Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 23 (2013) 24 – 35

Procedia
Computer Science

4th International Conference on Computational Systems-Biology and Bioinformatics, CSBio2013

A generic, service-based data integration framework applied to linking drugs & clinical trials

Chris Borckholder[a*], Andreas Heinzel[b], Yuriy Kaniovskyi[a], Siegfried Benkner[a], Arno Lukas[b], Bernd Mayer[b]

[a]Research Group Scientific Computing, University of Vienna, Währinger Straße 29, 1090 Vienna, Austria
[b]emergentec biodevelopment GmbH, Gersthofer Straße 29-31, 1180 Vienna, Austria

Abstract

This paper presents an integrated framework, the Vienna Cloud Environment, for building a generic, service-based data infrastructure for biomedical data integration needs. The infrastructure consists of diverse service types, which in an example scenario are used to expose and link several open drug and clinical trials data sources under a unified interface. As a result, the biomedical research community is offered a convenient way to query parameters related to drugs, for example generic compound names or brand names or their molecular targets, and associate these to their effective utilization in clinical trials. We develop and discuss several approaches on setting up such integration infrastructure and discuss implications of this framework exemplarily on an experimental evaluation of the drug-trial integration application. Practical recommendations and theoretical considerations inspired by the implementation and experimental evaluation of the drug-trials integration application are presented.

Keywords: service-oriented; data integration; mediation; framework; cloud; clinical trials; medical drugs

1. Introduction

Data richness in biomedical research has triggered the imminent need for consolidating the information available through scattered data sources. Among a manifold of specific scopes, the link between drugs and clinical trials in

* Corresponding author. Tel.: +43-1-4277-784 40 ; fax: +43-1-4277-8 784 40 .
Email address: Chris.Borckholder@univie.ac.at

doi:10.1016/j.procs.2013.10.005

order to consolidate information on drug use in specific clinical phenotype context, together with reports on drug efficacy as well as adverse drug effects substantially inform on drug mode of action. Although vast quantities of individual data sources covering drug and target information on the one hand and clinical trial (drug use) data on the other hand are available to the public, a single common interface for them is essentially missing. In practice, the task of aligning drugs to medical trials remains a tedious procedure of manual search through several distributed and heterogeneous data sources. More importantly, this problem of fragmented data sources is of generic nature in data-driven biomedical research, involving e.g. alignment of molecular features, their specifics in disease states (particularly regarding large scale analysis tools as enabled by the Omics technologies[1]), their involvement in specific molecular processes and pathways, reference functional data across species, and many more[2]. The underlying cause for this fragmentation is the intrinsic specialization of research domains, in turn seeing the emergence of a multitude of highly specialized content repositories[3]. Novel research concepts as systems biology and systems medicine, however, essentially rest on an integrated view of such specialized database content[4,5].

The Vienna Cloud Environment (VCE)[6], developed at the University of Vienna, offers tools for provisioning of a data infrastructure that allows an integrated view on distributed data sources. The infrastructure exposes data sources through web service technologies, and allows the establishment of a unified interface through a common, mediated schema. The VCE is utilized and further developed in the context of the European VPH-Share project[7] under the EU Virtual Physiological Human Initiative (VPH-I) with the objective of creating a biomedical data infrastructure comprising web services for data finding, extraction and processing[8–10]. The data involved are typically multi-scale in a qualitative as well as in a quantitative sense, and multi-modal. Its main objective is to integrate European healthcare data into a research platform for allowing analysis of risk factors in the context of disease prevention. In turn, it will contribute to improving stratification in therapy. To achieve these goals, the VPH-Share project builds a service-based IT infrastructure for integrating relevant data sources, and enables the formation and execution of medical workflows.

We demonstrate the suitability and capabilities of this infrastructure for the integration of heterogeneous data from the life science domain on the example of a *virtual data source* providing central, unified access to data from multiple, largely disjoint medical drug and clinical study databases. While in general, centralized access to the different data sources resolves the issue of data source discovery, the unified virtual schema opens up enhanced use of given data being otherwise only achievable via the use of multiple queries to the different databases and custom functionality for interlinking and filtering the individual results. On the one hand, the virtual schema allows to unify and expose information from multiple databases holding information on the same kind of real world entities via a single schema to the user, who in return will be able to query the unified set using a single query and does not even have to be aware of the fact that the actual data are in reality retrieved from multiple sources. On the other hand, the solution empowers users to formulate queries across boundaries of the underlying databases in the same way as they would query a single, conventional database. A prototypical data infrastructure involving datasets related to drugs and trials is set up in order to unify this knowledge base.

