View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Servicio de Coordinación de Bibliotecas de la Universidad Politécnica de Madrid

p-medicine: A Medical Informatics Platform for Integrated Large Scale Heterogeneous Patient Data

J. Marés[1], L. Shamardin[1], G. Weiler[2], A. Anguita[3], S. Sfakianakis[4], E. Neri[5], S.J. Zasada[1], N. Graf[6], P.V. Coveney[1*]

[1]University College London, London, United Kingdom; [2]Fraunhofer Institute for Biomedical Engineering, St. Ingbert, Germany; [3]Universidad Politécnica de Madrid, Madrid, Spain; [4]ICS-FORTH, Heraklion, Greece; [5]Custodix, Sint-Martens-Latem, Belgium; [2]University of Saarland, Saarland, Germany;

Abstract

Secure access to patient data is becoming of increasing importance, as medical informatics grows in significance, to both assist with population health studies, and patient specific medicine in support of treatment. However, assembling the many different types of data emanating from the clinic is in itself a difficulty, and doing so across national borders compounds the problem. In this paper we present our solution: an easy to use distributed informatics platform embedding a state of the art data warehouse incorporating a secure pseudonymisation system protecting access to personal healthcare data. Using this system, a whole range of patient derived data, from genomics to imaging to clinical records, can be assembled and linked, and then connected with analytics tools that help us to understand the data. Research performed in this environment will have immediate clinical impact for personalised patient healthcare.

Introduction

Secure access to patient data, coupled to analysis tools running on that data, promises to revolutionise the clinical decision making process for the treatment of a wide range of diseases. The problem of sharing clinical data presents a major hurdle to overcome if patient specific data analytics and computer-based modelling are to be developed for medical research purposes and, ultimately, incorporated into clinical practice.

Because of this, the data sources held by hospitals represent a major resource that is currently not adequately exploited, either by researchers or clinicians. Digitised patient data collected as part of routine clinical practice, and which can be used as input to a wide range of analytics techniques, initially resides in information systems based within the individual hospital where the data is acquired. These data include medical images obtained through techniques such as magnetic resonance imaging (MRI) or computed tomography (CT), biopsy microphotographs, DNA and RNA sequence data, proteomics, metabolomics, and other kinds of medical records.

The data held by clinical data systems can be used in at least two different ways: (1) to compose large, (pseudo)-anonymised datasets from multiple sources, in order to perform inference based machine learning and to structure and support clinical trials; (2) to run workflows in support of clinical decision making processes on individual patients.

To make this personal data available to scientists and clinicians, we have developed a data warehouse as part of the EU FP7 p-medicine project, coupled to a medical informatics platform, into which data sources from multiple hospitals can be aggregated to generate substantially larger data collections upon which more comprehensive data analytics can be performed. The benefit of this informatics platform is clear: not only do our solutions allow data from diverse sources to be linked and integrated, they also provide a common platform that scientists and clinicians can use to initiate analytics workflows based on that data, as they offer standards compliant interfaces and APIs into which many existing and future tools and services can be plugged. The flexibility of the triplestore approach (described below) also allows us to model data in new and innovative ways. These capabilities are key to meeting the needs of the new and rapidly growing field of personalised medicine. In this paper, we describe the components of our medical informatics platform, as well as how external tools can be interfaced to the warehouse to upload, download and analyse data. In our view, the flexibility and generic nature of this platform make it applicable to a whole range of medical informatics problems. It permits immediate biomedical and clinical research to be conducted. Its novelty resides in its ability to seamlessly integrate large scale heterogeneous medical

[*] Corresponding Author. E-mail: p.v.coveney@ucl.ac.uk.

data in a secure, federated and distributed environment, spanning all levels from local, regional to national and international scales.

Healthcare Information Management Systems

There are many different approaches to building electronic health record and hospital information systems. Systems such as PatientCentre from iSOFT[1] are deployed within an individual hospital to manage patient data, although the GP2GP system[2] in the UK allows GP practices to transfer records amongst themselves. There are emerging standards for integrated electronic health record systems, such as HL7[3], which facilitate data transfer between systems. Online "cloud" based systems such as Microsoft HealthVault[4] allow individual patients to store and manage their health records via an online service.

Clinical data management systems are designed to manage data coming from clinical trials, and thus are often used to federate data from multiple administrative domains. Systems such as the IBM Cognos platform[5] provide business intelligence services to pharmaceutical and life sciences companies conducting clinical trials. Microsoft's Amalga platform[6] brings historically disparate data together and makes it easy to search and gain insight from that data.

However, while these systems are designed to manage and integrate large amounts of data (potentially all of the patients treated in a hospital), they do not generally deal with sharing patient data for research purposes between multiple institutions, including the ones located in different countries. In addition, traditional electronic health care record systems do not go beyond basic data management to provide advanced analysis and decision support capabilities, which rely on high performance computing platforms currently out of the control of the administrators of the data management system.

