TMF Special Issue: Unlocking Data for Clinical Research - The German i2b2 Experience

T. Ganslandt[1]; S. Mate[2], Helbing K[3], U. Sax[3,4], H.U. Prokosch[1,2]

[1] Center for Medical Information and Communication, Erlangen University Hospital, Erlangen, Germany
[2] Chair of Medical Informatics, Friedrich-Alexander-University Erlangen-Nuremberg, Erlangen, Germany
[3] Department of Medical Informatics, University Medical Center Göttingen, Göttingen, Germany
[4] Division of Information Technology, University Medical Center Göttingen, Göttingen, Germany

Keywords

Medical Record Systems; Information Storage and Retrieval; Single-Source

Summary

Objective: Data from clinical care is increasingly being used for research purposes. The i2b2 platform has been introduced in some US research communities as a tool for data integration and querying be clinical users. The purpose of this project was to assess the applicability of i2b2 in Germany regarding use cases, functionality and integration with privacy enhancing tools.
Methods: A set of four research usage scenarios was chosen, including the transformation and import of ontology and fact data from existing clinical data collections into i2b2 v1.4 instances. Query performance was measured in comparison to native SQL queries. A setup and administration tool for i2b2 was developed. An extraction tool for CDISC ODM data was programmed. Interfaces for the TMF PID Generator and Pseudonymization Service were implemented.
Results: Data could be imported in all tested scenarios from various source systems, including the generation of i2b2 ontology definitions. Query performance in i2b2 was on average inferior to native SQL queries. A query restricting for co-occurring age and diagnosis items delivered different results between i2b2 and native SQL. A query containing temporal combinations of items could not be designed in i2b2. The integration of privacy enhancing tools into the import process was possible without modification of the i2b2 platform itself.
Conclusion: i2b2 provides an intuitive interface allowing clinical users to construct complex queries. Data from existing sources can be re-used leveraged by standards like CDISC ODM or through individual transformation scripts. Integration with existing privacy enhancing tools was straightforward. Limitations regarding query performance, temporal or post-coordinated criteria should be addressed. Even though i2b2 does not cover all possible aspects of data analysis, it is an important component for single-source implementation strategies.

Correspondence to:

Dr. med. Thomas Ganslandt
Center for Medical Information and Communication
Erlangen University Hospital
Krankenhausstr. 12, DE-91054 Erlangen
Germany
E-mail: thomas.ganslandt@uk-erlangen.de

Introduction

Increasing amounts of data are captured digitally during clinical routine care with the primary objective of supporting the care process. The need for making these data available for scientific re-use has been discussed extensively [1-3]. Thus, routine data could be beneficial throughout the whole research lifecycle, as a base for hypothesis generation, estimation of expected study cohort sizes, to enhance patient recruitment in ongoing studies and to reduce duplication of data entry, among other uses. Currently, the availability of these data for research purposes is limited, and projects trying to tap them are facing complex challenges: Data items are typically spread among disparate databases, including electronic medical records (EMR), laboratory and order entry systems or electronic data capture (EDC) systems for research, and have to be extracted and transformed into a common schema optimized for analysis. Records of the same patients across various systems have to be linked and de-duplicated to create added value. At the same time, data protection guidelines mandate the de-identification of patient data as well as strict access controls. Query interfaces have to be developed that allow clinical users to analyze complex datasets.

The "informatics for integrating biology and the bedside" project (i2b2) was funded by the NIH as one of seven National Centers for Biomedical Computing to provide a generic and scalable platform for the integration of clinical and research data[4;5]. The i2b2 platform uses a modular approach that provides several "cells" to carry out queries, export and visualize result data as well as generate additional data points through further analysis, e.g. natural language processing[6;7]. i2b2 has reached wide adoption in the United States and created an active user community[8-12]. i2b2 uses a generic Entity-Attribute-Value (EAV) database schema[13;14] that facilitates rapid integration with additional data sources as well as a simplified user interface that allows clinicians to formulate complex Boolean queries. Limitations of this approach have been discussed, including restrictions in the formulation of queries containing post-coordination as well as certain temporal or aggregated expressions[14;15].

