World Academy of Science, Engineering and Technology
International Journal of Medical and Health Sciences
Vol:1, No:7, 2007

The Resource Description Framework (RDF) as a Modern Structure for Medical Data

Gabriela Lindemann, Danilo Schmidt, Thomas Schrader, and Dietmar Keune

***Abstract*—**The amount and heterogeneity of data in biomedical research, notably in interdisciplinary fields, requires new methods for the collection, presentation and analysis of information. Important data from laboratory experiments as well as patient trials are available but come out of distributed resources. The Charité - University Hospital Berlin has established together with the German Research Foundation (DFG) a new information service centre for kidney diseases and transplantation (Open European Nephrology Science Centre - OpEN.SC). Beside a collaborative aspect to create new research groups every single partner or institution of this science information centre making his own data available is allowed to search the whole data pool of the various involved centres. A core task is the implementation of a non-restricting open data structure for the various different data sources. We decided to use a modern RDF model and in a first phase transformed original data coming from the web-based Electronic Patient Record database TBase©.

***Keywords*—**Medical databases, Resource Description Framework (RDF), metadata repository.

I. Introduction

TYPICAL problems of research in medical disciplines are apparent: heterogeneous data types with varying quality of automatically and "by hand" generated data, a limitation of the amount of cases of rare diseases and decentralized data retention inhibit the scientific work. In contrast to the efforts of different standardization organizations (HL7, OpenEHR, IEEE, ISO, CEN, …) no common accepted integrative standard exist.

To fill in this gap the German Research Foundation (DFG) established a new priority program "Centres of Research Information" in 2006, not only dedicated to medical data but of crucial importance especially in this field.

G. Lindemann holds a senior researchers position at the institute of computer science of Humboldt University of Berlin. Humboldt University of Berlin, Institute of Computer Science, visitors: Rudower Chausse 25, Unter den Linden 6, 10099 Berlin, Germany (e-mail: lindeman@informatik.hu-berlin.de, http://www2.informatik.hu-berlin.de/~lindeman).

D. Schmidt holds a research position at the Lab of Artificial Intelligence of the Institute of Computer Science of Humboldt University of Berlin, Unter den Linden 6, 10099 Berlin, Germany (e-mail: danilo.schmidt@charite.de).

T. Schrader holds a research position at the Department of Pathology, Charite Universitätsmedizin Berlin, Chariteplatz 1, 10117 Berlin, Germany (e-mail: thomas.schrader@charite.de).

D. Keune holds a research position at the Department of Pathology, Charite Universitätsmedizin Berlin, Chariteplatz 1, 10117 Berlin, Germany (e-mail: dietmar.keune@charite.de).

Within this framework an interdisciplinary group consisting of nephrologists and pathologists from the Charité – University Hospital Berlin and computer scientists from the Lab of Artificial Intelligence of the Institute of Computer Sciences of Humboldt University of Berlin found together and start the OpEN.SC project, which is now granted by the DFG.

The main task of OpEN.SC is the implementation of a metadata repository for clinical data, data of studies, literature and virtual slides. The system works as a data centre to improve the availability of standardized raw data including digitalized glass slides (Virtual Microscopy) and works at first as an interface for scientific publications [3]. Later on it can easily extended in several directions, e.g. as to serve as a part of an e-learning system for young physicians in the field [7].

OpEN.SC comprises different working packages. We developed a structured system architecture based on a Service Oriented Architecture (SOA) proposed by [1] which is designed with the help of BPEL a Business Process Execution Language. BPEL depicts the workflows of the scientific workplace connected to OpEN.SC and yields templates for the web services which are the core software modules of the system.

As presented in several papers before [2, 6] the meanwhile in 18 German transplant centres used web-based Electronic Patient Record database TBase© is our first data source. TBase© is based on a well-known "normal" relational database structure, what means, that it has fixed patterns of entities and their relations, see Fig. 1, what shows a part of this relational structure – only to get an impression. In OpEN.SC we integrate data sources of several involved partner clinics wherein the process of data messaging the private patient data were anonymoused.

Every of the clinical partners has their own data formats and structures. Therefore it was important to construct a database model which is independent from these resources. We decided to use the very flexible and open Resource Description Framework (RDF) which leads to a flat database structure consisting only of triples.

Fig. 1 Part of the TBase© data structure

II. Resource Description Framework (RDF)

RDF - the Resource Description Framework - is a model and language based on XML for representing information about resources of different types. It is particularly useful for storing metadata about shared resources.

Due to the W3C consortium: "RDF is intended for situations in which this information needs to be processed by applications, rather than being only displayed to people. RDF provides a common framework for expressing this information so it can be exchanged between applications without loss of meaning. Since it is a common framework, application designers can leverage the availability of common RDF parsers and processing tools. The ability to exchange information between different applications means that the information may be made available to applications other than those for which it was originally created." [4].