The caBIG project[7] has developed an informatics platform designed to share basic research, clinical trials imaging data, and biobanking samples between researchers, and allow them to use the data to run analytics techniques. However, caBIG's approach has been viewed as being too broad and technology led, and not designed around the needs of clinical users and researchers[8].

tranSMART[9] is an interesting recent development in the area of clinical research data management, and has found significant use in the pharmaceutical industry. tranSMART is a knowledge management platform that enables scientists to develop and refine research hypotheses by investigating correlations between genotypic and phenotypic data, and assessing their analytical results in the context of published literature and other work. However, tranSMART does have a number of limitations: it is focused around clinical records and molecular data and it does not integrate imaging data, nor does it provide access to high performance computation for data analytics.

i2b2 provides a scalable computational framework to address the bottleneck limiting the translation of genomic findings and hypotheses in model systems relevant to human health. New computational paradigms and methodologies are begin developed and tested as part of the i2b2 project in several disease cases. The server side of the software provides file and data repository facilities, as well as ontology and identity management tools. Plug-ins allow system capabilities to be expanded, with high performance computing capabilities for data analysis for example. Users access the system from client workbench and web portal tools.

System Design Overview

Our platform is organized inside a secure framework in order to ensure the confidentiality of the stored clinical data. Inside this framework there are five main components: the upload tools, the clinical trials manager (ObTiMA, described in more detail below), the data warehouse, the ontology annotator/data translator and the pseudonymisation services. Figure 1 shows a diagram of the system's design.

Upload tools and clinical trials manager components are the ones that interact with the world outside the secure framework but, at the same time, ensure that the connections with the inner components are secure. The data warehouse is the core component that stores all the data and makes it available through a public API that can be accessed only by secure connections. The ontology annotator/data translator is the component that takes complex data uploaded to the data warehouse and adapts it to conform to standard formats and protocols in order to make it easier to use and transforms it to a common vocabulary in order to enable the semantic integration of the data. Finally, the secure framework in which the system is embedded is responsible for authentication, authorization and accounting across the informatics platform. In addition, it ensures that data is pseudonymised when it is pushed into the data warehouse.

The following sections describe in detail the data ingestion workflow, the data warehouse, the pseudonymisation process, the integration of data and ontology markup and the use of the informatics platform.

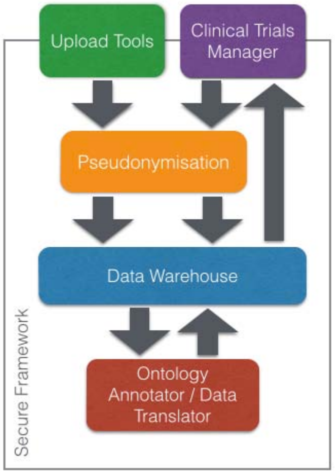

Figure 1. System design overview diagram. All traffic is secured by the authentication and authorization services in the secure framework. In addition, all data pushed into the data warehouse must pass through the pseudonymisation process to assure the privacy of the data stored in the warehouse. Finally, when data to be annotated is uploaded to the data warehouse, it triggers the ontology annotator and the data translator modules.

Data Ingestion Workflow

Health data from patients is collected by the treating physicians and stored within the treating hospitals. This data is exported by uploading it to the data warehouse. Data can also be entered into the platform's Clinical Trial Management System, ObTiMA (described in a following section). From there, data can then also be further transferred to the data warehouse. Researchers and users within the network of trust (the research domain) only have access to the *de facto* anonymous data in the data warehouse. Re-identification is only possible through the Trusted Third Party (TTP) towards the hospital.

The process of uploading new data in the data infrastructure consists of several steps: export of the data from primary sources, transformation of the data to a data format that is supported by the "upload infrastructure", pseudonymisation of the data to remove or obscure the patient identifiers, semantic annotation to attach "meaning" to the uploaded data, and semantic translation to transform the uploaded information to the semantic framework and its set of supported ontologies. The general flow of the data between the main components in the "data push" scenario is shown in Figure 2.

As in the well-known "Extract-Transform-Load" (ETL)[10] process, we can identify three main stages in the overall "upload data" scenario:

(i) Export of data from their original data sources. Our platform supports three different ways of exporting data from their sources. In the first case the data are already available on the user's computer where the user has full access to the files to be updated and the only foreseen complexity is to ensure that the data formats are supported by the upload infrastructure either directly or through a suitable transformation tool. In the second case, the data, in a multitude of formats, are stored within a "system" that allows the user to export them. In the third and most challenging case the data are stored in a Relational Database where the users need to know about the internal details of the source system, its relational schema, and must define in Structured Query Language (SQL) what is to be exported. In the latter case there are certain concerns raised with respect to the security and stability of the primary data providing systems (e.g. Hospital Information Systems) when accessed directly during their mission critical operation in the treatment domain. For these reasons the proposed architecture does not include in its design generic adapters or gateways for retrieving the information stored in the original data sources. Rather, such special data export adapters can be provided on case-by-case basis depending on the operational characteristics of a given data providing system.

(ii) Format of the uploaded data. In general the data format can be domain specific (e.g. DICOM images) and either structured or unstructured. For the raw binary formats such as medical images, the uploaded format can

be identical to the original one after the pseudonymisation process. For structured data, Comma Separated Values (CSV) can be used, a plain text format that stores tabular data. In the case of a relational database where multiple tables are to be exported, the adoption of the CSV format requires that multiple CSV files be created. Although there are certain, more expressive alternatives such as XML (Extensible Markup Language) and RDF, the CSV format provides a "lowest common denominator" for the exchange of clinical data and enjoys the best support within the pseudonymisation infrastructure; moreover, it is "natively" supported by spread sheet applications and all relational database management systems[11].