The German Technology and Method Platform for Networked Medical Research (TMF)[1] sponsored an IT strategy project in 2009 in order to identify relevant tools and platforms to support networked medical research in the near future. Within this project, the authors' objective was to assess the applicability of i2b2 within the German research context, including the identification of possible use cases and limitations, data protection issues and the integration with privacy enhancing tools developed by the TMF.

Methods

Installations of i2b2 v1.4 were carried out using both a preconfigured virtual machine (VM) as well as from scratch using source code provided on the i2b2 website[2]. Installations were set up on a VMware ESX™ virtualization platform (VMware Inc) with a single virtual CPU and 1GB RAM. A setup and administration tool was developed locally to simplify installation from source code.

The system was tested in 4 different usage scenarios:

A. Query frontend for a local clinical data warehouse
B. Research database for local prostate cancer project
C. Research database for multicenter dermatologic research network
D. Research database for long term storage

For scenario A, a Clinical Data Warehouse (Cognos BI™, IBM Inc) already established at Erlangen University Hospital was integrated with i2b2. Metadata for diagnoses (German ICD 10 GM 2010) and procedures (German OPS 2010) were converted from star schema dimension tables into the i2b2 ontology format using the IBM Cognos DataManager™ Extraction, Transformation and

[1] http://www.tmf-ev.de/ (retrieved 08/05/2010)
[2] https://www.i2b2.org/software/index.html (retrieved 08/05/2010)

Loading (ETL) tool. Patient demographics and fact data for diagnoses and procedures were transformed from warehouse fact tables into the i2b2 EAV representation using the same method. The TMF PID-Generator, a tool for pseudonymization and the robust linkage of patient demographic data[16-18] was integrated into the import process. Pseudonymization was carried out asynchronously to generate a patient list containing both identifying data and pseudonyms. The patient list was then joined to the demographic source data during conversion into the i2b2 format.

To verify query performance and capabilities within scenario A, a set of consecutive clinical selections were carried out in comparison of direct SQL requests on the clinical data warehouse (relational database schema) against requests through the i2b2 user interface (generic EAV database schema). Both the clinical data warehouse and the i2b2 project resided on the same Oracle™ database server and contained the same number of patients and diagnosis/procedure codes. The query was started by selecting a set of female patients presenting in 2009 with a diagnosis of breast cancer (ICD10-GM[3] code C50). The selection was then further restricted by a procedure code for radiation therapy (OPS[4] code 8-52), a procedure code for surgical breast excision or resection (OPS code 5-87) and a procedure code for chemotherapy (OPS code 8-54). The dataset was then restricted to the patients aged 30-49 years at diagnosis. Finally, additional restrictions were added for the chemotherapy to have occurred before the surgical procedure and the radiation to have taken place after surgery. All queries were carried out 5 times consecutively and the average runtime was computed.

For scenario B, a prostate cancer documentation based on the Erlangen University hospital EMR system (Soarian™, Siemens Inc) was integrated with i2b2. Ontology metadata was automatically generated from the EMR forms definition tables using a Perl script for conversion into the i2b2 ontology format. Fact data was extracted from the EMR system for script based conversion into the i2b2 EAV representation.

For scenario C, a locally developed EDC system used with the German Epidermolysis Bullosa Research Network[19] was integrated with i2b2. Ontology metadata was transferred manually into an Excel™ (Microsoft Inc) spreadsheet containing the concept and hierarchy followed by script-based conversion into the i2b2 ontology format.

For Scenario D the trial database of the Competence Network for Congenital Heart Defects (KN AHF) was integrated with i2b2. The trial database is based on the commercial EDC system SecuTrial™ (iAS GmbH). Data was exported using standard CDISC ODM (Operational Data Model)[5] 1.2 and 1.3 export files[20]. Ontology metadata was extracted and converted into SQL-statements suitable for populating the i2b2 ontology tables using a Java-based program. Fact data was similarly extracted from the ODM files and converted into individual SQL statements for each patient. The TMF pseudonymization service (PSD), a tool for the reversible pseudonymization of medical research data[17;18], was integrated into the import process. The integration consisted of Python programs importing the fact data SQL files generated by the ODM converter, sending them in XML form to the PSD web service, receiving the pseudonymized records from the PSD web service, parsing and storing them in SQL files ready for import into i2b2.

