SchizConnect: Virtual Data Integration in Neuroimaging

Jose Luis Ambite[1(✉)], Marcelo Tallis[1], Kathryn Alpert[2], David B. Keator[3], Margaret King[4], Drew Landis[4], George Konstantinidis[1], Vince D. Calhoun[4,5], Steven G. Potkin[3], Jessica A. Turner[4,6], and Lei Wang[2]

[1] University of Southern California, Los Angeles, CA, USA
{ambite, tallis, konstant}@isi.edu
[2] Northwestern University, Chicago, IL, USA
{k-alpert, leiwangl}@northwestern.edu
[3] University of California, Irvine, CA, USA
{dbkeator, sgpotkin}@uci.edu
[4] Mind Research Network, Albuquerque, NM, USA
{mking, dlandis, vdcalhoun}@mrn.org
[5] University of New Mexico, Albuquerque, NM, USA
[6] Georgia State University, Atlanta, GA, USA
jturner63@gsu.edu

Abstract. In many scientific domains, including neuroimaging studies, there is a need to obtain increasingly larger cohorts to achieve the desired statistical power for discovery. However, the economics of imaging studies make it unlikely that any single study or consortia can achieve the desired sample sizes. What is needed is an architecture that can easily incorporate additional studies as they become available. We present such architecture based on a virtual data integration approach, where data remains at the original sources, and is retrieved and harmonized in response to user queries. This is in contrast to approaches that move the data to a central warehouse. We implemented our approach in the SchizConnect system that integrates data from three neuroimaging consortia on Schizophrenia: FBIRN's Human Imaging Database (HID), MRN's Collaborative Imaging and Neuroinformatics System (COINS), and the NUSDAST project at XNAT Central. A portal providing harmonized access to these sources is publicly deployed at schizconnect.org.

Keywords: Data integration · Neuroimaging · Mediation · Schema mappings

1 Introduction

The study of complex diseases, such as Schizophrenia, requires the integration of data from multiple cohorts [1]. As a result, over the past decade we have witnessed the creation of many multi-site consortia, such as the Functional Biomedical Informatics Research Network (FBIRN) [2], the Mind Clinical Imaging Consortium (MCIC) [3], or

© Springer International Publishing Switzerland 2015
N. Ashish and J.-L. Ambite (Eds.): DILS 2015, LNBI 9162, pp. 37–51, 2015.
DOI: 10.1007/978-3-319-21843-4_4

the ENIGMA Network [4]. Within a consortium, researchers strive to harmonize the data. For example, FBIRN's Human Imaging Database (HID) [5] is a multi-site federated database where each site follows the same standard schema. However, across consortia harmonizing the data remains a challenge.

One approach to data integration, commonly called the *warehouse* approach, is to create a centralized repository with a uniform schema and data values. Data providers transform their data to the warehouse schema and formats, and move the data to the repository. An example of this approach within neuroscience is the National Database for Autism Research (NDAR) [6]. The warehouse approach is common in industry and in government and provides several advantages. The main ones are performance and stability. Since the data has been moved to a single repository, often a relational database, or other systems that allow for efficient query access, the performance of the system can be optimized by the addition of indices and restructuring of the data. Also, since the repository holds a copy of the original data, the life of the data can persist beyond the life of the original data generator. However, these strengths turn into disadvantages in more dynamic situations. First, the data in the warehouse is only as recent as the last update, so this approach may not be appropriate for data that is updated frequently. A more insidious problem is that once the schema of the warehouse has been defined and the data from the sources transformed and loaded under such schema, it becomes quite costly to evolve the warehouse if additional sources require changes to the schema.

An alternative approach to data integration, commonly called the *virtual data integration* or mediation approach, is to leave the data at the original sources, but map the source data to a harmonized virtual schema. These schema mappings are described declaratively by logical formulas. When the user specifies a query (expressed over the harmonized schema), the data integration system (also called a *mediator*) consults the schema mappings to identify the relevant data sources and to translate the query into the schemas used by each of the data sources. In addition, the system generates and optimizes a distributed query evaluation plan that accesses the sources and composes the answers to the user query. This approach has opposite advantages and disadvantages to the warehouse approach. The main advantages are data recency, ease of incorporation of new sources, and ease of restructuring the virtual schema. The user always gets the most recent data available since the answers to the user query are obtained live from the original data sources. Adding a new data source or changing the harmonized schema is accomplished by defining a set of declarative schema mappings. This process is often much simpler than reloading and/or restructuring a large warehouse. The fact that the schema mappings are a set of compact logical rules significantly lowers the cost of developing, maintaining and evolving the system. Conversely, a disadvantage of this system is that query performance generally cannot match that of a warehouse, since optimization options available in the centralized setting of a warehouse cannot be used in a distributed system. Nonetheless, as we will show in this paper, the virtual mediation approach can provide adequate performance.

Finally, the warehouse and the virtual data integration approaches are not mutually exclusive. The system can materialize the most stable data, but query in real time the data that changes more frequently.

For SchizConnect virtual data integration was preferable to data warehousing. First, it requires significantly less resources; essentially, just developing the web portal/query interface and hosting the mediator engine. There is no need for us to store and take care of large datasets locally. Second, it demands a minimum effort to integrate new data sources. In order to encourage data providers to participate in SchizConnect we required an approach that imposed minimum overhead to them. Finally, it does not require data providers to relinquish control of their data. Different data providers have different policies regarding data sharing and the virtual integration approach allows them to keep full control of who can access their data. Our mediator architecture allows for data sources to grant authorization to individual data requests based on the user's security credentials.

In this the paper we present how the virtual data integration approach has been applied to create the SchizConnect system, which is publicly available at www.schizconnect.org. First, we describe the data sources that have currently been integrated. Second, we present the behavior of the system from a user perspective, as an investigator interacting with the SchizConnect web portal. Third, we provide a technical description of the SchizConnect mediator process, including the definition of the harmonized schema, the schema mappings, the data value mappings, the query rewriting process, and the distributed query evaluation. Fourth, we provide some experimental results. Finally, we discuss related work, future work and conclusions.