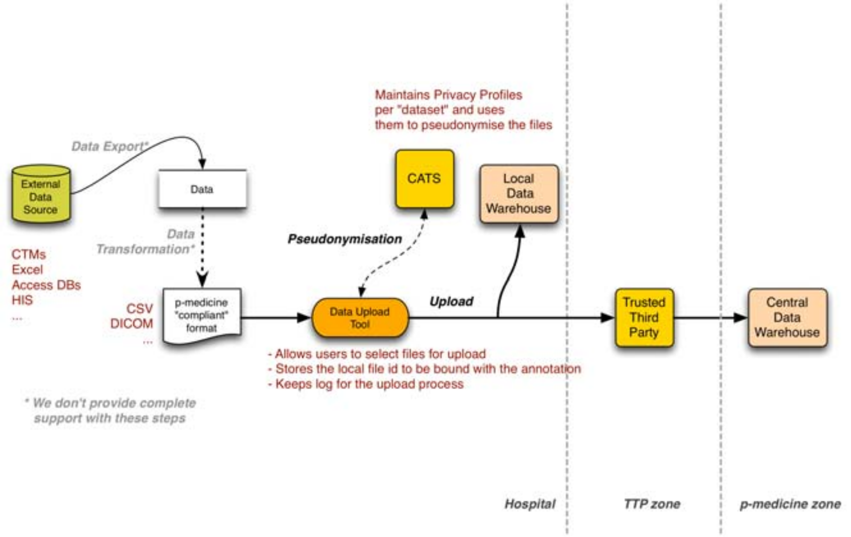
Maintains Privacy Profiles
per "dataset" and uses
them to pseudonymise the files
Data Export*
External Data Source
Data
CATS
Local Data Warehouse
Data Transformation*
Pseudonymisation
CTMs
Excel
Access DBs
HIS
...
CSV
DICOM
...
p-medicine "compliant" format
Data Upload Tool
Upload
Trusted Third Party
Central Data Warehouse
- Allows users to select files for upload
- Stores the local file id to be bound with the annotation
- Keeps log for the upload process
* We don't provide complete support with these steps
Hospital
TTP zone
p-medicine zone

Figure 2. The data upload process. Data is exported in a p-medicine "compliant" format before being pushed into the local data warehouse and the central data warehouse via the trusted third party.

(iii) Data anonymisation and upload. The Data Upload tool is a desktop GUI application for "pushing" the data through the security infrastructure to their final destination, which is the data warehouse. It provides a graphical user interface for the push services, a user-friendly tool hiding all the complexity behind its easy-to-use interface. Its main functionality is to allow a user to load files containing patient data in a variety of formats, including CSV or DICOM, perform a first round of anonymisation through the use of the Custodix Anonymization Tool Services (CATS), see the Pseudonymisation section below), and then upload the data to the data warehouse..

Data Warehouse

The data warehouse (DWH) provides the main information storage system at the core of our platform. It consists of three main components: the triplestore, the filestore and the imagestore. DWH is built to address the central notion of reproducible research. The key concept that is implemented to make research reproducible is the automatic versioning of triplestore contents: any modification of the triplestore creates a new and unique version of the triplestore (a modification of the triplestore is also called a transaction). All responses to the triplestore queries contain explicit information on the triplestore version used to fulfil the request. Any of the triplestore versions can be queried, but cannot be modified at any time.

Semantic data held in the triplestore originate from the "raw" source files, which may be stored in the DWH filestore. There is a well-defined process of extraction of semantic data from some of the common file formats; this data is transformed after extraction using the data annotation and data translation tools, and the semantically translated data is added to the DWH triplestore. Changes to the annotation of the source data are immediately reflected in the triplestore by rolling back the transaction containing the old data originating from this file's extraction and translation chain, and by creating a new transaction with new data.

Image data upload is not supported by the DWH; however, they are accessed via DWH since it acts as a protocol proxy between the DWH user and the associated Picture Archiving and Communication System (PACS) servers. Imaging metadata is automatically transformed into raw semantic form which can be used as input for the data annotation and translation chain (explained in a following section), equipped with matching triplestore update semantics similar to the filestore.

A data warehouse can be deployed centrally in the research domain or locally in the treatment domain. In the central data warehouse (CDW) the data has been pseudonymised twice (the second pseudonymisation performed by the TTP) and is protected by a contractual framework and access controls, meaning the data can be regarded as *de facto* anonymous and can be used for research. In a local data warehouse (LDW) the data is pseudonymised only once, with the pseudonymisation key kept by the hospital. That means re-pseudonymisation of the data in the LDWs and therefore reuse of that data for treatment purposes is possible, whereas for data stored in a CDW this is only possible under very limited circumstances. From the technical point of view, LDWs and CDWs are identical.

Triplestore, Filestore, Imagestore

The DWH triplestore API permits querying the triplestore contents using SPARQL queries. It also allows plain export of the statements matching the given criteria (subject, predicate and/or object match filters). Updates to the triplestore can be made either by direct upload/deletion of specified statements, or by using the SPARQL UPDATE language. Raw data and the final processed triples added to the triplestore are tightly coupled as the user is allowed to update the annotation description at any time and then the old triples generated by this file can be updated to the new format at any time. In addition, to maintain the versioning control, the new triples are put in a new version of the triplestore together with all other triples already existent. Each triple may appear in some triplestore version exactly once, so the concept of modification rollback (transaction rollback) requires a more precise definition.