[3] International Classification of Diseases (ICD10) with German Modifications, version 2010 (http://www.dimdi.de/static/de/klassi/diagnosen/icd10/, retrieved 08/10/2010)
[4] German Procedure Codes, version 2010 (http://www.dimdi.de/static/de/klassi/prozeduren/ops301/, retrieved 08/10/2010)
[5] http://www.cdisc.org/odm (retrieved 09/02/2010)

Results

The setup and administration tool developed in this project provided functions for the setup and initial configuration of a fully functional i2b2 server instance from source including the installation of required Linux packages. Administration functions covered the setup of i2b2 project instances including the configuration of related database schemas and i2b2 users. The tool was made available for public use on the TMF website[6] and the i2b2 Academic User Group (AUG)[7].

[place figure 1 about here]

Fig. 1 shows the process established in scenarios A-C for the conversion and import of ontology, demographic and fact data into the i2b2 project database. Ontology metadata was extracted directly from source databases in scenarios A and B and was prepared manually for scenario C. Table 1 presents the number of patients, ontology and fact records imported in each scenario as well as loading times. The import process was tested both with and without PID generator integration for demographic data in scenario A. Fig. 2 shows the performance of the PID generator against the number of coded demographic records. Table 2 shows the composition of the loading time for scenario D, including the pseudonymization service.

[place figure 2 about here]
[place table 2 about here]
[place table 3 about here]

Table 3 presents the record counts and runtimes from the SQL/i2b2 query comparison in scenario A. Fig. 3 shows a screenshot of a prostate cancer EMR form from scenario C in comparison to an i2b2 ontology hierarchy extracted from its metadata. Fig. 4 illustrates the import process established in scenario D for the extraction of ontology and fact data from ODM files and their subsequent processing through the TMF pseudonymization service.

[place figure 3 about here]
[place figure 4 about here]

Discussion

While the preconfigured i2b2 virtual machine download provides a quick method of experimenting with the platform, it poses restrictions regarding space, operating system updates and adherence to local IT guidelines. A dedicated installation from source should therefore be used for production environments. The setup and configuration from scratch, however, was complicated by dependencies on specific library versions and multiple interdependent configuration steps. Feedback gathered at two national i2b2 workshops indicated that the complex setup posed a serious obstacle to production use at several sites. By the creation of a dedicated setup and administration tool it was possible to automate this process and reduce installation and configuration times to a few minutes.

In all tested usage scenarios it was possible to integrate i2b2 with existing data sources. The simple hierarchical structure of the i2b2 ontology allowed a direct conversion of standard star schema dimension tables from the clinical data warehouse. Metadata definitions extracted from the EMR could be structured by form, field and value levels which provide easy recognition for i2b2 users experienced with the EMR. For data sources without readily available structured metadata, a manual spreadsheet could be constructed containing a similar form/field/value hierarchy.

---

[6] http://www.tmf-ev.de/Forum.aspx (registration required, retrieved 08/05/2010)
[7] http://www.i2b2aug.org/ (registration required, retrieved 09/01/2010)

Pseudonymization tools developed to meet national data protection requirements could be integrated seamlessly into the import process. While performance of the TMF PID generator is fast (average >2000 records per minute), pseudonymization of the full demographic dataset in scenario A took more than 5 hours. By using asynchronous integration, this additional time burden can be separated from the import workflow itself. Additionally, only new or modified records have to be pseudonymized once the initial dataset has been processed.

