*Informatics for Health: Connected Citizen-Led Wellness and Population Health*
*R. Randell et al. (Eds.)*

*doi:10.3233/978-1-61499-753-5-131*

# Linked Data Applications Through Ontology Based Data Access in Clinical Research

Ann-Kristin KOCK-SCHOPPENHAUER [a], Christian KAMANN [a, b],
Hannes ULRICH [a, b], Petra DUHM-HARBECK [a], Josef INGENERF [a, b, 1]
[a] *IT for Clinical Research, Lübeck (ITCR-L), University of Lübeck, Germany*
[b] *Institute of Medical Informatics, University of Lübeck, Germany*

**Abstract.** Clinical care and research data are widely dispersed in isolated systems based on heterogeneous data models. Biomedicine predominantly makes use of connected datasets based on the Semantic Web paradigm. Initiatives like Bio2RDF created Resource Description Framework (RDF) versions of Omics resources, enabling sophisticated Linked Data applications. In contrast, electronic healthcare records (EHR) data are generated and processed in diverse clinical subsystems within hospital information systems (HIS). Usually, each of them utilizes a relational database system with a different proprietary schema. Semantic integration and access to the data is hardly possible. This paper describes ways of using Ontology Based Data Access (OBDA) for bridging the semantic gap between existing raw data and user-oriented views supported by ontology-based queries. Based on mappings between entities of data schemas and ontologies data can be made available as materialized or virtualized RDF triples ready for querying and processing. Our experiments based on CentraXX for biobank and study management demonstrate the advantages of abstracting away from low level details and semantic mediation. Furthermore, it becomes clear that using a professional platform for Linked Data applications is recommended due to the inherent complexity, the inconvenience to confront end users with SPARQL, and scalability and performance issues.

**Keywords.** Semantic Web, Linked Data, Semantic Querying and Data Integration

## 1. Introduction

Within translational research there is a demand to semantically process and integrate clinical care and biomedical research data from different heterogeneous resources. Instead of schema matching approaches (e.g. for relational databases) the Semantic Web paradigm uses the Resource Description Format (RDF) for the flexible representation of facts together with a semantic layer for describing corresponding types and relationships by ontologies (RDFS, OWL). Efficient frameworks for distributed queries across multiple RDF data sources are used in many application areas; amongst others in biomedicine and less often in healthcare [2, 3]. In clinical care usually Relational Database Management Systems (RDBMS), i.e. mature products like Oracle, MS SQL or MySQL, are used.

Ontology-Based Data Access (OBDA) is a new paradigm for accessing and integrating data, whose key concept is to resort to a three-level architecture with an ontology,

[1] Corresponding author, Josef Ingenerf, Institut für Medizinische Informatik, Ratzeburgerallee 160, 23562 Lübeck, Germany; E-mail: ingenerf@imi.uni-luebeck.de.

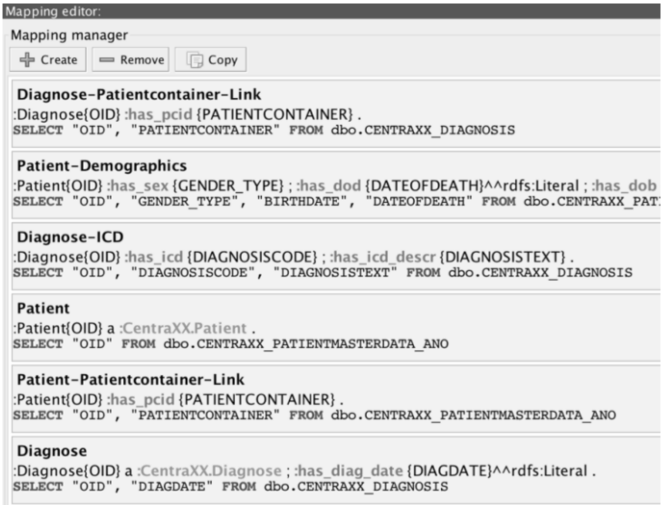
data sources, and mappings between both [4]. The ontology defines a high-level global schema of data sources and provides a vocabulary in terms of concepts, roles, i.e. binary relations and attributes for user queries. The mapping layer explicitly specifies the relationships between the domain concepts and the data sources. Afterwards an OBDA system rewrites such queries and ontologies into the vocabulary of the data sources and delegates the actual query evaluation to a suitable query answering system such as SQL for RDBMS. The ontology together with the mappings exposes a virtual RDF graph, which can be queried using SPARQL, the standard query language for RDF data. This virtual RDF graph can be materialized by using RDF triplestores, or alternatively it can be kept virtual and queried only during query execution.

