotype representation extension and a SDQL-based data repository are mandatory preconditions for applying our approach.

The novel DSL developed in this work reuses the conceptual abstract ODM entities to utilize known and well-defined vocabularies. The CDISC ODM is an established standard for exchanging and archiving clinical and translational research data. The LHA Data Portal is based on ODM to enable efficient but flexible storing and providing the data in a standardized manner. This data portal allows to query data by user-specified filter criteria. Each query is internally translated and exchanged between web client and portal server using SDQL. The SDQL enables an ODM-compliant retrieval of research data and can be used also outside of the LHA.

In contrast to similar approaches, we consider modelling of eligibility criteria as one aspect in a general ontological phenotyping framework (PhenoMan) and a set of criteria as a specification of a phenotype class [8]. PhenoMan is already used to compute and derive various phenotypes (such as estimated body surface area and scores like SOFA and Glasgow Coma Scale) in the SMITH project and for querying clinical data based on FHIR Search [24]. Our ontological framework demonstrated its applicability also for the LHA.

In SMITH, FHIR Search is used to query the integrated patient data. The FHIR Search Framework provides a range of operations and parameters to search for existing FHIR resources in the underlying repository. A study by Gulden et al. [25] has shown the advantages and disadvantages of using FHIR Search for specifying eligibility criteria. The investigation of representing phenotypic queries in different languages, such as FHIR Search or CQL [26], is a fundamental part of our current and future work.

The evaluation of the LHA is work in progress but the first results look very promising.

6. Conclusion

In order to represent structured phenotype algorithms and to perform queries on a research data repository based on the CDISC ODM standard, we introduced a novel domain-specific query language and an ontology-based method for modelling phenotypes and generating phenotypic queries. Our approach was successfully applied to represent phenotype algorithms in the Leipzig Health Atlas with the possibility to execute phenotypic queries within the LHA Data Portal.

Conflict of Interest

The authors state that they have no conflict of interests.

Acknowledgment

This work was supported by the German Federal Ministry of Education and Research (LHA: 031L0026, SMITH: 01ZZ1803A).

References

[1] The Leipzig Health Atlas, (n.d.). https://www.health-atlas.de/.
[2] F.A. Meineke, M. Löbe, and S. Stäubert, Introducing Technical Aspects of Research Data Management in the Leipzig Health Atlas, *Stud Health Technol Inform*. **247** (2018) 426–430. doi:10.3233/978-1-61499-852-5-426.
[3] M.D. Wilkinson, M. Dumontier, Ij.J. Aalbersberg, et al., The FAIR Guiding Principles for scientific data management and stewardship, *Sci Data*. **3** (2016) 1–9. doi:10.1038/sdata.2016.18.
[4] SEEK | FAIRDOM, (n.d.). https://fair-dom.org/platform/seek/.
[5] ISA framework, *ISA Tools*. (n.d.). https://isa-tools.org/.
[6] LHA Data Portal, (n.d.). https://dp.health-atlas.de/.
[7] T. Kirsten, A. Kiel, J. Wagner, M. Rühle, and M. Löffler, Selecting, Packaging, and Granting Access for Sharing Study Data – Experiences from the LIFE Study, *Workshop Digitale Prozesse Und Informationssysteme Im Forschungsdatenmanagement, Jahreskongress INFORMATIK 2017*. **P275** (2017). doi:10.18420/in2017_138.
[8] A. Uciteli, C. Beger, T. Kirsten, F.A. Meineke, and H. Herre, Ontological Modelling and Reasoning of Phenotypes, in: Proceedings of the Joint Ontology Workshops (JOWO) 2019, Episode V: The Styrian Autumn of Ontology, Graz, Austria, 2019, CEUR Workshop Proceedings, Vol. 2518, 2019.
[9] A. Bucur, J. van Leeuwen, N.-Z. Chen, et al., Cohort Selection and Management Application Leveraging Standards-based Semantic Interoperability and a Groovy DSL, *AMIA Jt Summits Transl Sci Proc*. **2016** (2016) 25–32.
[10] P.N. Robinson, Deep phenotyping for precision medicine, *Hum. Mutat.* **33** (2012) 777–780. doi:10.1002/humu.22080.
[11] Y. Zhang, G. Zhang, and Q. Shang, Computer-Aided Clinical Trial Recruitment Based on Domain-Specific Language Translation: A Case Study of Retinopathy of Prematurity, *Journal of Healthcare Engineering*. **2017** (2017). doi:10.1155/2017/7862672.
[12] C. Weng, S.W. Tu, I. Sim, and R. Richesson, Formal representation of eligibility criteria: A literature review, *Journal of Biomedical Informatics*. **43** (2010) 451–467. doi:10.1016/j.jbi.2009.12.004.
[13] H. Zhang, Z. He, X. He, et al., Computable Eligibility Criteria through Ontology-driven Data Access: A Case Study of Hepatitis C Virus Trials, *AMIA Annu Symp Proc*. **2018** (2018) 1601–1610.
[14] D. Calvanese, B. Cogrel, S. Komla-Ebri, et al., Ontop: Answering SPARQL queries over relational databases, *Semantic Web*. **8** (2017) 471–487. doi:10.3233/SW-160217.
[15] S.W. Tu, M. Peleg, S. Carini, et al., A Practical Method for Transforming Free-Text Eligibility Criteria into Computable Criteria, *J Biomed Inform*. **44** (2011) 239–250. doi:10.1016/j.jbi.2010.09.007.
[16] E. Chondrogiannis, V. Andronikou, A. Tagaris, E. Karanastasis, T. Varvarigou, and M. Tsuji, A novel semantic representation for eligibility criteria in clinical trials, *Journal of Biomedical Informatics*. **69** (2017) 10–23. doi:10.1016/j.jbi.2017.03.013.
[17] CDISC ODM, *CDISC*. (n.d.). https://www.cdisc.org/standards/data-exchange/odm.
[18] A. Kiel, J. Wagner, M. Rühle, and A. Twrdik, Lens - The system behind the LIFE Data Portal, in: 15th Leipzig Research Festival for Life Sciences, Leipzig, 2019.
[19] J. Wagner, Softwaregestützte Bereitstellung von epidemiologischen Forschungsdaten, Master Thesis, Leipzig University of Applied Sciences, 2016.
[20] Waist circumference and waist-hip ratio: report of a WHO Expert Consultation, WHO, 2008. www.who.int/nutrition/publications/obesity/WHO_report_waistcircumference_and_waisthip_ratio/en/.
[21] A. Winter, S. Stäubert, D. Ammon, et al., Smart Medical Information Technology for Healthcare (SMITH), *Methods Inf Med*. **57** (2018) e92–e105. doi:10.3414/ME18-02-0004.
[22] OWL API, (n.d.). http://owlcs.github.io/owlapi/.

[23] OpenAPI, *OpenAPI Initiative Registry*. (n.d.). http://spec.openapis.org/.
[24] FHIR Search, (n.d.). https://www.hl7.org/fhir/search.html.
[25] C. Gulden, S. Mate, H.-U. Prokosch, and S. Kraus, Investigating the Capabilities of FHIR Search for Clinical Trial Phenotyping, *Studies in Health Technology and Informatics*. (2018) 3–7. doi:10.3233/978-1-61499-896-9-3.
[26] Clinical Quality Language (CQL), (n.d.). https://cql.hl7.org/.