The query performance comparison (s. table 3) showed that the average runtime was generally longer in i2b2 than using native SQL queries with factors ranging between 5 to 10 times slower. The data warehouse tables used for native queries are modeled in a relational schema optimized for reporting, making use of indices and optimizer hints to gain maximum performance. i2b2 in comparison uses a generic schema, putting all dimensional data in a single table and all fact data in one additional large table. Data segmentation in a relationally modeled schema allows the database to gain speed by having to access only those tables containing data relevant to the query. It should be evaluated whether i2b2 query performance can be optimized by partitioning the fact data table, modifying indices or adding database-specific optimizer hints. It should be noted that the creation of the native SQL statements for the queries required detailed knowledge of the database structure as well as complex SQL commands including subselects and optimizer hints. Using the i2b2 frontend, all queries were constructed graphically, regardless of the underlying table structures.
The number of patients retrieved was identical between native SQL and i2b2 for the first 4 queries. Adding an age restriction in step 5 resulted in different patient counts. Further analysis revealed that i2b2 retrieved 3 additional patients because the age restriction was handled differently. In native SQL it was possible to combine the diagnosis and age restriction in the same statement. i2b2, however, treats each query item separately and combines them afterwards using boolean operators. The age restriction was thus not applied in co-occurrence with the breast cancer diagnosis, but rather as a separate set of all patients aged 30-49, regardless of diagnosis. As age is an important inclusion/exclusion criterion in many patient samples, i2b2 should be extended to allow restricting patient age in direct combination with other query items.
The final query step required the temporal combination of query items in the sense that they had to occur in a specific sequence over time. As it has been noted before[14], i2b2 does not provide functions to define this type of query, so this step could not be carried out in i2b2. Even though the addition of temporal constraints would be desirable, the increased complexity of the user interface should be balanced against overall usability. Alternatively, data could be preprocessed before importing to provide derived fact items containing the required temporal restrictions[14].

The implementation of a generic CDISC ODM parser and converter for i2b2 in scenario D allowed the automatic extraction of both ontology and fact data from a standardized format. As ODM is used widely e.g. in pharmaceutical trials[20], this tool could potentially be useful for many site looking at analyzing their study datasets in i2b2. Metadata available in the ODM files was, however, not in all cases sufficient to generate optimal ontology definitions: e.g. for numeric items the datatype itself is defined in the ODM file, but there is no information about suitable intervals for displaying valid choices in the ontology. i2b2 currently does not provide a means of modifying ontology data, apart from importing a new ontology dataset changed outside of the system. It should be considered to implement an ontology editor that can be pre-populated with metadata from source systems. Users could then adapt the ontology where needed and add details that could not be derived from source systems.

The TMF pseudonymization service could be integrated into the import process by the addition of sending and receiving programs accessing the PSD web service in scenario D. No modifications of the i2b2 systems had to be carried out. Integration of the TMF Pseudonymization Service resulted in an overhead of 65% for additional processing during loading time (s. table 2). The Pseudonymization Service currently accepts only individual patient records, which made it necessary to individually extract from ODM, encode into XML, decode the pseudonymized record from XML and load them into the i2b2 database for each patient. Extension the Pseudonymization Service to allow

batch processing should result in a relevant reduction of processing time. Also, records were loaded with individual SQL statements into the i2b2 database. In scenario A, much higher loading performance was achieved by using flat-file-based batch importing (Oracle SQL*Loader™), which could be implemented for the ODM/PSD pathway as well.
Role-based access controls became available in i2b2 v1.4 that can restrict user access to aggregated patient counts rather than individual, exportable records. This new feature can be used for a graduated approach, allowing a broader group of users to query a non-identified view of the database for relevant subsets and then request approval for full access. The platform, however, does not provide ways to generally restrict access to specific subsets (e.g. the patients of a single department). As a workaround, subsets can be extracted and copied into separate i2b2 project instances. When applied to large datasets, this approach would however greatly increase loading times as well as the complexity of database and user administration. The addition of fine-grained user permissions in i2b2 should therefore be considered.

It was demonstrated that i2b2 is a viable platform for data analysis in several real-world usage scenarios. Data re-use can be facilitated by leveraging standards (e.g. CDISC ODM) or simple conversion scripts. The platform's primary benefit consists in providing an intuitive graphical interface allowing clinical, non-IT users to construct complex queries based on a simple generic database schema. Side-effects of the simplified database schema, however, are performance issues on large datasets as well as restrictions concerning temporal or post-coordinated queries. These issues should be addressed in further revisions of the platform.
The integration of privacy-enhancing tools like the TMF PID-Generator and Pseudonymization Service into the import process was straightforward and did not require modifications to the platform. For use within multi-institutional setups, fine-grained user permissions should be implemented.
Generating ontology information from metadata available in source systems was possible in all tested scenarios. Metadata were, however, not in all cases sufficient to create ontology definitions capturing all aspects of the related data items. An ontology editor allowing users to supplement metadata imported from source systems would be a valuable addition.