2 Participating Data Sources

Currently, the SchizConnect system provides integrated access to the following sources of schizophrenia data, including demographics, cognitive and clinical assessments, and imaging data and metadata. These sources are also publicly available and have been extensively curated, documented, and subjected to quality assurance.

FBIRN Phase II @ UCI, http://fbirnbdr.nbirn.net:8080/BDR/ [2]. This study contains cross-sectional multisite data from 251 subjects, each with two visits. Data include structural and functional magnetic resonance imaging (sMRI, fMRI) scans collected on a variety of 1.5T and 3T scanners, including Sternberg Item Recognition Paradigm (SIRP) and Auditory Oddball paradigms, breath-hold and sensorimotor tasks. The data is stored in the HID system [5], which is powered by a PostgresSQL relational database located at the Univesity of California, Irvine. The SchizConnect mediator accesses HID using standard JDBC.

NUSDAST @ XNAT Central, central.xnat.org/REST/projects/NUDataSharing [7]. The Northwestern University Schizophrenia Data and Software Tool (NUSDAST) contains data from 368 subjects, the majority with longitudinal data ($\sim$2 years apart), include sMRI scans collected on a single Siemens 1.5T Vision scanner. The data is stored in XNAT central, a public repository of neuroimaging and clinical data, hosted at Washington University at Saint Louis. The site is built over the eXtensible Neuroimaging Archiving Toolkit (XNAT), a popular framework for neuroimaging data [8]. XNAT provides a REST web service interface. The mediator uses the search API, which accepts queries in an XNAT-specific XML format and returns results as a XML document.

COBRE & MCICShare @ COINS Data Exchange, coins.mrn.org [9]. The Collaborative Imaging and Neuroinformatics System (COINS), contains data from 198 and 212 subjects from the COBRE and MCICShare projects, respectively. Data for COBRE include sMRI and rest-state fMRI scans collected on a single 3T scanner. Data for the multisite MCICShare include sMRI, rest-state fMRI and dMRI scans, collected on 1.5T and 3T scanners. COINS required special handling in SchizConnect because the native COINS architecture involves dynamic data packaging following the query, which does not allow for data to be immediately returned to the query engine. With permission from the COINS executive committee, we duplicated the COINS data relevant to SchizConnect in a relational MySQL database at USC/ISI.

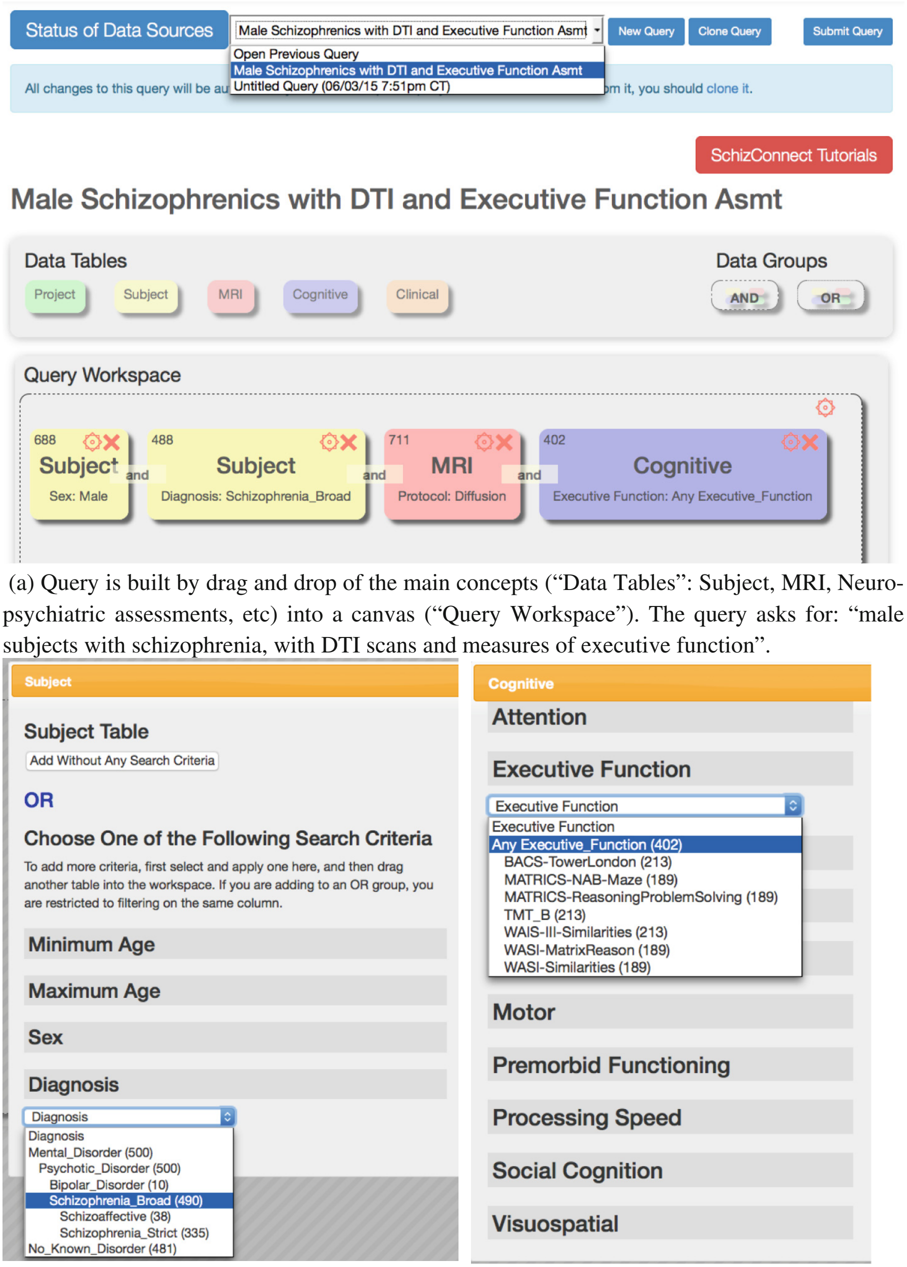
SchizConnect is positioned to become the largest neuroimaging resource for Schizophrenia, currently providing access to over 21 K images for over 1 K subjects, and expected to significantly grow as new sources are federated into the system.