We implement an application using the client API provided by the framework to access the mediated data on the molecular drug level and the clinical trials use level, which are then processed to establish a relationship between the uses of specific drugs in individual trials addressing clinical presentations. Having these relations at hand does not only allow identifying relevant information in one type of database starting from an entity in another database, e.g. all DrugBank[11] identifiers of drugs used in ongoing clinical studies, but also provides a possibility of answering more complex questions like which drugs being already used in a specific clinical setting are currently also evaluated for efficacy in new indications[12].

In the following we provide an overview of the framework that comprises data and mediation services needed for assembling the data infrastructure, complemented by the drug-trial application, where we specifically discuss approaches on establishing the relationships between the datasets.

2. Data and mediation services

VCE data services fall in two main categories: *data services* for querying individual data sources, and *data mediation services* for providing a common integrated schema on top of individual datasets.

VCE data and mediation services are developed on the basis of OGSA-DAI[13], a state-of-art middleware for service-based data access and integration, to facilitate a transparent data network. Domain specialists often resort to different tools and solutions within their activities. As such, data in biomedical domains are often encoded in different media and storage formats[14], consequently making the unification process a challenging task. Amongst other tasks, OGSA-DAI acts as a data source wrapper that supports access to different data storage technologies including relational databases, flat files and XML databases, which addresses the challenge of heterogeneous data sources.

2.1. VCE architecture

The architecture of VCE services is presented in Fig. 1, showing the integration and interaction of the OGSA-DAI middleware as well as the custom components that make up the VCE. The architecture roughly separates the framework into two tiers, the client- and server-side. The client provides the API for user-built applications. The API is built upon the OGSA-DAI client-toolkit and is used by the drug trials application to query the required data. The server-side tier contains the VCE data services that utilize OGSA-DAI to wrap data sources and expose them through a custom-built service interface. The mediation engine utilizes OGSA-DAI's Distributed Query Processing (DQP) to federate the data, and extends it by implementing mediation techniques. In addition to the service framework, we provide tools for deployment and the runtime environment through the service provisioning component.

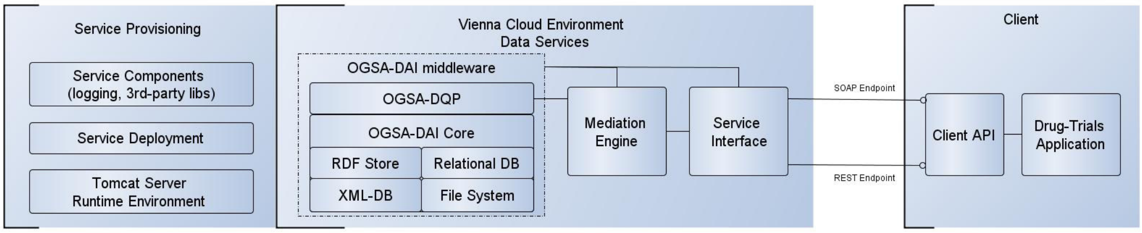

Fig. 1 VCE framework architecture. The application uses the client API, which interacts with data services through SOAP or RESTful web services. The data services use OGSA-DAI for data integration and mediation. Service provisioning comprises basic components for service deployment and the Apache Tomcat runtime environment.

2.2. Data services

A data service provides remote access to a data source via an OGSA-DAI compliant interface. Currently we only support read access, leaving the administration and management of the data source to data providers. The user is offered two well-established types of service interfaces, based on SOAP- and RESTful technologies. The REST-based interface allows the use of standard HTTP operations and execution through a simple browser, while the SOAP-based interface, implemented using the Apache CXF web service framework, utilizes the WS-* protocol stack to provide a well-defined and easily accessible endpoint for clients. Requests are allowed to be sent to either of them. The Web Service Description Document (WSDL) exposed by all services allows easy consumption and interoperation.

The query execution in VCE is based on tasks. A task represents an atomic execution and return of a single query. Employing this concept, we are able to audit executed queries on a task basis. Tasks within the VCE use OGSA-DAI mechanisms for accessing the data resource. Execution can be configured to either execute prepared statements or to utilize generic OGSA-DAI perform documents against a data source. OGSA-DAI utilizes XML perform documents, that are uploaded and executed against a data resource.

A convenient client API, based on the OGSA-DAI client-toolkit, provides the means to construct perform documents that describe OGSA-DAI workflows. These support the specification of the query and query-type, data transformation and delivery tasks. Due to the modular structure, all components, including the construction of workflows, can be extended on demand by utilizing the OGSA-DAI client-toolkit. This setup allows the construction of more complex workflows through the use of a wide range of data-centered activities. Currently the workflow builder supports SPARQL and SQL query types. The data transformation can either deliver results in the CSV or WebRowSet XML document.