Conclusions

i2b2 does not provide a "one-size-fits-all" solution for all possible analytic use cases. E.g. statistic analysis requires export and processing in a dedicated statistics package. Periodic reporting (e.g. of clinical coding data) should be carried out using dedicated business intelligence software that allows the execution of complex formatted reports. i2b2 does, however, provide a simple unified gateway for clinicians to access and re-use data that in the past has been available only in separate and often "closed" systems. It is therefore an important building block for future single-source implementation strategies.

Clinical Relevance Statement

Single-source strategies facilitate the re-use of data acquired in routine clinical care for research purposes. The availability of adequate tools for data integration and analysis positively impacts the cost-effectiveness, quality and timeliness of clinical research projects relying on such data.

Conflict of Interest

The authors have established in 08/2009 a memorandum of understanding with the i2b2 National Center for Biomedical Computing to collaborate on the further development, evaluation and dissemination of i2b2 in Germany.

Human Subject Research

The study was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects.

Acknowledgments

This project was supported in part by the grant KFO179 of the German Research Foundation (DFG) as well as by the grant *Kompetenznetz Angeborene Herzfehler (Competence Network for Congenital Heart Defects)* funded by the German Federal Ministry of Education and Research (BMBF), FKZ 01GI0210, and the grant *Netzwerk Epidermolysis Bullosa (German Epidermolysis Bullosa Network) funded by the German Federal Ministry of Education and Research (BMBF), FKZ 01GM0831*. The authors wish to thank Lars Reimann for his work on the integration of the TMF pseudonymization service, Steffen Zeiss, Roman Ostertag, Benedikt Schäffler and Christian Bauer for their work on the ODM import and Andreas Becker for his support on Clinical Data Warehouse integration.

Abbreviations

AUG: Academic User Group; CDISC: Clinical Data Interchange Standards Consortium; EAV: Entity-Attribute-Value; EDC: Electronic Data Capture; EMR: Electronic Medical Record; ETL: Extraction, Transformation and Loading; i2b2: Informatics for Integrating Biology and the Bedside; ICD: International Classification of Diseases; KN AHF: Competence Network for Congenital Heart Defects; ODM: Operational Data Model; OPS: Operation and Procedure Codes; PID: Patient Identifier; PSD: Pseudonymization Service; SQL: Structured Query Language; TMF: Technology and Method Platform for Networked Medical Research