3 The SchizConnect Web Portal

To understand the SchizConnect approach, it is best to start with the user experience at its web portal, schizconnect.org. The portal provides an intuitive graphical interface for investigators to query schizophrenia data across sources.

Consider a query for “male subjects with schizophrenia with DTI scans and measures of executive function”. An investigator constructs such query graphically by drag-and-drop of the main harmonized concepts into a canvas (Fig. 1(a)). Currently the supported concepts include Subject, MRI, Neuropsychiatric Assessments, and Clinical Assessments. Each concept has a number of attributes on which the user can make selections. Figure 1(b) shows the attributes of Subject, which include age, sex, and diagnosis, and a selection on the diagnosis attribute for subjects with schizophrenia in a broad sense. The values for diagnosis have a hierarchical structure and have been harmonized across the sources. In Sect. 4.3 we describe how the SchizConnect mediator classifies the subjects into these categories. Figure 1(c) shows the cognitive assessment concept (Neuropsych) and a selection on measures of executive function.

The results to this query appear in Figs. 2 and 3. The SchizConnect Portal shows the number of subjects, scans, and assessments that satisfy the query constraints, as well as a breakdown of the provenance of the data (Fig. 2). In this case, 117 images from 58 subjects come from the COBRE data source and 169 images from 82 subjects from MCICShare data source, for a total of 286 images and 6 distinct cognitive assessments of executive function for 140 subjects. Any investigator can obtain these



All changes to this query will be au … m it, you should clone it.

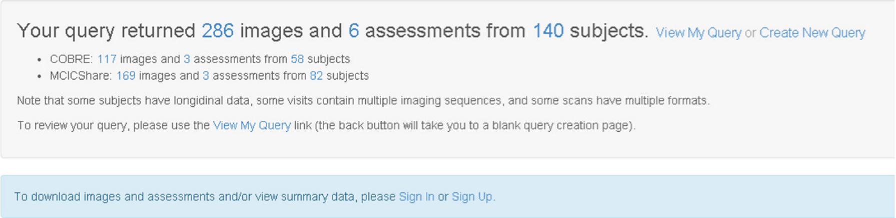
To add more criteria, first select and apply one here, and then drag another table into the workspace. If you are adding to an OR group, you are restricted to filtering on the same column.

(a) Query is built by drag and drop of the main concepts ("Data Tables": Subject, MRI, Neuropsychiatric assessments, etc) into a canvas ("Query Workspace"). The query asks for: "male subjects with schizophrenia, with DTI scans and measures of executive function".

(b) Selecting Subjects with a diagnosis of Schizophrenia in a broad sense.

(c) Selecting subjects with cognitive assessments of executive function

Fig. 1. Schizconnect portal: sample query using the harmonized schema and terminology. Each concept presents different attributes, some of which take hierarchical values, according to the SchizConnect harmonized terminologies.



Your query returned 286 images and 6 assessments from 140 subjects. View My Query or Create New Query

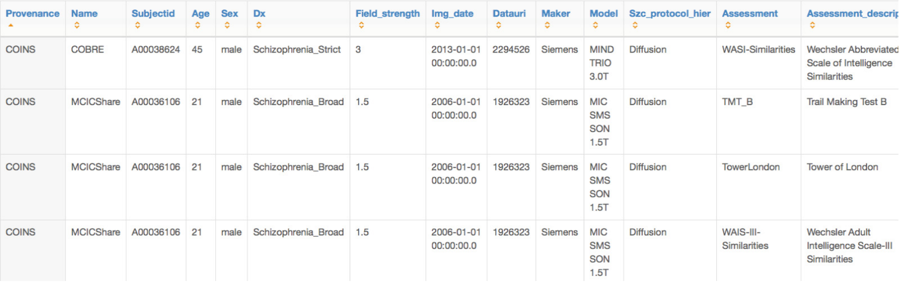
- COBRE: 117 images and 3 assessments from 58 subjects
- MCICShare: 169 images and 3 assessments from 82 subjects

Note that some subjects have longidinal data, some visits contain multiple imaging sequences, and some scans have multiple formats.

To review your query, please use the View My Query link (the back button will take you to a blank query creation page).

To download images and assessments and/or view summary data, please Sign In or Sign Up.

Fig. 2. The results of the query from Fig. 1. The user can then proceed to request the data from the different repositories.



Provenance	Name	Subjectid	Age	Sex	Dx	Field_strength	Img_date	Datauri	Maker	Model	Szc_protocol_hier	Assessment	Assessment_descrip
COINS	COBRE	A00038624	45	male	Schizophrenia_Strict	3	2013-01-01 00:00:00.0	2294526	Siemens	MIND TRIO 3.0T	Diffusion	WASI-Similarities	Wechsler Abbreviated Scale of Intelligence Similarities
COINS	MCICShare	A00036106	21	male	Schizophrenia_Broad	1.5	2006-01-01 00:00:00.0	1926323	Siemens	MIC SMS SON 1.5T	Diffusion	TMT_B	Trail Making Test B
COINS	MCICShare	A00036106	21	male	Schizophrenia_Broad	1.5	2006-01-01 00:00:00.0	1926323	Siemens	MIC SMS SON 1.5T	Diffusion	TowerLondon	Tower of London
COINS	MCICShare	A00036106	21	male	Schizophrenia_Broad	1.5	2006-01-01 00:00:00.0	1926323	Siemens	MIC SMS SON 1.5T	Diffusion	WAIS-III-Similarities	Wechsler Adult Intelligence Scale-III Similarities

Fig. 3. An excerpt individual-level results of the query from Fig. 1. To obtain individual level results the user needs to sign the appropriate data sharing agreements.

summary counts by visiting the schizconnect.org portal. After an investigator registers, logs into the system, and signs the data sharing agreements of the data providers, she can also retrieve the individual-level data, which include summary tables (Fig. 3), as well as links to download the images and full cognitive assessments for the selected subjects. The system remembers previously signed agreements and asks the investigator to sign additional ones when her query requires data from additional sources.