Filestore manages the storage of any type of file into DWH. Its API allows uploading and downloading a file just by providing the appropriate URL with the file identifier. In addition, when uploading a new file, whether a comma-separated value (CSV) file or an Access database, a mechanism to extract triples from the file is triggered and they are stored in the triplestore.

Finally, the imagestore manages the DICOM (Digital Imaging and Communications in Medicine) image storage while its API allows two different actions: downloading of DICOM image files and accessing the extracted triples pertaining to a DICOM image file. This is achieved by simply providing the appropriate image ID in the request URL.

Pseudonymisation

There is a clear distinction between the health treatment domain and the research domain. Medical data contain sensitive information about many patients' health and wellness. For research purposes, such data must be used in anonymised form whenever possible for legal and ethical reasons. Anonymisation is the best way to protect a patient's privacy. The platform should allow the re-identification of a patient when the research results reveal that a certain therapy would be highly effective for that given patient. Therefore data cannot be fully anonymised.

The secure framework is thus based on *de facto* anonymised data. Pseudonymous data can be regarded as *de facto* anonymous data[12,13] whenever the researchers working with the data do not possess the key linking it back to the patient and there are protection measures in place that prevent the researchers from trying to re-identify patients.

Where authorisation is given to upload data to the data warehouse, it is the clinician who triggers the transmission of the respective medical data to the warehouse. The data is first pseudonymised at the source, i.e. the hospital, and only then uploaded through the Trusted Third Party (TTP) into the data warehouse. The TTP anonymises the data through a second pseudonymisation round. The TTP, which does not dispose of any health data, also serves as a vault containing the link back to the patient if needed. The use of a TTP also assists in linking data from the same patient emanating from multiple sources.

Researchers and users within the network of trust (the research domain) only have access to the anonymised data in the data warehouse. Re-identification is only possible through the TTP and is provided unidirectionally to the source hospital.

In the current version of our health informatics platform, we make use of Custodix Anonymisation Tool Services (CATS), a set of tools and services responsible for the de-identification or anonymisation of patient data files, provided by the Belgian company Custodix, which acts as TTP. CATS anonymises or pseudonymises a data file based on a set of pre-configured transformation rules through so-called privacy profiles. The adequate definition of the transformation functions to be applied to an input file involves a thorough risk assessment. This is largely a manual task but, once defined, CATS can handle data without much effort. Once the transformations have been defined on a generic data model, all that the data uploader needs to do is map the data to that generic model. This two-step approach allows for uniform processing of data in different formats. It is also more convenient for setting up a project and provides a higher assurance level with respect to compliance.

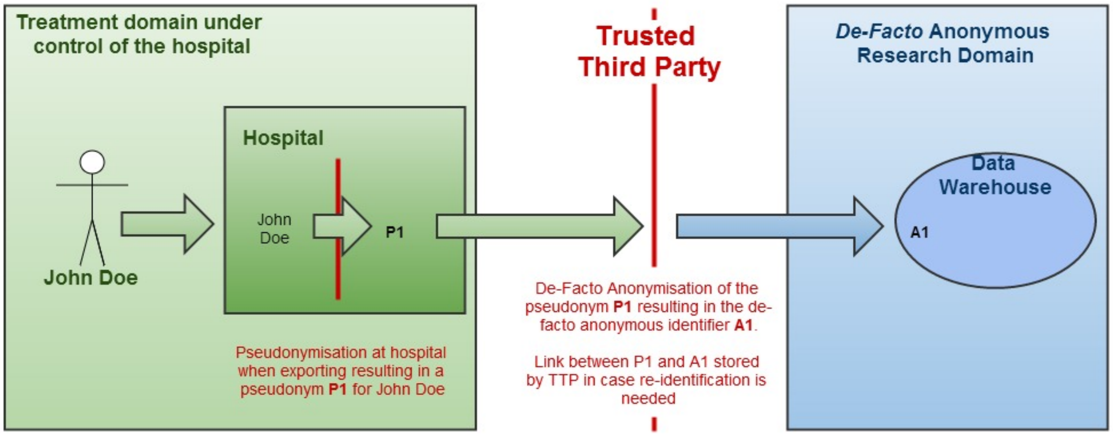
Treatment domain under control of the hospital
Hospital
John Doe
P1
John Doe
Pseudonymisation at hospital when exporting resulting in a pseudonym **P1** for John Doe
Trusted Third Party
De-Facto Anonymisation of the pseudonym **P1** resulting in the de-facto anonymous identifier **A1**.
Link between P1 and A1 stored by TTP in case re-identification is needed
De-Facto Anonymous Research Domain
Data Warehouse
A1

Figure 3. Pseudonymisation overview. Each user receives a pseudonym when data is pushed from the hospital. Then it is sent to the trusted third party (TTP) and it generates an anonymous identifier for this user, which is the one that is stored in the data warehouse. However, the TTP stores the link in case re-identification is needed.