Fig. 1: Ingest data flow for usage scenarios A-C

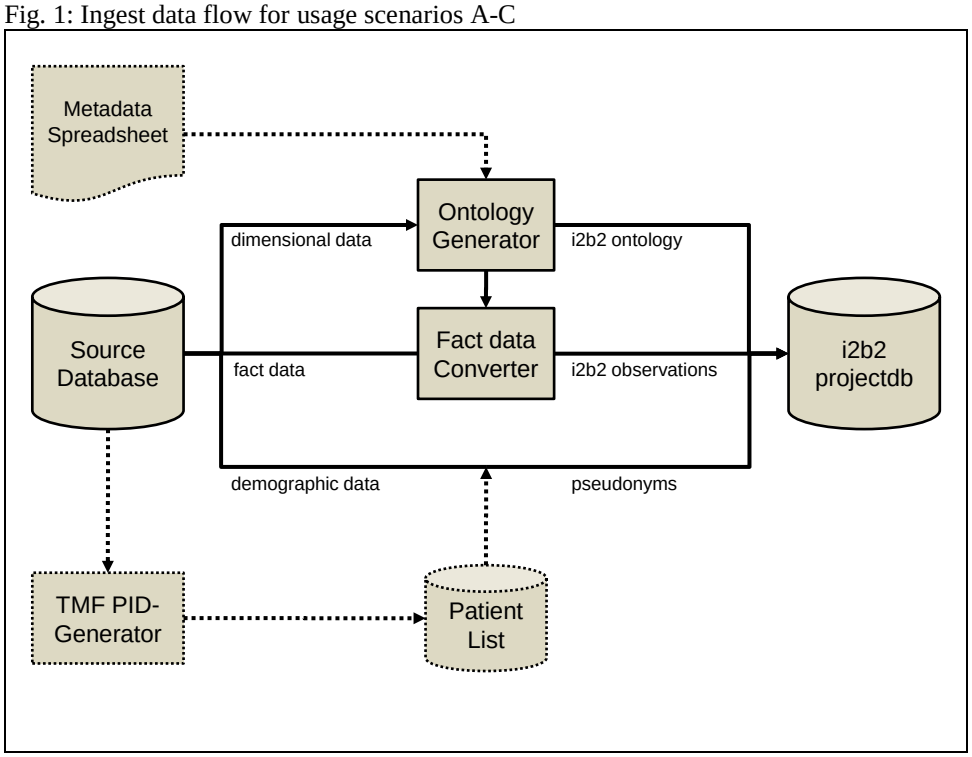

Fig. 2: Performance of TMF PID-Generator Pseudonymization tool

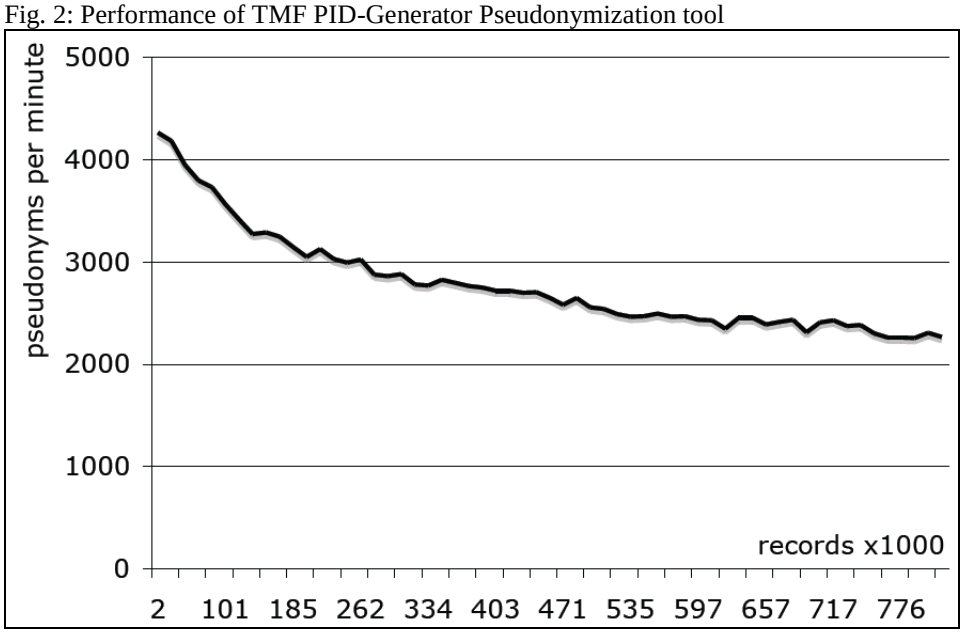

Fig. 3: Screenshots of prostate cancer EHR documentation form and corresponding i2b2 ontology elements: the left window shows the EHR system screen with a tumor histology form, the right window shows the corresponding generated ontology tree