2.3. Data mediation services

Data mediation services provide a unified virtual view on multiple heterogeneous data sources based on flexible data integration/mediation mechanisms. They follow the virtual data integration[15,16] technique, to establish *virtual data sources.* In contrast to data warehousing[17], these services adopt a loosely coupled, Service-oriented Architecture (SoA). The idea is to consolidate several sources in a real-time manner using a mediation schema, which maps data from different data sources to provide a common virtual data source. A single query is translated on the fly into individual queries for each separate data source, and data is then pulled directly from the original source ensuring that the results always reflect up-to-date data.

The data mediation is built upon OGSA-DAI's DQP engine. The DQP is a powerful component of OGSA-DAI that orchestrates the federation of data across the distributed data services. When processing a query, the DQP engine analyzes the schema of the underlying data services. It then builds an optimized query plan – only taking the required data into consideration - for each separate service and executes the resulting individual queries in parallel. The data is transferred into a temporary data sink of the mediation service. There it can be further processed, if required, before it is streamed back to the requester.

One drawback of the original DQP engine, however, is that the client has to be aware of the data distribution, since DQP requires mapping of the data sources to the requested data within the query. The VCE mediation engine overcomes this issue by relying on a flexible *mediation schema* that defines the data sources available for the decomposition of queries and the relations that map individual data sources to the mediated schema using the Global-as-View approach[18]. The same approach can also be used to create different virtual views on a single data source. Providing a global view through the same interface and access mechanisms for both data and mediation services implies a tight federation that offers schema, language and interface transparency.

2.4. Semantic access

Data and mediation services can be configured for semantic data access through SPARQL. SPARQL is a standard for querying Resource Description Framework (RDF) databases. Data services integrate the RDF extension of OGSA-DAI that gives them the capability to accept and process SPARQL. Consequently, OGSA-DAI supports exposing databases containing RDF triplets or SPARQL endpoints. As is the case with other data resources, data transformation and delivery activities are also available to the RDF resources. The result is returned in form of data tuples from the RDF graph. Furthermore the SPARQL-DQP extension for OGSA-DAI offers a query processor, which is able to execute SPARQL Federated Queries[19].

2.5. Deployment and configuration

Our service provisioning environment includes graphical and command line deployment tools for the automatic deployment of data and mediation services. The graphical deployment tool supports configuration of data services and the exposed data sources in a graphical manner. Users can expose one or more datasets, configure the type of data service, the details of data mediation, and additional service capabilities such as security features, if these are required.

The VCE data infrastructure relies on the concept of virtual appliances. Virtual appliances are preconfigured virtual images, deployed as virtual machines, containing the runtime environment and the software stack to allow scalability of the infrastructure on the Cloud. The infrastructure stipulates the deployment of the *data services* on servers in the vicinity of the data sources, since these have a direct connection to the data source. The *virtual data services*, on the other hand, are designed to be deployed in a Cloud environment to leverage the workload applied on them.

3. Drug trials application

We have developed an application on top of the VCE data infrastructure realizing a typical use case for biomedical researchers. With this application we aim to consolidate information available on drugs and clinical trials and allow users to traverse through the relations between the two.

3.1. Description of the datasets

Our application case rests on two data sources containing drug information, and one data source holding information on clinical studies.

The DrugBank database holds detailed information on 6,780 drugs and their corresponding drug targets. The entire DrugBank dataset is available for download in XML-format.

The Therapeutic Target Database (TTD)[20] revolves around information on therapeutic gene and protein targets as well as corresponding drugs modulating these targets. The TTD keeps track of 17,238 drugs (based on unique drug names). The data are published in plain text files. We retrieved the file holding raw information on TTD targets, the file holding cross-matching between TTD drugs and public databases, as well as the file holding synonyms for drugs and small molecules.

ClinicalTrials.gov[21] provides access to information on clinical studies. As of June 2013, ClinicalTrials.gov keeps track of 146,692 clinical studies, 114,454 unique interventions (based on the name of the intervention) of which 50,769 are drug interventions and 38,230 unique clinical conditions (again based on the naming of conditions). ClinicalTrials.gov provides a possibility to directly download study protocol records identified by a search request in XML format. A search for the wildcard character "*" was performed to retrieve study protocols for all clinical studies available in the ClinicalTrials.gov database.

These datasets all partially suffer from the following drawbacks:

- Redundancy: Since DrugBank and TTD collect independently of each other information about drugs and their respective targets, information on certain drugs will be redundant. In addition, the design of the TTD identifiers implies that different records for a single drug can exist.
- Consistency: Different names for a drug may be used throughout the different datasets, or designators may include typos, altogether impairing data comparability.
- Unstructured free text: Information is often recorded in the form of unstructured free text to ensure that even for exceptional cases all relevant facts can be stored. This holds especially true for the ClinicalTrials.gov dataset.
- No common identifiers: DrugBank, TTD and ClinicalTrials.gov do not share a set of common identifiers and as such relations between the data have to be established by other means.