Pseudonymisation functions typically generate a pseudonym based on a patient's identifying information (such as ID, name, date and place of birth). CATS therefore uses the Patient Identity Management System (PIMS) to issue pseudonyms. PIMS tries to assign the same pseudonym to the same patient by checking whether that patient has already been registered in the common PIMS database. A new pseudonym will only be assigned when there is no such patient registered so far. If the patient had already been registered the existing pseudonym will be attributed to him/her. Thus PIMS avoids the creation of different pseudonyms for the same patient (synonyms) as well as the creation of the same pseudonyms for different patients (homonyms).

Ideally, though, identifying information should never leave the source. Therefore CATS encrypts individual identifying attributes (name, data of birth, address, etc.) before sending them to PIMS. Due to the nature of cryptographic algorithms, very similar attributes (e.g. typographical errors) will be transformed to different encrypted values. Encryption does not maintain the similarity between records. For this reason, fault-tolerant matching (implemented by the matching engine in PIMS) on individual attributes is not possible. It is important to keep in mind that there is still a risk of re-identification when using encrypted attributes. Through statistical or frequency analysis techniques, re-identification of (parts of) encrypted attributes can still be achieved. To tackle the problem of matching encrypted records PIMS uses Q-grams[14] and bloom filters[15].

A Q-gram is a word with length Q that is a substring of a given word. Q-grams are used in fault-tolerant matching of words. The biggest disadvantage of this algorithm is the massive amount of subsets that are generated and encrypted at the source. Each encrypted subset must be sent to the matching service that will match with all previously received subsets. This requires substantial resources (network, storage, memory, etc.). A bloom filter is a compact data structure that allows checking for whether a given element is part of a collection, without the need to store the complete collection. Due to the compact representation false positives are possible. A bloom filter could indicate that an element is part of a collection but in reality it is not. On the other hand, false negatives can never occur. Bloom filters can be applied to investigate whether a given substring is part of a given word, without showing the complete word. To overcome the disadvantages of the Q-gram technique, we use bloom filters to match encrypted data. The algorithm itself is similar, but the collection of Q-grams for a given attribute is encrypted in a bloom filter instead of a collection of subsets.

Integration of Data and Ontology Markup

To apply meaning to the data in the data warehouse, and perform reasoning across datasets, all uploaded data is annotated with an ontology. To do this, we have coupled a data annotation tool, called Ontology Annotator (OA) with our data warehouse. The OA is a web-based tool for annotating the databases prior their integration in the DWH. The annotations generated with the OA consist of the semantic alignment of databases with HDOT (Health Data Ontology Trunk). This tool is aimed at end-users–scientists and clinicians as well as database administrators–who must have some comprehension of basic database concepts, but do not necessarily have expertise in the RDF paradigm. The OA receives as input the RDF schema of a database and provides a graphical interface that represents both the schema and the HDOT ontology elements in a graph-based representation. The output is an XML-based serialization of the semantic equivalences of elements of the database with elements of HDOT, as defined by the user. The resulting annotation is submitted to the DWH, and associated to the corresponding database. This information is later used to translate the database to an HDOT-compliant form at the entity and attribute levels.

The OA relies on a view-to-view alignment approach, as opposed to the classical element-to-element approach. This means that database annotations store pairs of semantically equivalent views, instead if single elements. While there can be found works that already apply this approach[16,17], these are restricted to tabular format, and ours is the first one that employs it with RDF sources, as described elsewhere[18]. The atomic elements mapped in our alignment format are RDF views, instead of classes or properties. This results in a novel capability to solve cases of semantic heterogeneity which cannot be handled by other approaches. As a consequence, a database annotation is formed by a set of pairs of RDF views, one belonging to the annotated database, and one to HDOT, which we refer to as *entries*. Figure 4 shows the structure of the annotations in the semantic layer of our platform.

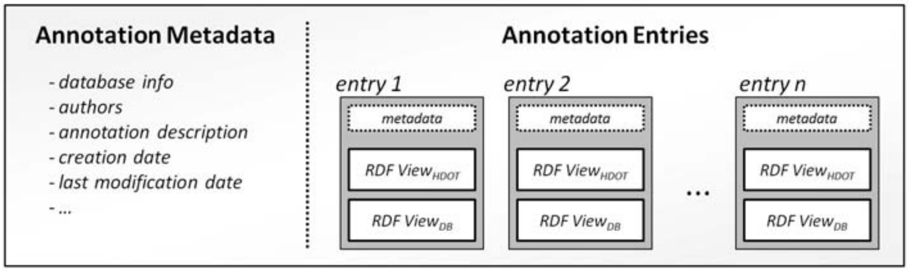

Figure 4. Structure of the annotations defined for each database. Semantic equivalences are defined with the set of entries, each containing an equivalence of a database view and an HDOT view.

This approach has been successfully applied in experiments with real data, where two heterogeneous datasets (one from the SIOP clinical trial, and one with miRNA data from some of the patients if this trial) where homogenized and merged. The former dataset contained clinical information of patients, with attributes such as age, the presence or absence of relapse, and patient identifier. The latter dataset was formed by a CSV file with 848 columns for each described miRNA molecule expression, and one column for the patient identifier. This was automatically translated to RDF by the DW. The two datasets schemas were annotated and properly aligned with the HDOT ontology. In the SIOP database, the RDF view for accessing the patients' identifiers (formed by two RDF paths, <"*patient_identifier*" → "*patient_identifier_patient_id*" → "*patient*"> and <"*patient_identifier*" → "*patient_identifier_value*" → "*string*">) was aligned with an HDOT view with similar semantics (formed by a single RDF path, <"*patient*" → "*denoted by*" → "*patient identifier*" → "*has value*" → "*string*">). In the miRNA dataset, the RDF view denoting patients' identifiers ("*Row*" → "*PatientID*" → "*string*") was aligned with the same HDOT view, thus achieving the effective merge of the two datasets in the DW triplestore.