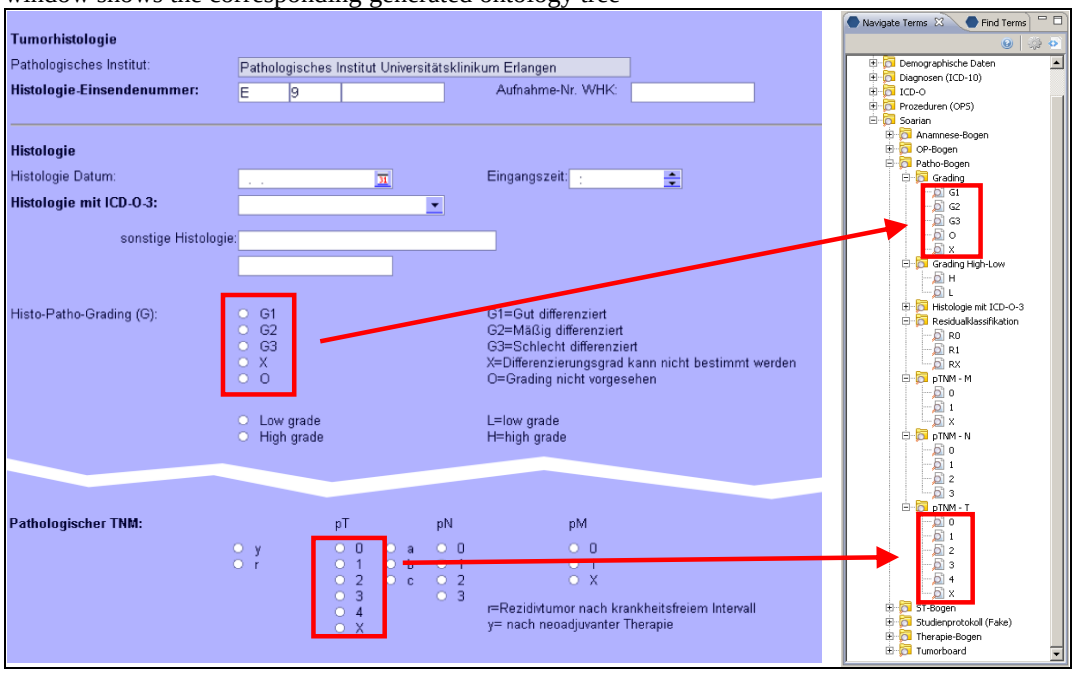

Fig. 4: Ingest workflow in scenario D with integration of TMF pseudonymization service: the ontology metadata (s. fig. 3 right window) is imported separately from the other data which is being de-identified in the lower branch using TMF tools.

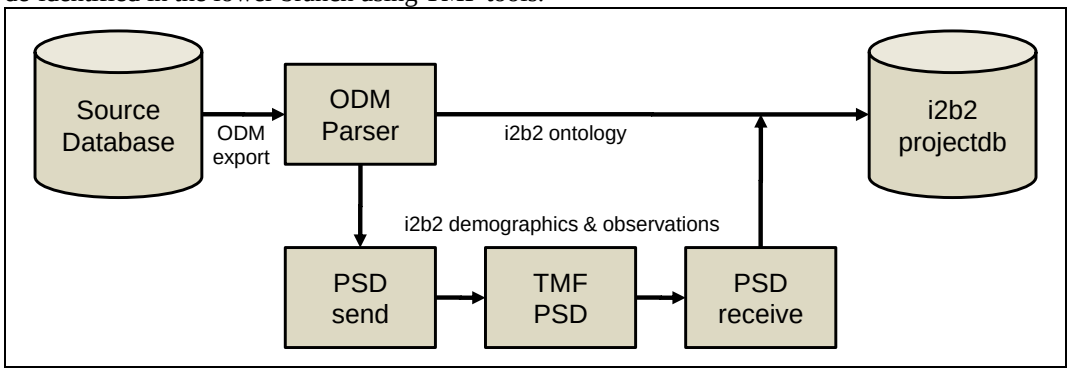

Table 1: Content structure and loading times in i2b2 usage scenarios

Scenario	Contents	Items	Concept Codes	Patients	Records	Loading time (mm:ss)
A Clinical Data Warehouse	Demographics Diagnoses Procedures	4	56,275	672,225	5,375,223	45:05
B Prostate Cancer Project	Demographics Med. History Surgery Pathology Selected Lab	46	232	121	2238	00:05
C Dermatology Research Network	Demographics Med. History Skin status	253	546	418	113,993	01:52
D Long-term research database	Demographics Diagnoses Procedures MRT/US	3195	94.117	143	54.534	33:44

Table 2: Composition of scenario D loading time

ODM->SQL	SQL->XML	PSEUD	XML->SQL	SQL->i2b2	Total time
01:25	00:03	13:15	00:03	18:57	33:44