3.2. Drug trials workflow

The proposed drug trial workflow consists of three major steps, depicted in Fig. 2. The first step consists of *parsing the web resources.* As described above the original datasets are publicly available as semi-structured data only. Therefore, that data has to be parsed in order to be exposed using a relational schema. In cases where data providers already have their data in relational format, this step can be omitted. As a second step, we *match the independent entities*. Implicit connections between the datasets are searched and are prepared to be integrated into a global view. The third and last workflow step is to *expose the data as an integrated data source*. Using a VCE data mediation service, a global view on the integrated data (incorporating the relations) is made accessible through web service endpoints.

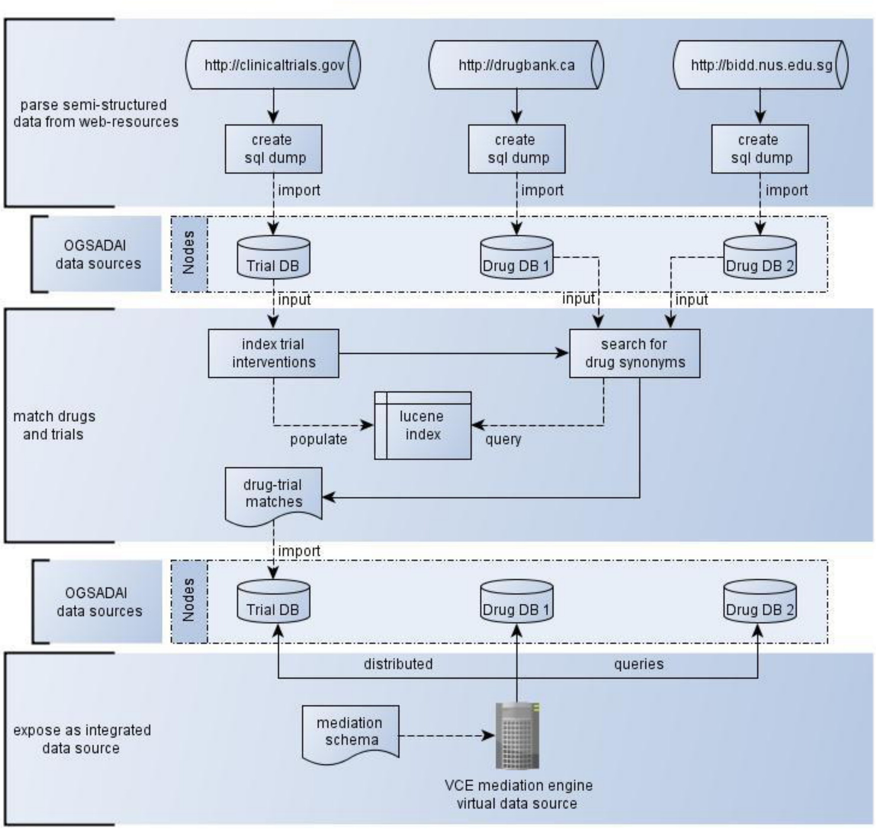

Fig. 2. Integration workflow for the drug trials data infrastructure. The three phases consist of (1) parsing semi-structured data sources and exposing them as data services, (2) match drug to trial data sources through Apache Lucene, and (3) expose aggregated data through a virtual data service.

Initially MySQL databases with required information from the three selected datasets were built. Therefore, the structure of the downloaded datasets was manually analyzed and information of relevance for allowing execution of our drug-trial workflow was identified. Based on the DrugBank dataset a relational schema for accommodating the drug data relevant for this application case was designed. Subsequently, the required information was parsed out of DrugBank's XML data file and the three TTD files, mapped onto the relational schema, and imported into two separate MySQL databases, one for each source.

In a similar fashion a relational schema for information on clinical studies was designed and populated with data from ClinicalTrials.gov study records.

The databases established from the parsed web resources are exposed as data resources in distinct OGSA-DAI containers. They are, therefore, accessible through SOAP web service endpoints. On top of this service-based data infrastructure, the VCE mediation engine provides access to integrated data. Thus, the client is relieved of the need to know the exact location and the layout of each individual dataset. Instead it interacts with the global, virtual schema established through a mediation schema.

Fig. 3 presents an overview of the global schema that has been established as part of our drug-trial use case. There are four virtual relations in that schema (*Trial*, *Drug*, *Target* and *Synonym*) and each of them is built from actual relations available in the OGSA-DAI data nodes set up beforehand. Fig. 3 also sketches how the actual data is combined in order to create the virtual relations. The integration tasks of accessing the trial data (a query is adapted and passed to the actual trial database) and accessing the unified drug data (a query is executed in parallel on both underlying drug databases and the results are combined) is the genuine purpose of the mediation engine. However, connecting an individual drug with a trial is not straightforward, as neither of the original resources provides appropriate foreign keys or join tables. To overcome this problem, we devised a matching process to make the implicit connections in the raw data visible.