The developed approach for annotating RDF databases is extremely generic, and allows translating any possible data model to HDOT, as long as it is modeled in RDF. The limit is actually imposed by the availability of appropriate concepts and relations in the HDOT ontology, which can be easily extended if required.

View-based annotations produced by the Ontology Annotator are used as input by the data translation process to generate HDOT-based data, achieving the desired semantic integration of the information. The output of this process is a set of RDF triples which are loaded into the DWH triplestore. These triples include provenance information and a timestamp, to allow tracking the merged data in the DWH. Translation of each annotated database is performed independently from each other. The failure in the translation of one database does not affect others, as the only consequence is that the triplestore does not receive triples from that database (upon this event, the user that created the annotation is duly notified). Figure 5 depicts the translation process performed by the data translator.

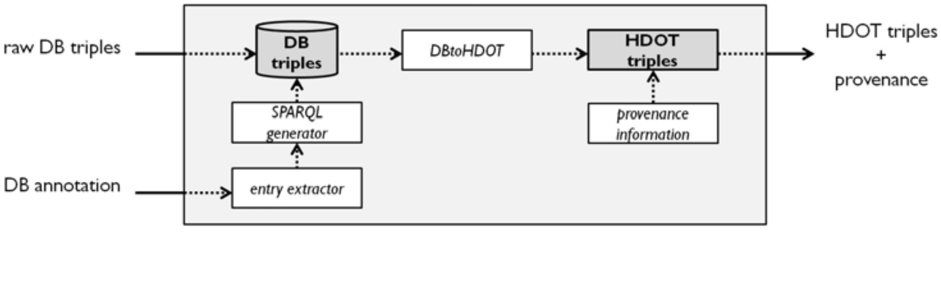

Figure 5. The data translation process uses database annotations as input. Resulting data conform to a common vocabulary provided by a cancer-related ontology. Provenance information is added to the transformed data.

Use of the Data Warehouse

The data warehouse outlined in this paper is designed to provide a flexible, robust, scalable, federated, distributed informatics platform within which to store heterogeneous medical data, ranging from clinical records, to medical images, to genomic data, in a linked, patient oriented format. In addition it has an easy use and deployment. In order to access this data, the warehouse provides a RESTful API to access, query and manipulate data. The API means that the warehouse can act as a back end to a whole range of data analytics tools, from generic statistical packages such as R, to bespoke processing tools (e.g. MINFI, LUMI, LIMMA, BWA, PICARD, SAMTOOLS, GALAXY, etc). It also makes it possible to quickly and easily tailor specific interfaces, such as web portals or mobile apps, to different groups of data warehouse users. This means that we can respond rapidly to the requirements of users to couple in the interface and analytics tools required by project researchers. By way of an example, we discuss how we have coupled our data warehouse to a clinical trials management platform.

The Ontology-based Trial Management Application[19] (ObTiMA) system is intended to support clinicians in both designing and conducting clinical trials in a user-friendly way. It has the ability to integrate an ontology into the trial building process, which makes it an extremely attractive platform for managing clinical trials. The HDOT ontology has been integrated into ObTiMA and a trial chairman is enabled to design a Case Report Form (CRF) compliant with the HDOT ontology. Therefore, data that is collected during a trial can be easily mapped to terms in the HDOT ontology. In order to integrate ObTiMA with our data warehouse, two modules have been developed for ObTiMA, which enable users to exploit the features of the platform and integrate ObTiMA with the DWH, namely the sync and the push services. Sync services enable the reuse of data stored in hospital information systems in running multicentric trials in ObTiMA. They facilitate clinical research by enabling a single entry of data for both research and healthcare avoiding redundant data entry and maximize the use of information from healthcare for research purposes. Push services have been developed in order that data collected via ontology-based CRFs in ObTiMA can be pushed into the DWH and provided in a format compliant with HDOT ontology, fully automatically without any manual preprocessing or annotation steps.

Sync Services

The sync services enable the reuse of data stored in hospital information systems in running multicentric trials in ObTiMA. In such trials several hospitals, with different heterogeneous hospital information systems (HISs), are involved, where the patient data is stored that can be reused. In such settings many authors have endorsed reusing electronic health record data in clinical trial management systems to enhance clinical trial processes and especially to avoid redundant data entry into these systems[20,21]. There are several projects that aim to link electronic health record (EHR) and clinical trial management systems (CTMS)[22,23], focusing especially on approaches to avoid redundant data entry. One of the main obstacles is the lack of semantic interoperability between different HIS systems and between HIS systems and trial management systems, partly resulting from a lack of harmonized semantic standards in the areas of health care and clinical research. The ObTiMA sync services solve these problems by utilizing the platform's semantic layer and retrieving data from the DWH. This approach has the advantage that it does not require for each new trial a full manual mapping of the data models in the EHR onto the CRFs of the trial as required in current approaches. Furthermore, it does not put restrictions on the hospital information systems used, such as compliance to standard data sets that are in general not fulfilled by current systems.