The design of RDF is generally intended to meet the following goals:

- having a simple data model
- having formal semantics and provable inference
- using an extensible URI (Uniform Resource Identifiers)-based vocabulary
- using an XML-based syntax and XML schema data types
- allowing anyone to make statements about any resource

see [5].

Expressions in RDF are represented as triples in general consisting of a subject, a predicate (also called a property) that denotes a relationship to an object.

A usual representation of the property that holds between subject and object could be as a row in a table in a relational database. Then the table has two columns, corresponding to the subject and the object of the RDF triple. The name of the table corresponds to the predicate of the RDF triple. Relational databases has an arbitrary number of columns and a row expresses a 1:n relationship between entities. Such a row, or relation, has to be decomposed for representation in a RDF triple structure, what we will describe in detail in the next chapter.

III. RDF Data Structure in OpEN.SC

In discussion with our medical partners we divided the complete TBase© domain and additional available data e.g. coming from the VMscope (virtual slides) in seven sub-domains according to considerations about a logical arrangement of medical and further data of patients. These are:

- Examination Data
- Diagnostic Data
- Treatment Data
- Basic Data (of a patient)
- Administrative Data (e.g. case number in a hospital, SAP data)
- External Data (e.g. virtual slides, discharge letters)
- Project Data (internal)

where the main medical data are in the first three domains.

The transformation of original TBase© data into the RDF structure of the OpEN.SC database is done via a request from OpEN.SC side out. Then a web service is initiated which compiles the TBase© data in the new structure. For our purposes and clearer understanding we re-named the general RDF-terms subject, predicate and object in resource, property and value. Let us give an example:

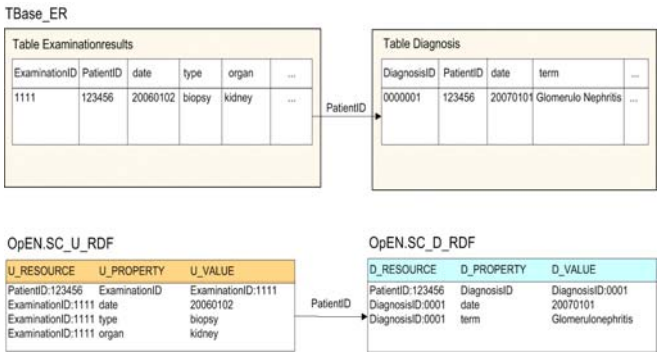

Fig. 2 Transformation of TBase© data into the OpEN.SC RDF structure

In Fig. 2 we demonstrate a short example of the transformation of fictive data records of TBase© where the internal entity relationship between the table *Examinationresults* and the table *Diagnosis* is depicted by the link PatientID.

The PatientID yields always the connection between the datasets belonging to one patient and distributed in the different sub-domains. In the original relational TBase©

database rows can have a large number of attributes – here outlined as three points.

Each of such a single attribute leads to a single row in the RDF structure. So the RDF structure increases in the depth in opposite to classical relational structures which grow in breadth when including additional features.

When transforming the relational TBase© structure into the RDF structure the next new row in the RDF-base contains the PatientID as resource semantically interpretable as mark for the beginning of a new set of data for this patient. In our instance we show in the left-hand side the transformation of a TBase© row concerning one examination – a biopsy of the kidney at the 1. February 2006 - with ID 1111. Examinations belong in our model to the U_Domain and consequently their datasets were stored in the OpEN.SC_U_RDF table. The following rows in the RDF-base have as resource the ExaminationID, as property the name of the succeeding columns in TBase© and as value the corresponding real attribute value.

The same procedure is done when transforming a row of the TBase© table Diagnosis into columns of the OpEN.SC_D_RDF database belonging to the diagnosis domain D_Domain.

In general, the identifier in sub-domains – these are primary keys in the relational structure – become the content of resource, each heading of a column becomes a property and each entry in the table denotes a value in the RDF structure.

Although it seems at a first look that the RDF structure is larger than that of the TBase© there is only a marginal difference in the memory space. At the moment the complete transformation of the whole TBase© data of Charité – Campus Mitte with the medical records of about 2500 patients comprises about 17.000.000 RDF triples. Moreover, the performance of the RDF database is high. A request to the basic data of a patient lasts 4 milliseconds, a complex request with respect to a special examination and a diagnostic course of a patient lasts 13 seconds. As hardware we use a server with Intel®Xenon®CPU, 5130 @ 2.00 GHz and 2,00 GB RAM with operating system Windows Server 2003 and the ORACLE 10g (10.2.0.1.0) database environment.