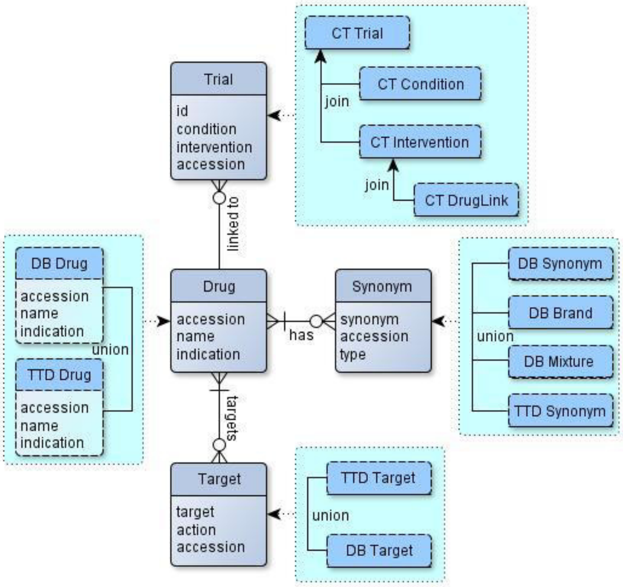

Fig. 3. Visualization of the global unified drug trial schema.

The clinical trial schema created from the parsed web resources contains an intervention field, where drug treatments being evaluated in a trial are described. The content of this field is non-structured, 'free text' description, e.g. "Prilosec® 40 mg". To connect a drug to the corresponding trials, the intervention fields are searched for drug names, their synonyms, brand names and mixture names. The general assumption is that if a drug is mentioned in the description of a trial's intervention description it is centrally connected to the trial's objective, i.e. testing the drug in a specific clinical condition.

There are two main factors complicating this matching process:

- (Full) text searches are costly database operations.
- A single drug may have several synonyms (including brand names, well known alternative names and mixed components) leading to a multiplication of search terms.

To overcome these obstacles, an Apache Lucene based workflow was implemented. As it is shown in Fig. 2, a Lucene index is populated with pairs composed of intervention id and intervention description from the TrialDB. Similarly, lists of drug ids and corresponding designators are extracted from the DrugDBs. Each drug designator is then used to execute a query on the Lucene index, yielding pairs of intervention and drug IDs, where the given intervention's text contains a designator of the given drug. The links identified between interventions and drugs are exported in form of an SQL-dump and, for performance reasons, embedded as a N:M join table into the TrialDB.

Behind the mediation engine implemented within the virtual data services the independent databases are hidden from the user. Incorporating the links found in the matching step, the mediation schema defines a uniform, global view on the distinct drug and trial databases. The client is able to query the global view, e.g. by limiting the scope to a specific set of trials, and benefits from the links found in the matching step. Using well-known SQL mechanics it is possible to explore the established knowledge base and to discover relevant relationships. At the same time the client does not need to be aware of the complexity of the integration or the physical layout of the underlying data sources.

3.3. Matching approaches

Apart from the performance gains achieved by using a Lucene index for the matching, the specialized text search facilities provided by the framework and the modular design of the workflow itself allow a refinement of the matching process.

The simplest and most obvious implementation is searching for exact matches of drug synonyms on (parts of) an intervention description (*identity matching*). This is realized using Lucene's PhraseQuery. A matching intervention text must contain a given drug designator exactly as is.

Fuzzy queries are an alternative to the exact matching. When issuing this kind of query to Lucene, a matching intervention is allowed to contain the given drug designator in a deviated form. The degree of deviation can be controlled through a fuzziness factor (essentially describing the edit distance needed to transform a word found into the search term provided). Two specific variants of fuzzy queries were used:

- Fixed factor - one (configurable) factor value (e.g. 0.8) is used for all drug designator searches.
- Fixed deviation count - the factor value is adapted to each designator's length in order to achieve a fixed edit distance over all search terms (max_dev_count / length_of_term).

Each variant allows setting a minimum term length, where the algorithm falls back to exact querying for designators that are shorter. Optionally, Lucene also allows setting a common prefix length for matches, i.e. a fuzzy matched intervention may only deviate after the common prefix from the searched designator.

Furthermore, a small number of drug designators are too generic, to serve as a search term (e.g. one brand name is "Control"). A dictionary of common English words[22] is used to filter these out.