4 The SchizConnect Mediator

The SchizConnect Web Portal presents a unified view of the data at the different sources, as if it was coming from in a single database. However, the data is not stored at the portal, but it remains at the original sources, structured under their original schemas. The SchizConnect mediator provides a *virtual* harmonized schema, over which the portal issues queries. Given a user query, over the harmonized schema, the mediator determines which sources have relevant data, translates the user query to the schemas of the sources, and constructs, optimizes, and executes a distributed query evaluation plan that computes the answers to the user query by accessing the data sources in real time. The SchizConnect mediator builds upon the BIRN Mediator [10]. In this section, we describe each of the components of the mediator that make this data harmonization and query processing possible.

4.1 SchizConnect Domain Schema

In order to integrate data from disparate sources, we need to understand the semantics of the data, and how different schema elements at different sources related to other elements. The common approach to specific such semantics is to map the schema of each source to a common harmonized schema (also called the target, or domain, or global schema) [11]. This common schema is a degree of freedom for the designer of the integration system.

```
project(provenance, name, projectid, description)
subject(provenance, subjectid, age, sex, dx)
in_project(provenance, subjectid, projectid)
imaging_protocol( provenance, subjectid, szc_protocol, img_date, notes,
                  datauri, maker, model, field_strength)
cognitive_assessment(provenance, study, subjectid, szc_assessment, description)
cognitive_assessment_data( provenance, study, subjectid, szc_assessment,
                           question_id, question_value)
clinical_assessment(provenance, study, subjectid, szc_assessment, description )
clinical_assessment_data( provenance, study, subjectid, szc_assessment,
                          question_id, question_value)
```

Fig. 4. SchizConnect current domain model.

It does not need to include every schema element present in the sources; just those elements useful for the purposes of the integration problem at hand. The design of the common schema is a balance between minimalism, that is, only include elements that exist in the sources and that are needed to answer the current query load, and generality, that is, a schema design that can easily be extended to model additional sources and query types. Our philosophy leans towards minimalism. Instead of attempting to model the neuroimaging domain wholesale, we build the common schema incrementally as we find sources that provide data for the desired concepts in the domain.

The current domain schema in SchizConnect follows the relational model and is composed of the following predicates (Fig. 4):

Project contains the name and description of the studies in the data sources.

Subject contains demographic and diagnostic information for individual participants, including “subject id”, “age”, “sex” and “diagnosis”.

Imaging Protocol (MRI) contains information on MRIs a subject has, including the type of the scan and metadata about the scanner. The values of the protocol attribute are organized hierarchically (cf. Sect. 4.3).

Cognitive Assessment contains information on which subjects have which neuropsychological assessments. The values of the “assessment” attribute are also organized hierarchically (cf. Sect. 4.3).

Cognitive Assessment Data contains full information on the assessments including the values for each measure in each assessment for each subject.

Clinical Assessment and **Clinical Assessment Data** contain assessments for different symptoms in the subjects.

The first attribute in each of the domain predicates is “provenance”, which records which source provided the data elements (see Fig. 3).

4.2 SchizConnect Schema Mappings

The SchizConnect domain predicates, shown in Fig. 4, provide a consistent view of the data available from the sources. However, the mediator does not pre-compute such data as in a warehouse, but obtains these data on-the-fly from the sources at query time. For this process, the mediator uses a set of declarative schema mappings, which define how predicates from the source schema relate to predicates in the domain schema. These mappings are usually logical implications of the form:

$$\forall \vec{X}, \vec{Y}, \Phi_S(\vec{X}, \vec{Y}) \rightarrow \exists \vec{Z}, \Psi_G(\vec{X}, \vec{Z})$$

with a conjunctive antecedent (Φ_S) over predicates from the source schemas (S), and a conjunctive consequent (Ψ_G) over predicates from the domain schema (G). These mappings are also known as source-to-target tuple-generating dependencies (st-tgds) in the database theory literature [12]. The SchizConnect mediator supports full conjunctive st-tgds (aka GLAV rules) [13], but so far the domain and schema mappings we have developed only needed to be Global-as-View (GAV) rules [11], which are st-tgds with a single predicate in the consequent.

Some sample schema mappings appear in Fig. 5. We use a logical syntax for the rules. We show domain predicates in bold (e.g., **subject**) and source predicates in italics (e.g., *HIDPSQLResource_nc_subjexperiment*). The first rule states that the source XNAT provides data for subjects. More precisely, that invoking the source predicate *XnatSubjectResource_xnat__subjectData*, and then joining the results with the *MappingsMySQLResource_dx_mappings* source predicate (which are located at different sources, XNAT and a MySQL db), yields the domain predicate **subject**. A shared variable in the antecedent of a rule (e.g., SRC_DX) denotes an equi-join condition. Other type of conditions can be included in antecedents by adding relational predicates (e.g., the selection ‘nc_experiment_uniqueid = 9610’ in the fourth rule). Variables in the consequent denote projections over data sources.

Rules with the same consequent denote *union*. For example, in Fig. 5 the domain predicate **subject** is obtained as the union of three rule, one for each data source (XNAT, COINS, and HID). Note how each of the rules includes a constant in the consequent to denote the provenance of the data (i.e., “XNAT”).