IV. Outlook

In the development of the Open European Nephrology Science Center a basic and essential task is the design and implementation of the underlying database. Because of the planned inclusion of the different sources of involved and future partner hospitals the data structure of our system must represent information in a minimally constraining, flexible way. For that reason we decided to use the RDF model. First experiments with respect to the performance of the OpEN.SC RDF database are encouraging. Furthermore, we implemented a web service module what allows to transform data from a "normal" relational database into the RDF format.

In the near future data from the External Data domain e.g. the virtual slides coming from VMscope or discharge letters from the physicians in free text will be included in the RDF database by a value that is a link to the original source. Beside this, we will perform a series of "stress-tests" with our RDF model.

Acknowledgment

This work is partly granted by the Deutsche Forschungsgemeinschaft (DFG) under the main point program "Centers of Research Information". OpEN.SC is a collaboration of the Department of Computer Science - Artificial Intelligence, the Institute of Pathology, the Institute of Medical Informatics and the four Departments of Nephrology of the Charité in Berlin as well as associated partners from abroad.

References

[1] D. Krafzig, K. Banke, D. Slama. "Enterprise SOA. Service Oriented Architecture Best Practices." *Prentice Hall PTR*, 2004.

[2] G. Lindemann, L. Fritsche. "Web-Based Patient Records - The Design of TBase2." *In Bruch*, Köckerling, Bouchard, Schug-Paß (Eds.) New Aspects of High Technology in Medicine; Seiten 409-414; 2000.

[3] G. Lindemann, T. Schrader, D. Schmidt. "The Aim of the Open European Nephrology Science Center (OpEN.SC) – First Steps." *In: Proceedings of "Concurrency, Specification & Programming" – CS&P2006*; Vol. I, II, III. Informatik Berichte 206, Institute of Computer Science, Humboldt University Berlin, ISSN: 0863-95X, Germany, 2006.

[4] F. Manola, E. Miller: http://www.w3.org/TR/2004/REC-rdf-primer-20040210/ , visited June 2007.

[5] B. McBride: http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/ , visited June 2007.

[6] E. Paslaru Bontas, S.Tietz, R.Tolksdorf, and T.Schrader. "Generation and Management of a Medical Ontology in a Semantic Web Retrieval System." [3290 / 2004]. *Lecture Notes in Computer Science*. R.Meersmann and Z.Tari, 2004.

[7] A.-B. M. Salem. "Intelligent Technologies for Medical e-Learning." In: Proceedings of the Third International Conference on Intelligent Computing and Information Systems. Cairo, Egypt, 2007

[8] T.Schrader, S.Niepage, T.Leuthold, K.Saeger, S.Hellmig, and P.Hufnagl. "Implementation of a pathology report compiler with integrated diagnostic path functionality in the Diagnostic Virtual Microscopy." *Pathol Res Pract* 200[4], 356. 2004.

[9] T.Schrader, P.Tietz, C.Tennstedt, M.Schwabe, T.-N.Nguyen-Dobinsky and P.Hufnagl. "Attributing images of interdisciplinary medical examinations in databases." *E J Pathol* 032 032-006. 2003.

[10] K. Schröter, G. Lindemann, L. Fritsche: "TBase2 – A web-based Electronic Patient Record." *Fundamenta Informaticae* 43, 343-353, IOS Press, Amsterdam, 2000.

Gabriela Lindemann has studied mathematics at Humboldt University of Berlin and received her diploma and PhD degree there in the field of applying discrete optimization methods in VLSI circuit design. With the foundation of the Institute of Computer Science of Humboldt University she entered a research position at the Lab of Artificial Intelligence there. Since more than 15 years she has been working in several applications of AI in medical informatics.

During the last years Dr. Lindemann has published in the fields of Multi-Agent Systems, Case-Based Reasoning and medical computer science. Moreover she was co-organizer and has been a member of programming committees of a lot of international conferences. She is a member of the academic senat of Humboldt University of Berlin.

Danilo Schmidt studied computer science at Humboldt University of Berlin. In 2006 he got his Diploma degree. With the beginning of the year 2007 Dipl.-Inf. Danilo Schmidt starts his business career within the project OpEN.SC. Within this project he is especially responsible for the part of the specification and implementation of the "Intelligent Catalogue", a tool which allows a fast access to the meta-data repository and what is based on principles of Case-Based Reasoning coming from Artificial Intelligence.

Thomas Schrader is a board certified pathologist and a computer scientist in the Department of Pathology at the Charite Medical School in Berlin, Germany. He is the program director of the Open European Nephrology Science Center and of various other projects in Telemedicine including in Virtual Microcopy.

Dietmar Keune is engineer for data processing. He has studied at the High-School for Engineering in Dresden (later included in the Technical University of Dresden, Germany). In Dipl.-Ing. Keune's business life he has been working in several jobs as programmer, project manager, network administrator, scientist and lecturer. In the OpEN.SC project he is the specialist for data-base specification, implementation, management and maintenance (e-mail: dietmar.keune@charite.de).