4. Experimental evaluation

The data sources on drugs and clinical trials described above were exposed and brought to a common schema through the presented data integration framework. Due to the specific characteristics of the data, a perfect matching for reaching a 100% rate of true positive and true negative matches is not feasible. Therefore, we evaluated five different parameterizations of our matching process:

- *Identity matching*
- *Fixed factor* – 0.8 fuzziness factor, minimal term length of 6, no prefix.
- *Fixed factor* – 0.8 fuzziness factor, minimal term length of 6, common prefix of 3 characters.
- *Fixed deviation count* – edit distance of up to 2, minimal term length of 6, no prefix.
- *Fixed deviation count* – edit distance of up to 2, minimal term length of 6, common prefix of 3 characters.

As a baseline reference, we also executed a SQL equi-join resulting in a full string comparison, i.e. a drug designator must match the intervention text exactly.

4.1. Preliminary evaluation of matching methods

The type of matching method used for identifying drugs in free text description of clinical trial drug interventions has a considerable impact on the number of resulting mappings. Full string comparison requiring the intervention description being identical to the drug name, synonym, brand name or mixture name of a drug entry allows linking only 6% of clinical trial drug interventions to respective drug entries from the two drug datasets. This low number of matches is explained by the fact that free text description of interventions, as found in ClinicalTrials.gov data, tends to include additional information, such as dosing information (e.g. Doxazosin 4mg) and drug delivery form (e.g. Cefprozil 500 mg Tablets), or that multiple drugs are listed in a single expression (e.g. 1mg Glimepiride/10mg Atorvastatin FDC). Therefore, matching methods checking if a drug designator is a part of the description rather than requiring them to be equal to the description will naturally find more positive matches in this given setting. Detailed numbers on interventions linked to drugs and resulting intervention drug pairs for the different matching methods are presented in Table 1. In addition, false positive (FP) rates estimated by manually reviewing 500 randomly picked matches generated by each of the different matching methods are provided. Drug intervention pairs made up of a drug and an intervention free text description referring to that certain drug (either to that drug alone, as part of a combination therapy or as part of a list of alternative drugs tested) were considered as true positives. Whereas, pairs of a drug and intervention free text description referring to a different drug or to a placebo used instead of a certain drug were considered as false positives.

Table 1. Provided are the number of drug interventions linked to any drug, the resulting intervention drug pairs, as well as FP rate estimates for each of the different matching methods employed.

Matching method	Interventions linked	Intervention drug pairs	FP rate
Full string comparison (equi-join)	3489	5109	-
Identity matching	30811	80507	12%
Fixed factor (0.8) – prefix 3	33072	94047	19%
Fixed factor (0.8) – no prefix	33805	106514	24%
Fixed deviation count (2) – prefix 3	35663	127351	36%
Fixed deviation count (2) – no prefix	38256	239435	69%

4.2. Discussion

The intrinsic specialization of individual research domains has led to the emergence of a multitude of highly specialized content repositories. However, these repositories are rarely linked with each other leading to a high grade of fragmentation of knowledge. Our data mediation framework is specifically aiming at leveraging knowledge being trapped in different distinct data sources by providing a technological foundation for uniting them in a common virtual schema. The framework is modular by design, allowing each part to evolve separately to meet the needs of the context currently in focus. Due to the support for both the JAX-WS and the JAX-RS standards for building service-oriented applications, client applications remain loosely coupled to our framework. OGSA-DAI and our extensions to it, offer several different options on data unification, the most powerful being the mediation engine. Nonetheless, the option of using more low-level OGSA-DAI components remains and enables the incorporation of different types of data sources, e.g. RDF or XML databases.

Public domain databases on medical drugs and clinical studies are highly specialized databases in principle both holding information being of relevance in the context of one another. Using the deployment tools available within the VCE service provisioning environment, exposing the individual drug and trial databases as data services was straightforward. Also, at least for the two medical drug databases, creation of the XML based configuration files defining the virtual, global schema itself, and the mappings from the virtual relations to the actual relations from the underlying databases was directly amenable. However, upon closer inspection of the clinical trial data it became evident that defining the mapping for the virtual relation linking clinical studies with drugs was not obvious, since an equi-join was not sufficient in this case. Therefore, before final integration became feasible, an intermediate step for identifying respective drugs based on the study intervention description was required. The evaluation of exact and approximate string matching methods showed that requiring the drug name to be a substring of the study intervention is the favorable procedure allowing linking eight times more interventions to drugs as opposed to conventional string equality approach usually used as join criteria. Keeping in mind that the intervention description in ClinicalTrials.gov is a free text field and that it typically not only contains the drug name, but also dosage information, the type of delivery as well as further comments, this gain is not surprising. Nevertheless, 12% of all study drug pairs generated by using this approach were false positive associations. This finding can be explained by the fact that, requiring the drug designator to be a substring of the intervention description, inevitably leads to false positive associations when a drug name is a substring of another drug name. Relaxing the matching criteria even further by just requiring approximate matches and, thereby, compensating for typos and spelling differences, allows recovering even more matches between study interventions and drugs. However, those additional matches come at the cost of high false positive rates. The majority of false positives are attributed to the drug designators being similar to each other. For example, the two drug names Mepivacaine and Bupivacaine, obviously different strings, are similar to each other in terms of the naive edit distance. The same also holds true for the example of Cyclosporine and Cycloserine. However, while in the case of the former example the use of an additional prefix filter requiring both words to have a common prefix would prevent this false positive match, it would not have any effect on the second example. Even if the approximate string matching allowed assigning ten times more interventions to at least one drug than the equi-join, for 25% of all drug interventions still no corresponding drug was identified. The reasons for this finding are multifaceted, ranging from typos in intervention names (e.g. Wafarin instead of Warfarin), errors in the data (e.g. instead of the actual medication just the letter A is used in the intervention field to refer to the drug name in the further study description text), general descriptions not referring to a specific medication (e.g. usual long term cardiac medications) or the fact that an evaluated agent is not present in any of the medical drug databases (e.g. Volasertib). The data quality issue is not unique to the clinical study data. At least in one of the medical drug databases, synonyms made up of a single character were found. Even though the integration of drugs and trials was hampered by technical and data quality issues, the resulting integrated data set available through the data mediation service opens up new possibilities. For example, the task of identifying all drugs for which bronchitis is not mentioned in their respective indication description, but are evaluated in clinical trials on bronchitis can now be achieved with a single SQL query.