The relevant data flow for the sync services is shown in Figure 6 (left). In order to enable the reuse of data from the HIS in each hospital a local data warehouse (LDW) needs to be installed into which the data from the HIS is pushed. During this step the data is pseudonymised and semantically integrated. The sync services retrieve data from the LDWs into ObTiMA. Since the data is already integrated compliant with HDOT, the data can be mapped semi-automatically into the ontology-based CRFs of the clinical trial. Although the data in the LDW is pseudonymised, it is ensured by the Patient Identity Management Services (PIMS) of the platform's security infrastructure that appropriate data for patients in ObTiMA can be found in the LDW. The retrieved HIS data with the information as to which CRF item they are mapped into and information about its origin is shown to the clinician for confirmation. Only data originating from a hospital information system or electronic health record will be shown. The user then needs to review and confirm the data. The user can then either allow the data to be stored in the associated CRF or delete the data. Data that has been accepted are stored in the associated CRFs in the same way as if the user would have entered it manually.

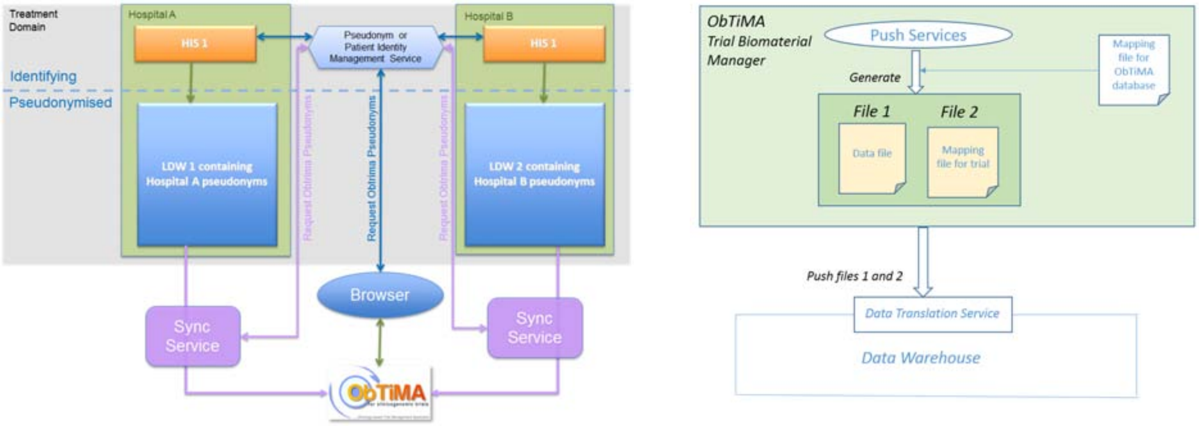
Treatment Domain
Hospital A
HIS 1
Pseudonym or Patient Identity Management Service
Hospital B
HIS 1
Identifying
Pseudonymised
LDW 1 containing Hospital A pseudonyms
Request Obtima Pseudonyms
Request Obtima Pseudonyms
Request Obtima Pseudonyms
LDW 2 containing Hospital B pseudonyms
Browser
Sync Service
Sync Service
ObTiMA
ObTiMA
Trial Biomaterial Manager
Push Services
Mapping file for ObTiMA database
Generate
File 1
File 2
Data file
Mapping file for trial
Push files 1 and 2
Data Translation Service
Data Warehouse

Figure 6. Data flow for sync services (left) and push services (right). The sync services retrieve data from the LDW installed in the hospital. The Patient Identity Management Service ensures that pseudonyms can be mapped to patients in ObTiMA. The push services are able to push trial data from ObTiMA either to the LDW or DWH. To translate trial data to a format compliant with HDOT the Data Translation Services are used.

Push Services

ObTiMA is seamlessly integrated with the DWH via push services whereby the data collected in ontology-based CRFs is ingested into the DWH and provided in a format compliant with the HDOT ontology without any manual pre-processing or annotation step. Pushing the data can be easily triggered by a trial chairman from the ObTiMA user interface. The process of pushing data from ObTiMA into the DWH is depicted in Figure 6 (right). In order to translate the data from the ObTiMA database to a format compliant with HDOT that can be stored in the DWH the Data Translation Services of the platform’s semantic layer are used. These services require two input files: a “data file” and a “mapping file” for the selected trial. The “data file” contains the trial data that will be pushed in the form of RDF triples. The “mapping file” contains the ontology-annotations necessary to translate the trial data into an HDOT compliant format. The two files are automatically generated and sent to the ontology annotator and data translator tools and stored in the DWH.

Conclusion

The digitization of patient health care records, and the diversity of data sources comprising these, make imperative the development of easy-to-use, standardised health informatics platforms. The system we have presented in this paper is designed to do just that, linking pseudonymised patient data from multiple clinical sources, on which analytics and modelling tools may be applied. The flexible, distributed nature of our system makes it highly robust and scalable, and the use of pseudonymisation means that, unlike many similar platforms, should results from the analytic processess we apply to our data be found to have an impact on an individual patient, we can, via the trusted third party, feed back those results to the clinicians treating the patient.