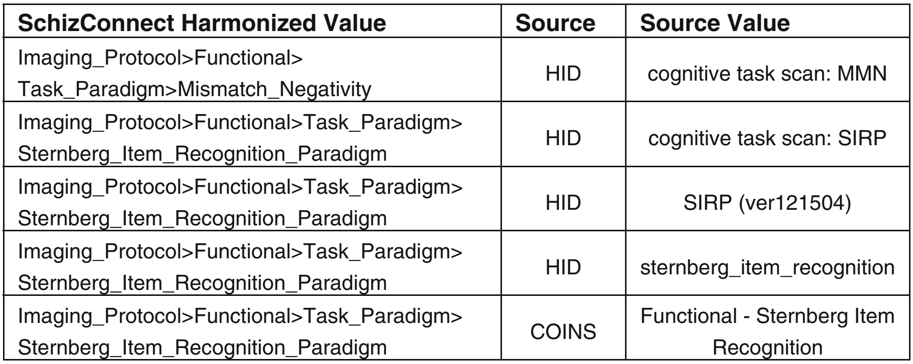
Our mediator language allows for non-recursive logic programs. For example, the third rule in Fig. 5 states that the **subject** domain predicate for HID is constructed by the *join* of 3 domain predicates: ***subject_age***, ***subject_sex***, and ***subject_dx.*** The next two rules show how the diagnoses for the subjects (***subject_dx***) in the HID source are calculated based on specific values for the assessments as stored in the original HID tables. For example, a subject with values of 3 and 1 in questions P47 and P53 of the SCID assessment, resp., is assigned a diagnosis of schizophrenia in the strict sense.

Finally, the last two rules show how to obtain the **imaging_protocol** domain predicate for the HID and XNAT sources. Normalization of the imaging protocol and

```
subject("XNAT", SUBJECT_ID, AGE, SEX, DX) <-
   XnatSubjectResource_xnat__subjectData(project, SUBJECT_ID, AGE, SEX, SRC_DX, QS) ^
   MappingsMySQLResource_dx_mappings(DX, "NUSDAST", 777, SRC_DX, id)

subject("COINS", SUBJECT_ID, AGE, SEX, DX) <-
   COINSMySQLResource_subjects_v( SUBJECT_ID, SEX, yob, SRC_DX, STUDY_ID, AGE) ^
   MappingsMySQLResource_dx_mappings(DX, "COINS", STUDY_ID, SRC_DX, id)

subject("HID", SUBJECTID, AGE, SEX, DX) <-
   subject_age("HID", SUBJECTID, AGE) ^ subject_sex("HID", SUBJECTID, SEX) ^
   subject_dx("HID", SUBJECTID, DX)
…
subject_dx("HID",SUBJECTID, 'No_Known_Disorder') <-
   HIDPSQLResource_nc_subjexperiment( uniqueid, tableid, owner, modtime, moduser,
            nc_experiment_uniqueid, SUBJECTID, nc_researchgroup_uniqueid) ^
   (nc_researchgroup_uniqueid IN [9612,4292] ) ^ (nc_experiment_uniqueid = 9610)
…
subject_dx("HID",SUBJECTID,
'Mental_Disorder>Psychotic_Disorder>Schizophrenia_Broad>Schizophrenia_Strict') <-
   HIDPSQLResource_nc_subjexperiment( uniqueid, tableid, owner, modtime, moduser,
            nc_experiment_uniqueid, SUBJECTID, nc_researchgroup_uniqueid) ^
   (nc_researchgroup_uniqueid = 9611 ) ^ (nc_experiment_uniqueid = 9610) ^
   HIDPSQLResource_nc_assessmentinteger( tableid1, nc_assessmentdata_uniqueid1,
            scoreorder1, owner1, modtime1, moduser1, textvalue1, textnormvalue1,
            comments1, DATAVALUE1, datanormvalue1, storedassessmentid1,
            ASSESSMENTID1, SCORENAME1, scoretype1, ISVALIDATED1, isranked1,
            SUBJECTID, entryid1, keyerid1, raterid1, classification1, uniqueid1) ^
   (ASSESSMENTID1 = 16415) ^ (SCORENAME1 = "SCID_P47") ^ (DATAVALUE1 = 3) ^
   HIDPSQLResource_nc_assessmentinteger( tableid2, nc_assessmentdata_uniqueid2,
            scoreorder2, owner2, modtime2, moduser2, textvalue2, textnormvalue2,
            comments2, DATAVALUE2, datanormvalue2, storedassessmentid2,
            ASSESSMENTID2, SCORENAME2, scoretype2, ISVALIDATED2, isranked2,
            SUBJECTID, entryid2, keyerid2, raterid2, classification2, uniqueid2) ^
   (ASSESSMENTID2 = 16415) ^ (SCORENAME2 = "SCID_P53") ^ (DATAVALUE2 = 1) ^
   (ISVALIDATED1 = "TRUE") ^ (ISVALIDATED2 = "TRUE")

imaging_protocol("HID", SUBJECTID, SZC_PROTOCOL_HIER, DATE, NOTES, DATAURI,
                        MAKER, MODEL, FIELD_STRENGTH) <-
   HIDPSQLResource_nc_scannersbyscan ( SUBJECTID, componentid, segmentid,
            SOURCE_PROTOCOL, DATE, nc_colequipment_uniqueid, SOURCE_MAKE,
            SOURCE_MODEL, DATAURI, NOTES) ^
   MappingsMySQLResource_protocol_mappings( SZC_PROTOCOL_HIER, "HID",
                                            SOURCE_PROTOCOL, ID1) ^
   MappingsMySQLResource_scanner_mappings( MAKER, MODEL, FIELD_STRENGTH, "HID",
                                            SOURCE_MAKE, SOURCE_MODEL, ID2)

imaging_protocol("XNAT", SUBJECTID, SZC_PROTOCOL_HIER, DATE, SCAN_ID,
                              DATA_URI, "SIEMENS", "VISION 1.5T", 1.5) <-
   XnatMRSessionResource_xnat__mrSessionData( SUBJECTID, IMAGE_ID, SESSION_ID,
                              DATE, SCANNER, SCAN_ID, SCAN_TYPE, quarantine_status) ^
   MappingsMySQLResource_protocol_mappings( SZC_PROTOCOL_HIER, "NUSDAST",
                                            SCAN_TYPE, ID1) ^
   Concat(IMAGE_ID, "/scans/", SCAN_ID, DATA_URI)
```

Fig. 5. SchizConnect schema mappings.

scanners values is achieved by joining with additional mapping tables (e.g., *MappingsMySQLResource_protocol_mappings*). The mediator also supports functional sources, such as concatenation (*Concat* in the last rule in Fig. 5). In general, the designer can define arbitrary Java functions and use them in the schema mappings to perform complex value transformations.