## 2. Methods and Material

In the following we present several ways to adopt the OBDA approach by accessing data from the RDBMS-based CentraXX system for biobank and study management [5].

### 2.1. Ontop Used As a Plugin within Protégé

Ontop is one of the most popular OBDA systems [6]. This open source software is available amongst others as plugin for the ontology editor Protégé. First, a domain ontology with relevant concepts like patients, encounters and diagnoses and their relationships are defined. Second, original data sources are connected and mappings are managed. The mapping includes how classes of instances and relationships are mapped to the database entries by SQL statements, see Fig 1. Finally, SPARQL queries are created and executed by *Quest*, a query answering engine with OWL 2 QL/RDFS entailment.



```
Diagnose-Patientcontainer-Link
:Diagnose{OID} :has_pcid {PATIENTCONTAINER} .
SELECT "OID", "PATIENTCONTAINER" FROM dbo.CENTRAXX_DIAGNOSIS

Patient-Demographics
:Patient{OID} :has_sex {GENDER_TYPE} ; :has_dod {DATEOFDEATH}^^rdfs:Literal ; :has_dob
SELECT "OID", "GENDER_TYPE", "BIRTHDATE", "DATEOFDEATH" FROM dbo.CENTRAXX_PAT

Diagnose-ICD
:Diagnose{OID} :has_icd {DIAGNOSISCODE} ; :has_icd_descr {DIAGNOSISTEXT} .
SELECT "OID", "DIAGNOSISCODE", "DIAGNOSISTEXT" FROM dbo.CENTRAXX_DIAGNOSIS

Patient
:Patient{OID} a :CentraXX.Patient .
SELECT "OID" FROM dbo.CENTRAXX_PATIENTMASTERDATA_ANO

Patient-Patientcontainer-Link
:Patient{OID} :has_pcid {PATIENTCONTAINER} .
SELECT "OID", "PATIENTCONTAINER" FROM dbo.CENTRAXX_PATIENTMASTERDATA_ANO

Diagnose
:Diagnose{OID} a :CentraXX.Diagnose ; :has_diag_date {DIAGDATE}^^rdfs:Literal .
SELECT "OID", "DIAGDATE" FROM dbo.CENTRAXX_DIAGNOSIS
```

**Figure 1.** Realizing OBDA with the Ontop plugin within Protégé, applied to the CentraXX database.

The mappings expressed in W3C standard R2RML (RDB to RDF Mapping) can be constructed semi-automatically by domain experts or by using a bootstrapper that creates the ontology and mappings automatically by analyzing the database schema. Furthermore, Ontop works with Teiid as open source Java software for data virtualization, used for federating different heterogeneous RDBMS behind one JDBC interface. The use of Ontop as plugin in Protégé was not fully sufficient, for mainly two reasons: First, the end users of the envisioned OBDA-based query system should not be forced to enter SPARQL queries. Second, there is a limitation when trying to follow up with linked data applications based on the resulting RDF triples.

### *2.2. Optique Platform with OptiqueVQS as a Visual Query System*

Optique (Scalable End-user Access to Big Data) is an EU-funded project where novel solutions based on the OBDA idea have been developed [7]. The Optique Platform has been made available as an app that can be installed and deployed within the Information Workbench, see chapter 2.3. The platform features a visual query system (VQS) where query dialogues are rendered based on the ontology, see Fig 2. However, in spite of the impressive potential of the VQS, the flexibility with regard to the desired query frontend was not sufficient, since this can mainly be influenced by cumbersome modifications of the ontology.

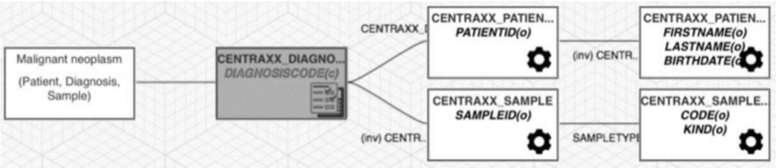


**Figure 2**. Screenshot of the query interface generated by OptiqueVQS.

### *2.3. Information Workbench (IWB) - A Platform for Linked Data Applications*

For flexibility reasons we decided to work directly with the Information Workbench (IWB) [8]. The software provides a generic frontend for customizable user interfaces based on Semantic Wiki technologies, enriched with a large set of widgets for data access, navigation, visualization, analytics and data mashups with external data sources [9]. It can be further customized and extended for domain specific applications through a SDK. Techniques for OBDA from chapter 2.1 and 2.2 like Ontop are included in the platform.