In contrast to the similar systems presented at the beginning of this paper such as tranSMART and i2b2, our system is characterized by an advanced data warehouse, which is capable of automatically processing a variety of uploaded file formats to extract relevant data and store it in our data triples. Unlike other systems, our platform is designed from the ground up to integrate heterogeneous data types, including imaging, genomic and clinical records data. Our system also provides unique capabilities to trace the provenance of research data and to roll back commits on the data held in our warehouse. Having a central knowledge base with such heterogeneous types of data makes the data warehouse a novel and powerful way to do research. This ability to combine data of widely varying types and perform analyses on them facilitates the opportunity to make entirely new medical discoveries.

Acknowledgements

The work reported here has been funded in part by the EU FP7 p-medicine (no FP7-ICT-2009-270089) project.

References

1. iSOFT PatientCentre. Available from: http://www.isofthealth.com/en- GB/Solutions/UKCentre.aspx.
2. GP2GP. Available from: http://www.connectingforhealth.nhs.uk/systemsandservices/ gpsupport/gp2gp
3. R. Dolin, L. Alschuler, C. Beebe, P. Biron, S. Boyer, D. Essin, E. Kimber, T. Lincoln, J. Mattison. The HL7 clinical document architecture. Journal of the American Medical Informatics Association Vol. 8, Num. 6 (2001) pp. 552-569
4. Microsoft HealthVault. Available from: http://www.healthvault.com/personal/index.aspx
5. IBM Cognos. Available from: http://www-01.ibm.com/software/data/cognos/clinical-trial-management-software.html
6. Microsoft Amalga. Available from: http://www.microsoft.com/en- us/microsofthealth/products/microsoft-amalga.aspx
7. D. Fenstermacher, C. Street, T. McSherry, V. Nayak, C. Overby, M. Feldman. The cancer biomedical informatics grid (caBIG TM). Engineering in Medicine and Biology Society. IEEE-EMBS (2005), pp. 743-746
8. A. Califano, A.M. Chinnaiyan, G.M. Duyk, S.S. Gambhir, T. Hubbard, D.J. Lipman, L.D. Stein, J.Y. Wang, O.T. Bartlett, C.L. Harris. An assessment of the impact of the NCI Cancer Biomedical Informatics GRID (caBIG) (2011). Available from: http://deainfo.nci.nih.gov/advisory/bsa/bsa0311/caBIGfinalReport.pdf
9. B. Athey, M. Braxenthaler, M. Haas, Y. Guo. tranSMART: An Open Source and Community-Driven Informatics and Data Sharing Platform for Clinical and Translational Research. AMIA Summit on Translational Science, Vol. 2013, (2013) pp. 6-8
10. P. Vassiliadis, A. Simitsis. Extraction Transformation and Loading. In Encyclopedia of Database Systems, Springer (2009) pp. 1095-1101
11. CSV on the Web Working Group. Available from: http://www.w3.org/2013/csvw/
12. N. Forgó. Ethical and Legal Requirements for Transnational Genetic Research. Hart Publishing (2010)
13. P.V. Coveney, V. Diaz, P. Hunter, M. Viceconti. Computational Biomedicine. Oxford University Press, (2014)
14. T. Churches, P. Christen. Blind Data Linkage Using n-gram Similarity Comparisons. In Advances in Knowledge Discovery and Data Mining, LNCS 3056, Springer (2004) pp. 121-126
15. R. Schnell, T. Bachteler, J. Reiher. Privacy-preserving record linkage using Bloom filters. In BMC Medical Informatics and Decision Making, Vol. 9, Num. 1, Biomedical Central (2009), pp. 1-11
16. C. Knoblock, P. Szekely, J. Ambite, A. Goel, S. Gupta, K. Lerman, M. Muslea, M. Taheriyan, P. Mallick. Semi-automatically Mapping Structured Sources into the Semantic Web. In The Semantic Web: Research and Applications, LNCS 7295, Springer (2012), pp. 375-390
17. R. Parundekar, C. Knoblock, J. Ambite. Discovering Concept Coverings in Ontologies of Linked Data Sources. In The Semantic Web: ISWD 2012, LNCS 7649, Springer (2012), pp. 427-443
18. A. Anguita, M. García-Remesal, D. de la Iglesia, N. Graf, V. Maojo. Toward a View-oriented Approach for Aligning RDF-based Biomedical Repositories. Methods of Information in Medicine, Vol. 53, Num. 4 (2014)
19. Ontology-based Trial Management Application. Available from: http://obtima.org
20. J. Powell, I. Buchan. Electronic Health Records should support clinical research. Journal of Medical Internet Research, Vol. 7, Num. 1, (2005)
21. H. Stenzhorn, G. Weiler, M. Brochhausen, F. Schera, V. Kritsotakis, M. Tsiknakis, S. Kiefer and N. Graf. The ObTiMA System - Ontology-based Managing of Clinical Trials. In Proceedings of the 13th World Congress on Health (Medical) Informatics (Medinfo 2010). Studies in Health Technology and Informatics Series, Vol. 160 (2010) pp. 1090-1094
22. CDISC Healthcare Link Initiative. Available from: http://www.cdisc.org/healthcare-link
23. R. Kush, L. Alschuler, R. Ruggeri, S. Cassells, N. Gupta, L. Bain, K.Claise, M. Shah, M. Nahm. Implementing Single Source: The STARBRITE Proof-of-Concept Study. Journal of the American Medical Association, Vol. 14, Num. 5 (2007), pp. 662-673