The here presented data mediation framework extends the range of already available software systems. Among them BioMart[23] is the most prominent in the area of bioinformatics for providing unified access to geographically distributed data sources. The strength of our solution lies in the fact that (1) individual databases can be directly exposed as data services; (2) the mediation schema used to centrally integrate these multiple data sources is specified in a declarative fashion, and (3) data are always fetched on-the-fly from the corresponding data service. The possibility to quickly add another data source to a common global schema or to adapt it to new requirements is a key requirement in a setting where multiple new data sources are created each year. Being able to easily create and host data services not only reduces the work required by database owners, but also ensures that they always have full control over their own data and as such significantly reduces the entry barrier to collaborative efforts. Data integration in this manner may finally allow putting the data currently trapped in individual domain specific databases into context, thereby rendering analysis concepts resting on an integrative perspective such as system biology or systems medicine more amenable.

5. Related work

Over the last couple of years, multiple projects emerged aiming to build a data infrastructure in order to tackle the issue of interlinking biomedical data. Many different approaches, including data mediation, data warehousing and semantics based on the Linked Data concept were applied.

The @neurIST project[24] created a data infrastructure based on data and mediation services aimed to study cerebral aneurisms on a relational basis. The project focused on challenging tasks of managing heterogeneous data sources, while taking into account their privacy and security. The European project Hypergenes[25] followed a data warehouse approach exposing data sources through semantic access, while Linking Open Drug Data (LODD)[26] adapted the Linked Data methodology in order to provide users a portal for navigation through vast amount of drug-related data. The linkage in LODD is based on RDF triples. These projects follow a common goal of unifying biomedical data. The VCE allows us, however, to offer users several perspectives onto the data. This is achieved by providing coarse- and fine-grained views on the data sources, as well as their semantic representation. In addition to the data infrastructure, we have implemented an application, which will allow an analysis of the relations and thus the impacts of drugs to the results of clinical trials.

The BioMart data management software and data mining toolset originated from the Ensembl project and it is focused on biological databases. Similar to our approach it enables the integration of independent, distributed datasets that are exposed as services. In contrast to the VCE approach, BioMart resembles a distributed *data warehouse* and individual datasets are required to be transformed into a predefined, query-optimized relational schema. Queries are expressed as a combination of requested attributes and filters to be applied on them, while the VCE accepts plain SQL queries.

6. Conclusion and future plans

In this paper we focused on using the VCE data integration framework for exposing and mediating drugs and clinical trial data sources for establishing relations between these two kinds of databases. Our example implementation represents the foundation for achieving multiple objectives such as enabling a unified and multi-modal access to data sources, semantic access to data, and management of the services. The architecture is structured into several tiers, and, thanks to the modular implementation, it allows an extension of any of these tiers.

Improvements may include semantic access (RDF) to integrated data, technical optimization to bolster the performance and an extended matching process to enhance linking of datasets. We will also consider exposing a refined matching process as a workflow in the VPH-Share project, to ease and encourage reuse of our components for similar use cases.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme under grant agreement #269978 (VPH-Share Project) and under grant agreement #202222 (Predict-IV).