Table 3: Query performance and capabilities for scenario A

Query Stage	Native SQL		i2b2	
	Patients retrieved	Runtime (sec)	Patients retrieved	Runtime (sec)
Female, Breast Cancer (ICD C50) in 2009	1081	9	1081	47
+ Radiation therapy (OPS 8-52)	384	5	384	57
+ Surgical Breast excision/resection (OPS 5-87)	194	8	194	39
+ Chemotherapy (OPS 8-54)	55	5	55	49
+ age group 30-49 at diagnosis	18	4	21	53
+ Chemotherapy pre & radiation post surgery	10	4	-	-

References

1. Prokosch HU, Ganslandt T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. Methods Inf Med 2009; 48(1):38-44.

2. Ohmann C, Kuchinke W. Future developments of medical informatics from the viewpoint of networked clinical research. Interoperability and integration. Methods Inf Med 2009; 48(1):45-54.

3. Kush R, Alschuler L, Ruggeri R, Cassells S, Gupta N, Bain L et al. Implementing Single Source: the STARBRITE proof-of-concept study. J Am Med Inform Assoc 2007; 14(5):662-673.

4. Murphy SN, Mendis ME, Berkowitz DA, Kohane I, Chueh HC. Integration of clinical and genetic data in the i2b2 architecture. AMIA Annu Symp Proc 2006;1040.

5. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010; 17(2):124-130.

6. Mendis M, Wattanasin N, Kuttan R, Pan W, Philips L, Hackett K et al. Integration of Hive and cell software in the i2b2 architecture. AMIA Annu Symp Proc 2007;1048.

7. Mendis M, Phillips LC, Kuttan R, Pan W, Gainer V, Kohane I et al. Integrating outside modules into the i2b2 architecture. AMIA Annu Symp Proc 2008;1054.

8. Goryachev S, Sordo M, Zeng QT. A suite of natural language processing tools developed for the I2B2 project. AMIA Annu Symp Proc 2006;931.

9. Gainer V, Hackett K, Mendis M, Kuttan R, Pan W, Phillips LC et al. Using the i2b2 hive for clinical discovery: an example. AMIA Annu Symp Proc 2007;959.

10. Uzuner O. Second i2b2 workshop on natural language processing challenges for clinical records. AMIA Annu Symp Proc 2008;1252-1253.

11. Heinze DT, Morsch ML, Potter BC, Sheffer RE, Jr. Medical i2b2 NLP smoking challenge: the A-Life system architecture and methodology. J Am Med Inform Assoc 2008; 15(1):40-43.

12. Childs LC, Enelow R, Simonsen L, Heintzelman NH, Kowalski KM, Taylor RJ. Description of a rule-based system for the i2b2 challenge in natural language processing for clinical data. J Am Med Inform Assoc 2009; 16(4):571-575.

13. Nadkarni PM, Brandt C. Data extraction and ad hoc query of an entity-attribute-value database. J Am Med Inform Assoc 1998; 5(6):511-527.

14. Deshmukh VG, Meystre SM, Mitchell JA. Evaluating the informatics for integrating biology and the bedside system for clinical research. BMC Med Res Methodol 2009; 9:70.

15. Meystre SM, Deshmukh VG, Mitchell J. A clinical use case to evaluate the i2b2 Hive: predicting asthma exacerbations. AMIA Annu Symp Proc 2009; 2009:442-446.

16. Faldum A, Pommerening K. An optimal code for patient identifiers. Comput Methods Programs Biomed 2005; 79(1):81-88.

17. Pommerening K, Reng M. Secondary use of the EHR via pseudonymisation. Stud Health Technol Inform 2004; 103:441-446.

18. Helbing K, Demiroglu SY, Rakebrandt F, Pommerening K, Rienhoff O, Sax U. A Data Protection Scheme for Medical Research Networks. Review after Five Years of Operation. Methods Inf Med 2010; 49(5).

19. Klein A, Prokosch HU, Muller M, Ganslandt T. Experiences with an interoperable data acquisition platform for multi-centric research networks based on HL7 CDA. Methods Inf Med 2007; 46(5):580-585.

20. Kuchinke W, Wiegelmann S, Verplancke P, Ohmann C. Extended cooperation in clinical studies through exchange of CDISC metadata between different study software solutions. Methods Inf Med 2006; 45(4):441-446.