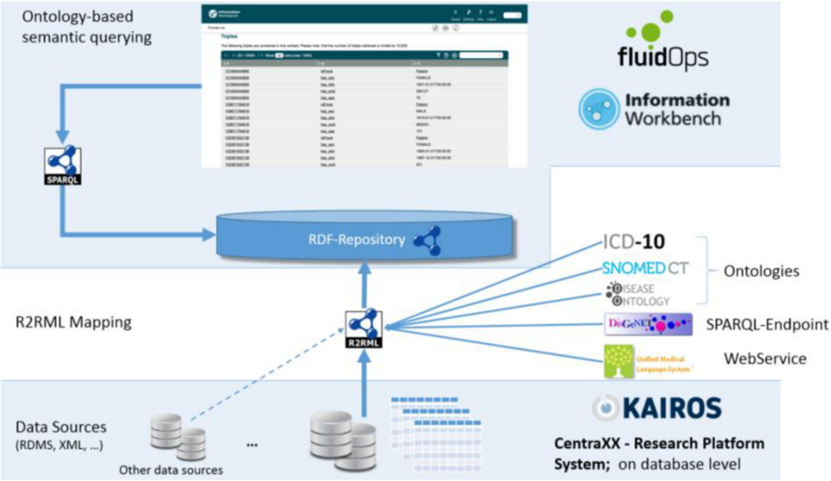
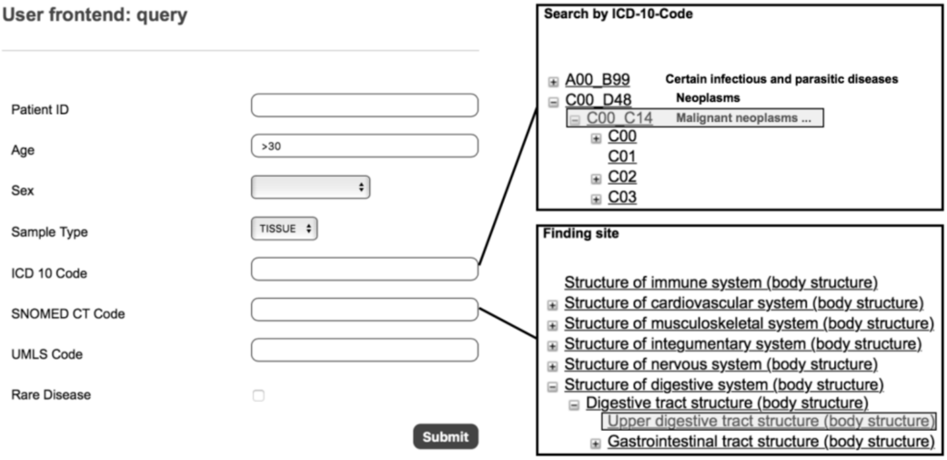


**Figure 3.** Information Workbench for linked data applications based on relational data from CentraXX.

For federated queries, i.e. sending decomposed SPARQL subqueries to various data services and integrating the results virtually, the FedX module is available [10]. The Information Workbench is available as a Community Edition under an open source license as well as an Enterprise Edition with a commercial license. We used IWB for creating a demonstrator that provides an ontology based query frontend for accessing RDBMS data of the CentraXX system enriched by linked data, see Fig 3.

## 3. Results and Discussion

The IWB greatly facilitates the creation of an ontology with concepts, relationships and attributes of interest and relevant R2RML-mappings to the relational database used for example by CentraXX. This allows to materialize corresponding RDF triples internally.



User frontend: query

Patient ID:
Age: >30
Sex:
Sample Type: TISSUE
ICD 10 Code:
SNOMED CT Code:
UMLS Code:
Rare Disease: □

Search by ICD-10-Code

- ⊞ A00_B99 Certain infectious and parasitic diseases
- ⊟ C00_D48 Neoplasms
  - ⊟ C00_C14 Malignant neoplasms ...
    - ⊞ C00
    - C01
    - ⊞ C02
    - ⊞ C03

Finding site

- Structure of immune system (body structure)
- ⊞ Structure of cardiovascular system (body structure)
- ⊞ Structure of musculoskeletal system (body structure)
- ⊞ Structure of integumentary system (body structure)
- ⊞ Structure of nervous system (body structure)
- ⊟ Structure of digestive system (body structure)
  - ⊟ Digestive tract structure (body structure)
    - Upper digestive tract structure (body structure)
    - ⊞ Gastrointestinal tract structure (body structure)

**Figure 4.** User frontend and technical details of the demonstrator using the IWB

By using further ontologies like the Disease Ontology [11] or the Unified Medical Language System (UMLS) [12] the triples could be semantically enriched by ontological mappings, especially taking the ICD-10 codes within the CentraXX data into consideration. By extracting finding sites from mapped disorder concepts in SNOMED CT and adding this anatomical codes to the RDF repository. For example, it becomes possible to query all data from patients or samples that are located at the digestive tract. An example for such a query is shown in Fig 4. Compared to ICD-10 based retrieval this is an example for the kind of added values that should be further explored. Additionally, it is possible to use annotated Disease Ontology or UMLS codes for accessing linked data of interest like disease associated genes from the SPARQL endpoint of DisGenNET [13]. There are much more linked data of interest that we might include similarly, e.g. accessing literature from MEDLINE.