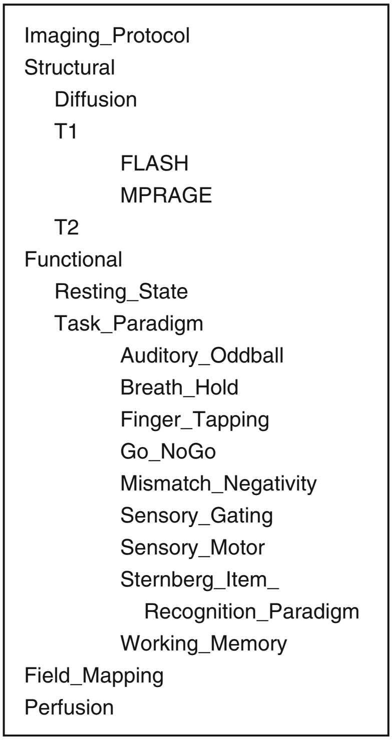


SchizConnect Harmonized Value	Source	Source Value
Imaging_Protocol>Functional> Task_Paradigm>Mismatch_Negativity	HID	cognitive task scan: MMN
Imaging_Protocol>Functional>Task_Paradigm> Sternberg_Item_Recognition_Paradigm	HID	cognitive task scan: SIRP
Imaging_Protocol>Functional>Task_Paradigm> Sternberg_Item_Recognition_Paradigm	HID	SIRP (ver121504)
Imaging_Protocol>Functional>Task_Paradigm> Sternberg_Item_Recognition_Paradigm	HID	sternberg_item_recognition
Imaging_Protocol>Functional>Task_Paradigm> Sternberg_Item_Recognition_Paradigm	COINS	Functional - Sternberg Item Recognition

Fig. 6. SchizConnect value mappings.

4.3 SchizConnect Value Mappings

In addition to mapping the schemas of the sources into the SchizConnect domain schema, we also harmonized the values for the attributes. This was achieved by developing mapping tables that relate values used in the sources with harmonized values in SchizConnect. These tables are stored in a separate relational database, which is treated as a regular data source for the mediator. For example, the source predicate *MappingsMySQLResource_protocol_mappings* stores the mappings for imaging protocols. Some sample mappings for this predicate appear in Fig. 6. Note that even within the same source, there are often several different values/codes for the same concept. For example, HID has several different codes for the Sternberg Item Recognition Paradigm protocol (since HID contains multiple substudies performed at different times, and no attempt at enforcing common values across substudies was made).



```
Imaging_Protocol
Structural
   Diffusion
   T1
         FLASH
         MPRAGE
   T2
Functional
   Resting_State
   Task_Paradigm
         Auditory_Oddball
         Breath_Hold
         Finger_Tapping
         Go_NoGo
         Mismatch_Negativity
         Sensory_Gating
         Sensory_Motor
         Sternberg_Item_
           Recognition_Paradigm
         Working_Memory
Field_Mapping
Perfusion
```

Fig. 7. Imaging protocol taxonomy

Many of the harmonized values in SchizConnect have a hierarchical structure. For example the current hierarchy for the imaging protocol appears in Fig. 7. Similar hierarchies for the diagnosis and cognitive assessments appear (partially) in Figs. 1(b) and 1(c). These hierarchies are easily extensible by updating the mapping tables.

The design of the harmonized values takes into account existing ontologies. A companion paper [16] describes in detail this design and the mapping of the SchizConnect value taxonomies to concepts in NeuroLex and other well-known ontologies.

4.4 Query Rewriting

Given a user query, the mediator uses the schema mappings defined for the application domain, to translate the query from the virtual domain schema into an executable query over the source schemas, a process called query rewriting. For GAV schema mappings, such as those in Fig. 5, query rewriting amounts to rule unfolding and simplification. We have also developed algorithms for query rewriting under LAV schema mappings [13] and GLAV rules, but they are not used in the current modeling of the SchizConnect domain.

We will describe the rewriting process by example. Consider a user query for all the available T1 scans: select * from imaging_protocol where szc_protocol like '%T1 %', and the schema mappings for **imaging_protocol** in Fig. 5. The rewritten query, expressed in SQL, appears in Fig. 8. This query is built by unfolding the definitions of imaging_protocol according to the schema mapping rules. In general, for GAV rewriting the system unifies each domain predicate with the corresponding consequent of the GAV rule (i.e., with the same predicate) and replaces it with the antecedent of the rule. After this unfolding process, the source-level queries are logically minimized to avoid probably redundant predicates (i.e., source invocations). For this simple example, the rewritten query is a union of conjunctive queries over the sources providing the data, including joins with the mapping sources to produce harmonized values, as we described in Sect. 4.3. The schema mapping for COINS and the corresponding portion of the rewriting is not shown for brevity.