References

1. Schneider MV, Orchard S. Omics technologies, data and bioinformatics principles. *Methods in Molecular Biology* 2011;**719**:3–30.
2. Heinzel A, Fechete R, Söllner J, Perco P, Heinze G, Oberbauer R, et al. Data Graphs for Linking Clinical Phenotype and Molecular Feature Space. *International Journal of Systems Biology and Biomedical Technologies* 2012;**1**:11–25.
3. Fernández-Suárez XM, Galperin MY. The 2013 Nucleic Acids Research Database Issue and the online molecular biology database collection. *Nucleic Acids Research* 2013;**41**:D1–7.
4. Heinzel A, Fechete R, Mühlberger I, Perco P, Mayer B, Lukas A. Molecular models of the cardiorenal syndrome. *Electrophoresis* 2013;**34**:1649–56.
5. Wiesinger M, Mayer B, Jennings P, Lukas A. Comparative analysis of perturbed molecular pathways identified in in vitro and in vivo toxicology studies. *Toxicology in Vitro : an International Journal Published in Association with BIBRA* 2012;**26**:956–62.
6. Köhler M, Benkner S. VCE - A Versatile Cloud Environment for Scientific Applications. In: The Seventh International Conference on Autonomic and Autonomous Systems (ICAS 2011). Kyoto; 2011.
7. Virtual Physiological Human: Sharing for Healthcare - A Research Environment. http://www.vph-share.eu/, accessed: June 2013.
8. Benkner S, Bisbal J, Engelbrecht G, Hose R, Kaniovskyi Y, Koehler M, et al. Towards collaborative data management in the VPH-Share Project. In: Proceedings of the Intl. Workshop on Cloud Computing Projects and Initiatives, in conjunction with Euro-Par 2011. Bordeaux; 2011.
9. Koehler M, Knight R, Benkner S, Kaniovskyi Y, Wood S. The VPH-Share data management platform: Enabling collaborative data management for the virtual physiological human community. In: Proceedings of the 8th International Conference on Semantics, Knowledge & Grids. Beijing; 2012.
10. Benkner S, Borckholder C, Bubak M, Kaniovskyi Y, Knight R, Koehler M, et al. A Cloud-based framework for collaborative data management in the VPH-Share Project. In: Proceedings of the Intl. Workshop on Cloud Computing Projects and Initiatives, in conjunction with AINA. Barcelona; 2013.
11. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, et al. DrugBank 3.0: a comprehensive resource for "omics" research on drugs. *Nucleic Acids Research* 2011;**39**:D1035–41.
12. Bernthaler A, Mönks K, Mühlberger I, Mayer B, Perco P, Oberbauer R. Linking molecular feature space and disease terms for the immunosuppressive drug rapamycin. *Molecular Biosystems* 2011;**7**:2863–71.
13. Antonioletti M, Atkinson M, Baxter R, Borley A, Chue Hong NP, Collins B, et al. The design and implementation of Grid database services in OGSA-DAI. *Concurrency and Computation: Practice and Experience* 2005;**17**:357–76.
14. Chervitz SA, Deutsch EW, Field D, Parkinson H, Quackenbush J, Rocca-Serra P, et al. Data standards for Omics data: the basis of data sharing and reuse. *Methods in Molecular Biology* 2011;**719**:31–69.
15. Sheth AP, Larson JA. Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys* 1990;**22**:183–236.
16. Lenzerini M. Data integration. In: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems - PODS '02. New York; 2002.
17. Rifaie M, Kianmehr K, Alhajj R, Ridley MJ. Data warehouse architecture and design. In: 2008 IEEE International Conference on Information Reuse and Integration. Las Vegas; 2008.
18. Ullmann J. Information Integration using Logical Views. In: Proceedings of the 6th International Conference on Database Theoryvol. 1186. Berlin; 1997.
19. Buil-Aranda C, Arenas M, Corcho O. Semantics and Optimization of the SPARQL 1.1 Federation Extension. In: Proceedings of the 8th extended semantic web conference on the semanic web: research and applications - Volume Part II. Berlin; 2011.
20. Zhu F, Shi Z, Qin C, Tao L, Liu X, Xu F, et al. Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Research* 2012;**40**:D1128–36.

21. ClinicalTrials.gov. http://www.clinicaltrials.gov, accessed: June 2013.
22. Spell Checker Oriented Word Lists. http://wordlist.sourceforge.net/, accessed: June 2013.
23. Kasprzyk A. BioMart: driving a paradigm change in biological data management. *Database : the Journal of Biological Databases and Curation* 2011;**2011**:bar049.
24. Rajasekaran H, Hasselmeyer P, Iacono L, Fingberg J, Summers P, Benkner S, et al. @neurIST - Towards a System Architecture for Advanced Disease Managment through Integration of Heterogeneous Data, Computing, and Complex Processing Services. In: IEEE International Symposium on Computer-Based Medical Systems. Jyväskylä; 2008.
25. European Project Hypergenes. http://www.hypergenes.eu, accessed: June 2013.
26. Linking Open Drug Data (LODD). http://www.w3.org/wiki/HCLSIG/LODD, accessed: June 2013.