On the userinterface, the customization is done in a completely declarative way, resorting to a rich pool of widgets and creating template pages in wiki syntax, which are associated with elements of domain ontologies. In our preliminary experiments this significantly simplifies and speeds up the application development. The OBDA paradigm is a promising approach within clinical research informatics because a lot of existing applications are based on relational database systems.

## Acknowledgement

The authors would like to thank fluid Operations AG and KAIROS GmbH for their support. The approach presented in this paper is implemented on a CentraXX test environment, but can be easily transferred to the system in use as well as to many other EHR systems of interest.

## References

[1] Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J: *Bio2RDF: towards a mashup to build bioinformatics knowledge systems*. J Biomed Inform. 2008; 41(5):706-16.

[2] Carmen Legaz-Garcia MD, Minarro-Gimenez JA, Menarguez-Tortosa M, Fernandez-Breis JT: *Generation of open biomedical datasets through ontology-driven transformation and integration processes*. J Biomed Semantics. 2016; 7:32. 10.1186/s13326-016-0075-z.

[3] Hussain S, Ouagne D, Sadou E, Dart T, Jaulent MC, De Vloed B, Colaert D, Daniel C: *EHR4CR: A Semantic Web Based Interoperability Approach for Reusing Electronic Healthcare Records in Protocol Feasibility Studies*. In: Proceedings of the 5th International Workshop on Semantic Web Applications and Tools for Life Sciences; 2012. p. http://ceur-ws.org/Vol-952/paper_31.pdf.

[4] Liaw ST, Taggart J, Yu H, de Lusignan S, Kuziemsky C, Hayen A: *Integrating electronic health record information to support integrated care: practical application of ontologies to improve the accuracy of diabetes disease registers*. J Biomed Inform. 2014; 52:364-72. 10.1016/j.jbi.2014.07.016

[5] CentraXX - KAIROS GmbH, [Internet, cited 19 October 2016], Available from: http://www.kairos.de/centraxx/.

[6] Calvanese D, Cogrel B, Komla-Ebri S, Kontchakov R, Lanti D, Rezk M, Rodriguez-Muro M, Xiao G: *Ontop: Answering SPARQL queries over relational databases*. Semantic Web. 2016:1-17 [in press].

[7] Giese M, Soylu A, Vega-Gorgojo G, Waaler A, Haase P, Jiménez-Ruiz E, Lanti D, Rezk M, Xiao G, Özccep ÖL, Rosati R: *Optique: Zooming in on Big Data*. IEEE Computer. 2015; 48(3):60-67.

[8] Information Workbench - Fluid Operations AG, [Internet, cited 19 October 2016], Available from: https://www.fluidops.com/en/products/information_workbench/.

[9] Gossen A, Haase P, Hütter C, Meier M, Nikolov A, Pinkel C, Schmidt M, Schwarte A: *The Information Workbench - A Platform for Linked Data Applications*. Semantic Web 2016:1-7 [in press].

[10] Schwarte A, Haase P, Hose K, Schenkel R, Schmidt M: *FedX: Optimization Techniques for Federated Query Processing on Linked Data*. In: The Semantic Web - ISWC 2011 - 10th Internat. Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part I: Springer; 2011. p. 601-616.

[11] Schriml LM, Arze C, Nadendla S, Chang YW, Mazaitis M, Felix V, Feng G, Kibbe WA: *Disease Ontology: a backbone for disease semantic integration*. Nucleic Acids Res. 2012; 40 (Database issue): D940-6. 10.1093/nar/gkr972

[12] Bodenreider O: *Biomedical ontologies in action: role in knowledge management, data integration and decision support*. Yearb Med Inform. 2008:67-79.

[13] Queralt-Rosinach N, Pinero J, Bravo A, Sanz F, Furlong LI: *DisGeNET-RDF: harnessing the innovative power of the Semantic Web to explore the genetic basis of diseases*. Bioinformatics. 2016; 32(14):2236-8. 10.1093/bioinformatics/btw214