ONTOLOGY-BASED SYSTEM FOR CLINICAL TRIAL DATA MANAGEMENT

Sandra Geisler[1,2], Andreas Brauers[1], Christoph Quix[2], Anke Schmeink[1]

[1]Philips Research Europe, PO 500145, 52085 Aachen, Germany

[2]Informatik 5, RWTH Aachen University, Germany

Abstract

We developed an ontology-based system for the design and integration of clinical trial data management in a convenient and flexible way. A reference ontology serves as basis for both, the generation of clinical trial databases and the integration of data from various data sources into this database. We evaluated the usability of the system using the test case of a specific pilot trial on medical devices.

1 Introduction

Companies in the field of medical technology are concerned with the development of novel medical devices, which contribute to the development and improvement of diagnostics, therapy, prevention and monitoring of diverse diseases. For this purpose clinical trials and preliminary pilot trials are conducted. During the trials, data of interest is acquired in various formats by means of the developed devices, by questionnaires, case report forms and more. Afterwards the gathered data is processed and statistically analyzed.

Commercial data management systems for clinical trials are very expensive and provide a plethora of functionalities too comprehensive for smaller pilot trials [9]. Most of the existing systems are based on Entity-Attribute-Value (EAV) database design. This prohibits efficient database management through, e.g., indexing, partitioning, or query optimization and hamper data analysis and ad-hoc querying. Therefore, the proposed solution uses ontologies for the efficient, flexible and user-friendly design of the database and the integration process for the clinical trial data.

Often clinical trials are conducted by multidisciplinary teams. The ontology based approach provides a common basis on a high abstraction level, which allows users to design the databases and integration processes conceptually independent of physical design issues such as indexes and keys.

Furthermore, available systems often do not offer the possibility to design and implement data import from different sources, which is a crucial requirement especially for trials conducted with novel medical devices. The proposed system will offer the functionality to design the integration of various data sources also in a flexible and ontology-based way.

2 Approach

The proposed system is based on ontologies. According to Gruber in [1] an ontology is a "formal, explicit specification of a shared conceptualization". It describes a domain of interest in a machine readable and semantic way, i.e., the concepts of the domain, relationships among them and constraints can be expressed and mostly a larger community agrees upon it.

Ontologies are used in the fields of artificial intelligence, knowledge engineering and the Semantic Web and are developed for variety of domains, including biomedicine or physics.

The modeling and use of ontologies in the field of database creation and data integration offers several benefits. Ontologies describe domains on a high abstraction level and can therefore be easily understood by and discussed with domain experts without detailed database knowledge. The strengths of ontologies lie especially in the possibility to build a consistent and formal vocabulary, which cannot only be used for the definition of the structure and meaning of data stored in a database, but also be reused, to interoperate with and build applications based on this vocabulary [3]. Additionally, the ontology language used in this work, OWL, is standardized by the World Wide Web Consortium and offers a high level of expressivity.

The Clinical Trial Data Management Ontology (short CTDMO), developed by the authors, serves as a basis to create databases for clinical trials in a convenient and flexible way and to define data import modules based on that ontology. The overall system architecture and the workflow within the system are depicted in Figure 1.

For a new trial the users firstly extend the CTDM ontology with concepts, properties and individuals not already included in the model. This is done in an external OWL ontology editor, like Protégé, developed by the Stanford University School of Medicine (see [5]). In a later step, the user has to create an ontology for every kind of data source which will be imported into the planned database. During this step,

---

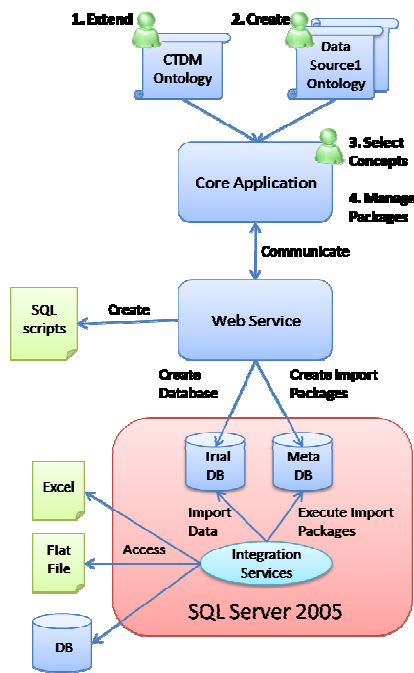the user will be assisted by the core application of the system.

Figure 1: Design Workflow

Afterwards, (Figure 1, Step 3) the user loads the CTDM ontology into the user interface of the core application. The user interface enables the user to select the concepts from the ontology he wishes to be represented in the prospective trial database, i.e., describing the data, which will be stored.

The application presents a preview of the database schema to the user. More database experienced users can also define views for the database, which will be included into the database creation process. The creation of the database can then either be executed directly by connecting to the respective server or the user can save SQL scripts which include the code to create the database by executing the scripts.