4.5 Distributed Query Engine

Once the mediator has translated the user domain query into a source-level query (i.e., involving only source predicates), it must generate, optimize and execute a distributed query evaluation plan. Our current query engine is based on the Open Grid Services Architecture (OGSA) Distributed Access and Integration (DAI), and Distributed Query Processing (DQP) projects [14]. OGSA-DAI is a streaming dataflow workflow evaluation engine that includes a library of connectors to many types of common data sources such as databases and web services. Each data source is wrapped and presents a uniform interface as a Globus [15] grid web service. OGSA-DQP is a distributed query evaluation engine implemented on top of OGSA-DAI. In response to a SQL query, OGSA-DQP constructs a query evaluation plan to answer such query. The evaluation

```
(SELECT  'HID' as provenance, T6.subjectid as subjectid, T4.szc_protocol_hier as
   szc_protocol_hier, T6.date as img_date, T6.description as notes, T6.datauri as datauri,
   T2.maker as maker, T2.model as model, T2.field_strength as field_strength
FROM MappingsMySQLResource_scanner_mappings T2,
        MappingsMySQLResource_protocol_mappings T4,
        HIDPSQLResource_nc_scannersbyscan_mview T6
WHERE T2.source_make=T6.source_make AND T2.source_model=T6.source_model AND
   T2.source = 'HID' AND T4.source_protocol=T6.source_protocol AND T4.source = 'HID' AND
   T4.szc_protocol_hier LIKE '%T1%')
UNION
(SELECT 'XNAT' as provenance, T10.SUBJECT_ID as subjectid,
   T8.szc_protocol_hier as szc_protocol_hier, T10.SCAN_DATE as img_date,
   T10.SCAN_ID as notes, Concat(T10.IMAGE_ID,'/scans/',T10.SCAN_ID) as datauri,
   'SIEMENS' as maker, 'VISION 1.5T' as model, 1.5 as field_strength
FROM MappingsMySQLResource_protocol_mappings T8,
        XnatMRSessionResource_xnat__mrSessionData T10
WHERE T8.source_protocol=T10.SCAN_TYPE AND T8.source = 'NUSDAST' AND
           T8.szc_protocol_hier LIKE '%T1%')
```

Fig. 8. Executable query over the source schemas.

plan is implemented as an OGSA-DAI workflow, where the workflow activities correspond to relational algebra operations. The OGSA-DQP query optimizer partitions the workflow across multiple sources attempting to push as much of the evaluation of subqueries to remote sources. OGSA-DQP currently supports distributed SQL queries over tables in multiple sources. The OGSA-DAI/DQP architecture is modular and allows for the incorporation of new optimization algorithms, as well as mediator (query rewriting) modules, as plug-ins for new source types into the system.

We improved the OGSA-DAI/DQP query engine by adding a module to gather cost statistics from the sources, including table sizes and selectivity parameters, and by developing a cost-based query optimizer based on these statistics, as well as several other enhancements to specific optimization steps. The query plan optimizer proceeds in two phases. First, it applies a sequence of classical query plan transformations, such as pushing selection operations closer to their data sources, grouping operations on the same source and pushing subqueries to sources with query evaluation capabilities. Second, it searches how join operations can be ordered to minimize the cost of the overall plan. For complex queries, such as those described in Sect. 5 that involve conjunctive queries with 10–20 predicates, the enhanced cost-based optimizer produced plans that improved execution time by orders of magnitude.

4.6 Source Wrappers

The mediator can access sources of different types, including relational databases, such as HID, and web service APIs, such as XNAT. The actual data sources are wrapped as

OGSA-DAI resources. OGSA-DAI provides a common extensible framework to add new types of data sources.

For each non-relational source, we develop a *wrapper* that takes as input a SQL query (over predicates that encapsulate the data from the source), and translates this SQL query into the native query language of the source. Symmetrically, the wrapper takes data results from the source in their original format and converts them into relational tuples that can flow through the query engine.

For SchizConnect, we developed such a wrapper for XNAT. Consider the query:

select * from XnatMRSessionResource_xnat__mrSessionData where scan_type = 'T1'

This query invokes the wrapper for XNAT (see also the rewritten query in Fig. 8). This SQL query is translated to the native query language of the XNAT search service API, which is expressed as an XML document. The XNAT web service returns the results also as an XML document. The wrapper parses this document and translates it into relational tuples, following the schema of XnatMRSessionResource_xnat__mrSessionData. Now a uniform relational result, it is processed by the query engine as the data from any other source.

5 Experimental Results

The system is publicly deployed at SchizConnect.org. The web front-end is hosted at Northwestern University, the mediator is hosted at USC/ISI, and the sources are at UCI (the HID PostgreSQL DB), Washington University at Saint Louis (XNAT Central), and at USC/ISI (the MySQL database that hosts the replica COINS data).

Despite its nationwide distribution, the system performs well. We show some performance results for a representative set of queries in Fig. 9. The table of results is structured as follows. The first column is just the query id. The next two columns show the size of the tested domain query, and the specific predicates involved. All the tested domain queries are conjunctive. The following two columns show the structure and size of the resulting rewritten source-level query, which is generally much larger than the domain (user) query. The last two columns show the number of tuples in the answer to the user query and the total time in seconds to compute the answers (i.e., from sending the query to the mediator to returning the results to the user). For example, the fourth row shows the results for a domain query that asks for subjects with two assessments (of verbal episodic memory: HVLT-Delay and HVLT-Immediate), with two imaging protocols (T1, and sensory motor scans). The query involves the join of 7 domain predicates; namely, subject (s), in_project (ip), project (p), two instances of imaging_protocol (i), and two instances of cognitive_assessement (ca). The resulting rewritten query is a union of 5 conjunctive queries, each involving 16, 18, 17, 10, and 10 source predicates, respectively, for a total of 71 source predicates. The query returns 722 tuples and takes 12.1 s to complete.

The queries shown identify the subjects, imaging protocol, cognitive assessments, etc., satisfying the desired constraints, and return the desired data. However, the performance results in Fig. 9 do not include the transfer of the actual image files. For example, the seventh query asks for all the metadata about the 21447 imaging protocols currently accessible through SchizConnect from all the sources, which the mediator does return. However, the size of corresponding images is several hundred GBs ($\sim$173 GB compressed). So, when the user query identifies the subjects and scans of interest, SchizConnect schedules separate grid-ftp, ftp, and http connections to the original sources to obtain and package the images for the query subjects. In contrast, the cognitive and clinical assessment data are retrieved directly through the mediator, since these are smaller datasets. For example, the third query in Fig. 9, shows that asking for all the data on 13 cognitive assessments for all subjects produces a result set of 9318 tuples, which are returned in 8.9 s.

	Domain Query		Source-level Query		Result Size (#tuples)	Time (s)
	Size (#p)	Preds	Structure	Size (#p)		
1	6	ip, p, 4ca	U3CQ (10, 10,10)	30	189	7.8
2	1	s	U5CQ (4, 6, 2, 2)	14	1091	8.2
3	1	cad (13)	U2CQ (2, 3)	5	9318	8.9
4	7	s,ip,p,2i,2ca	U5CQ (16,18,17,10,10)	71	722	12.1
5	5	p,ca,i,s,ip	U5CQ (11, 13, 12,7,7)	50	1094	15.9
6	4	s,ip,p,i	U5CQ (9, 11, 10, 5, 5)	40	1462	17.3
7	1	i	U3CQ (3, 2, 1)	6	21447	18.7
8	4	s,ip,p,i	U5CQ (9,11,10,5,5)	40	19112	24.5

Fig. 9. Experimental results

The computation cost is a combination of the number final and intermediate results needed to compute the query, the number of sources involved, and the complexity of the rewritten queries, with large and more complex queries often taking more time, but not in a simple relationship.

6 Discussion

We have presented SchizConnect, a virtual data integration approach that provides semantically-consistent, harmonized access to several leading neuroimaging data sources. The mediation architecture is driven by declarative schema mappings that make the system easier to develop, maintain and extend. Our virtual approach allows the creation of large data resources at a fraction of the cost of competing approaches.

The system is publicly available at **SchizConnect.org**. Since its initial deployment in September 2014, the number of users, queries and image downloads has grown steadily (with over 50 registered users as of May 2015).

We are currently extending the coverage of different types data, specifically clinical assessments. We also plan to incorporate additional schizophrenia studies to SchizConnect. Finally, we plan to improve the underlying data integration architecture, specifically the performance of the query optimizer and adding a more expressive representational language for the domain schema, such as OWL2 QL.

Acknowledgements. SchizConnect is supported by a grant from the National Institutes of Health (NIH/NIMH), 5U01MH097435 to L. Wang, JL. Ambite, S.G. Potkin and J.A.Turner. The work on COINS is also supported by 5P20GM103472 (NIGMS) to V.D. Calhoun.

References

1. Turner, J.A.: The rise of large-scale imaging studies in psychiatry. GigaScience **3**, 29 (2014)
2. Glover, G.H., et al.: Function biomedical informatics research network recommendations for prospective multicenter functional MRI studies. J. Magn. Reson. Imaging JMRI **36**, 39–54 (2012)
3. King, M.D., Wood, D., Miller, B., Kelly, R., Landis, D., Courtney, W., Wang, R., Turner, J. A., Calhoun, V.D.: Automated collection of imaging and phenotypic data to centralized and distributed data repositories. Front. Neuroinform. **8**, 60 (2014)
4. Thompson, P.M., et al.: The ENIGMA consortium: large-scale collaborative analyses of neuroimaging and genetic data. Brain Imaging Behav. **8**, 153–182 (2014)
5. Keator, D.B., et al.: A national human neuroimaging collaboratory enabled by the biomedical informatics research network (BIRN). IEEE Trans. Inf. Technol. Biomed. **12**, 162–172 (2008)
6. Hall, D., Huerta, M.F., McAuliffe, M.J., Farber, G.K.: Sharing heterogeneous data: the national database for autism research. Neuroinformatics **10**, 331–339 (2012)
7. Wang, L., et al.: Northwestern University Schizophrenia Data and Software Tool (NUSDAST). Frontiers in Neuroinformatics **7**, 25 (2013)
8. Marcus, D.S., Olsen, T., Ramaratnam, M., Buckner, M.L.: The extensible neuroimaging archive toolkit (XNAT): An informatics platform for managing, exploring, and sharing neuroimaging data. Neuroinformatics **5**, 11–34 (2005)
9. Scott, A., Courtney, W., Wood, D., de la Garza, R., Lane, S., King, M., Wang, R., Roberts, J., Turner, J.A., Calhoun, V.D.: COINS: An innovative informatics and neuroimaging tool suite built for large heterogeneous datasets. Front. Neuroinform. **5**, 33 (2011)
10. Ashish, N., Ambite, J.L., Muslea, M., Turner, J.: Neuroscience Data Integration through Mediation: An (F)BIRN Case Study. Front. Neuroinform. **4**, 118 (2010)
11. Doan, A., Halevy, A., Ives, Z.: Principles of Data Integration. Morgan Kauffman, Waltham (2012)
12. Fagin, R., Kolaitis, P.G., Miller, R.J., Popa, L.: Data Exchange: Semantics and query answering. In: Calvanese, D., Lenzerini, M., Motwani, R. (eds.) ICDT 2003. LNCS, vol. 2572, pp. 207–224. Springer, Heidelberg (2002)
13. Konstantinidis, G., Ambite, J.: Scalable query rewriting: a graph-based approach. In: SIGMOD Conference, pp. 97–108. ACM (2011)
14. Grant, A., Antonioletti, M., Hume, A.C., Krause, A., Dobrzelecki, B., Jackson, M.J., Parsons, M., Atkinson, M.P., Theocharopoulos, E.: OGSA-DAI: Middleware for data integration: selected applications. In: Fourth IEEE International Conference on eScience (2008)
15. The Globus Project (1997). http://www.globus.org
16. Turner, et al.: Terminology development towards harmonizing multiple clinical neuroimaging research repositories. In: Proceedings of DILS 2015 (2015)