In a fourth step, the user can define data integration modules for each data source from which he wants to import data into the new database. These are described by additional ontologies which then have to be mapped by the user in the application with the concepts and properties he selected from the CTDM ontology. Based on the mapping and some additional information the user provided to the system, import modules for each data source are created.

To balance load, make the architecture more flexible and encapsulate database access, the core application communicates with a web service, which is responsible for creating the databases and data integration modules. The architecture in this project is based on SQL Server 2005. It hosts and manages the created databases and stores the import pack-

ages. The import packages are based on the Integration Services technology of SQL Server and executed by the respective service on the server.

2.1 The CDTM Ontology

The proposed system is based on a reference ontology called CDTMO, which was developed by the authors. Common requirements for an ontology to be suitable as a reference ontology for the system have been defined in the project:

1. Extendable, to easily add concepts and properties describing data of new studies

2. Concise and user-friendly structure to enable fast orientation

3. Modular design (concepts can be reused to describe other concepts)

4. High quality regarding the consistency, documentation and comprehensibility

Furthermore, interviews with domain experts, reviews of literature, regulations and proposed data exchange standards have been used for knowledge acquisition and to find out, which basic concepts have to be included in the ontology. Based on these requirements, existing ontologies describing clinical trials or containing knowledge about this domain have been assessed.

The outcome of the analysis showed that none of the examined ontologies could fulfill the requirements completely. Hence, the CTDM ontology was developed from scratch, using a top-down approach, i.e., starting at a very high abstraction level and stopping at a level where the users can start to extend the ontology with concepts and properties specific for their pilot trial.

In Figure 2 the root concepts of the CTDM ontology are shown. The modularity of the ontology has been achieved by separating single information objects, like "Weight" or "Gender", from concepts which make use of these information, like "Measurement". Crucial concepts which are likely to be extended by creating subconcepts for new studies, in the project called "Hot Spots", are for example "Medical Device", "Measurement", "Calculation" or "Subject".

The practicability of the ontology is evaluated by domain experts extending it for a concrete pilot study (see section 3).

Two different approaches to evaluate the usability are made. Firstly, the domain experts read a guideline, describing the ontology and giving a short introduction to ontology basics. Afterwards, the experts try to extend the ontology on their own. In a second test a short oral introduction to ontologies and the CDTMO is given and the domain experts extend the ontology afterwards in recurring modeling sessions.
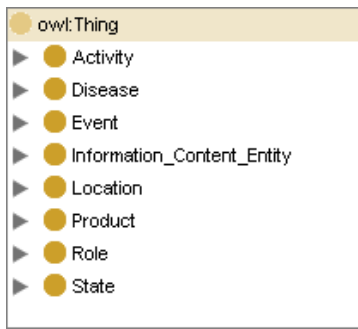
Figure 2: The CTDMO Root Concepts

After each modeling session the experts are interviewed. A questionnaire has been developed based on ontology quality criteria proposed in [2].

2.2 Ontology-based Database Schema Creation

There have been several approaches to create databases based on ontologies. As mentioned in the previous sections, ontologies have some crucial advantages in comparison to other semantic modeling techniques used for conceptual database design, including the ability to share and reuse them. In the design and creation of relational databases based on ontologies, three categories of approaches can be distinguished according to the degree of automating the process of database schema creation. These approaches also differ according to the used steps. Some approaches directly transfer an ontology into a database schema, e.g., in [4], whereas others map the ontology first to an ER-Model and then generate the relational database schema based on this model (e.g. in [6]).

We decided to use a direct and semi-automatic approach based on [7]. The CTDM ontology is loaded into a user interface and the user can select the concepts he wants to be represented in the new trial database.

The selection enables the user to use a comprehensive and semantic complete ontology and at the same time only include the concepts in the database he needs.

The following approach to convert the OWL ontology into a relational schema is used:

1. Conversion of the concept hierarchy: each selected concept is converted to a table. If a superconcept was selected, two tables are created which have the same primary key.

2. Conversion of datatype properties: each property is converted into an attribute of the respective table.

3. Conversion of object properties: depending on the constraints the user defined for the object property, the property is converted to a respective relationship. For 1:n-relationships a foreign key attribute is inserted and the corresponding constraint created. For n:m-relationships an intermediate table is created and its columns represent foreign keys to the related tables. For 1:1-relationships the attributes of the referenced concept are added to the table of the concept representing the domain of the property.

4. Individuals are converted to data rows, which are inserted into the new database.

In the Core Application the user is presented a preview of the new database based on the conversion of the selected concepts (see Figure 3).
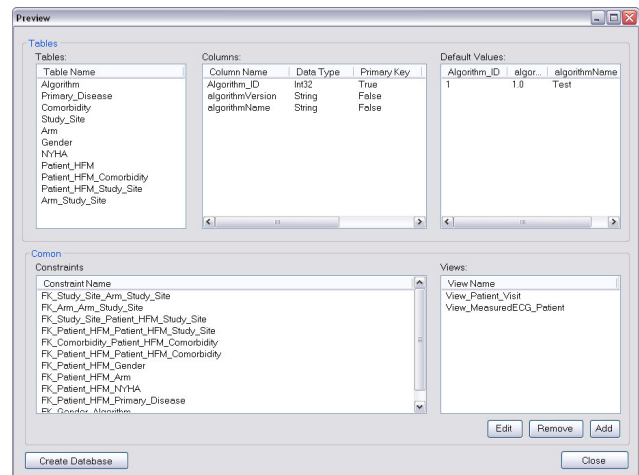

Figure 3: The Database Preview

2.3 Ontology-based Data Integration

To make the trial data accessible at one central point for statistical analysis or preparation for the submission to respective authorities, it needs to be integrated ideally into a single database. Approaches in data integration can be distinguished according to different aspects. One aspect is the manifestation of integration of the data sources. This can be described with the terms On-Demand Integration, where data sources are integrated just when a user or system queries a framework of data sources. The requested data is then acquired from each data source separately and afterwards integrated into a single result. On the other hand, the In-Advance Integration copies, consolidates and integrates data from the data sources into a single database, which can be queried afterwards.

In data integration by means of ontologies, according to [8], there exist three approaches, how integration can be established: the single ontology approach, the multiple ontology approach and the hybrid approach. The single ontology approach assumes that a global ontology exists, by which concepts of all local schemes and the target global schema are de-

scribed. The multiple ontology approach is based on defining a separate ontology for each of the data sources and one for the global schema, meaning that they do not share the same vocabulary.

For this work, we use a hybrid approach combined with In-Advance Integration. A reference ontology, the CTDMO, is used as a basis for describing the target database as well as the data sources, but for each of them a single ontology is used. Furthermore, the data integration is executed in advance.

In the core application the user can create new import packages for the SQL Server Integration Services by providing the respective data source ontology to the application. The application can also assist him to create this ontology. To create the import package, he maps concepts and properties of the data source ontology to the selected concepts and properties of the CTDMO in a visual way. After providing some additional information about the new package, the web service will create the package based on the mappings and the additional information automatically.

3 The Case Study

To verify the applicability of the proposed approach and to give a proof of concept, the proposed system will be used in a pilot trial conducted in the MyHeart subproject Heart Failure Management (HFM). MyHeart is a Philips-led FP6 Integrated Project, aiming at developing intelligent systems for the prevention and monitoring of cardiovascular diseases (CVD), the leading cause of death in the western world.

The pilot trial of the HFM project is conducted by the Philips Medical Signal Processing (MSP) group in cooperation with the University Hospital Aachen. The involved parties include mathematicians, electronic engineers, physicists and computer scientists of the MSP group and physicians of the University Hospital.

In the trial both, new and common medical devices are used to measure certain parameters of the study participants and have various output formats, e.g., flat files or binary data. Furthermore, data is acquired manually by physicians on site.

4 Conclusion and Future Work

We developed an ontology-based system, which enables users to create databases and data integration packages in a flexible and comfortable way. The CDTM ontology, developed during the project, serves as a reference ontology. The usability of the approach is assessed by using the system for the data management of a concrete pilot trial.

First evaluations of the CTDM ontology showed that the domain experts need a personal introduction to the ontology basics and the concepts of the CTDM ontology on the higher abstraction levels to be able

to extend it. The guideline without any further explanation was not sufficient.

The CTDM ontology and the system in general is intended to cover many aspects of acquired data in clinical studies, but further tests with other pilot trials have to be made, to see, if the system is flexible and comprehensive enough to be suitable for different kinds of trials (e.g. drug trials).

Future work on the system has to consider functionality to adapt changes to created databases and to offer the ability to also convert constraints for ontology concepts and properties into respective structures in a relational schema.

Acknowledgement

References

[1] Gruber, T. A translation approach to portable ontology specifications..*Knowledge Acquisition*, Volume 5, pp. 199-220, 1993.

[2] Gomez-Perez, Ontology Evaluation. In *Handbook on Ontologies.* Springer Verlag, pp. 250-273, 2004.

[3] Jean, S. et al. Domain Ontologies: A Database-Oriented Analysis. In *Proc. of Web Information Systems and Technologies*, pp. 341-351, 2006.

[4] Li, H et. al. Model Driven Laboratory Information Management Systems. In *AMIA 2006 Symposium Proceedings*, pp. 484-488, 2006.

[5] The Protégé Project, http://protege.stanford.edu

[6] Trinkunas, J. et. al. A Graph Oriented Model For Ontology Transformation Into Conceptual Data Model. *Information Technology And Control*, 36, pp. 126-132, 2007.

[7] Vysniauskas, E. et. al. Transforming Ontology Representation from OWL to Relational Database. In *Information Technology And Control*, 35, pp. 333-343, 2006.

[8] Wache, H. et. al. Ontology-Based Integration of Information - A Survey of Existing Approaches. In *Proc. Intl. Workshop on Ontologies and Information Sharing*, pp. 108-117, 2001.

[9] Weaver, M. CTMS Procurement: The seven deadly sins. *Clinical Research Focus,* 17(2), pp. 18-20, 2006.