frontiers in
NEUROINFORMATICS

ORIGINAL RESEARCH ARTICLE
published: 28 December 2010
doi: 10.3389/fninf.2010.00118

Neuroscience data integration through mediation: an (F)BIRN case study

Naveen Ashish[1]*, José Luis Ambite[2], Maria Muslea[2] and Jessica A. Turner[3]

[1] Calit2, University of California at Irvine, Irvine, CA, USA
[2] Information Sciences Institute, University of Southern California, Los Angeles, CA, USA
[3] Mind Research Network, Albuquerque, NM, USA

Edited by:
Jan G. Bjaalie, University of Oslo, Norway

Reviewed by:
Daniel Gardner, Weill Cornell Medical College, USA
Luis Marenco, Yale University, USA

***Correspondence:**
Naveen Ashish, University of California at Irvine, 4308 Calit2 Building, Irvine, CA 92697, USA.
e-mail: ashish@ics.uci.edu

We describe an application of the BIRN mediator to the integration of neuroscience experimental data sources. The BIRN mediator is a general purpose solution to the problem of providing integrated, semantically-consistent access to biomedical data from multiple, distributed, heterogeneous data sources. The system follows the mediation approach, where the data remains at the sources, providers maintain control of the data, and the integration system retrieves data from the sources in real-time in response to client queries. Our aim with this paper is to illustrate how domain-specific data integration applications can be developed quickly and in a principled way by using our general mediation technology. We describe in detail the integration of two leading, but radically different, experimental neuroscience sources, namely, the human imaging database, a relational database, and the eXtensible neuroimaging archive toolkit, an XML web services system. We discuss the steps, sources of complexity, effort, and time required to build such applications, as well as outline directions of ongoing and future research on biomedical data integration.

Keywords: data integration, neuroinformatics, heterogeneous sources

INTRODUCTION

Our work in information integration is in the context of the *Biomedical Information Research Network* (BIRN)[1] and reflects a cross institutional collaboration to address a key practical problem, that of integrated information access to information in multiple sources. By integrated data access we broadly mean that a user, for instance a scientific investigator, is abstracted from the fact that data of interest to her may reside in heterogeneous and geographically distributed data sources. Rather she is able to go to a single interface and access data seamlessly *as if* it were one single harmonized data source. Database use has become pervasive in the scientific community; however the burden of integrating different kinds of information when required, is at present on the investigator. We provide data integration solutions that alleviate the investigator from such information gathering burdens.

There are many interesting aspects of our approach and experience which we believe are important and useful to share with the community at this time. First, our solution is based on a robust, grid-based, general purpose, and state of the art data integration technology developed using a *mediator* architecture and implemented using an open and *grid-enabled* framework. The solutions we are developing in BIRN have benefited from the close involvement of domain experts in areas such as neuroscience. Further, we have explored many diverse data integration applications in neuroscience, non-human primate, and cardiovascular informatics, demonstrating the general purpose applicability of our technology and approach.

In this paper we describe our technology solution and approach, taking the context of a data integration application from the *Function BIRN* (FBIRN) domain (Keator et al., 2008) which is focused on making multi-site functional MRI studies a reality.

NEEDS OF USERS AND INVESTIGATORS

Neuroscience and biomedical data resources exist in many places today, and the ability to combine data across sources requires data integration methods. Use cases such as the ones defined for mouse functional genomics (Gruenberger et al., 2010) highlight what many biomedical researchers could use: The ability to query multiple databases to find the relevant data for particular scientific questions. In neuroimaging research, for example, there have been many structural imaging studies, and some of those datasets are being made publicly available. The ability to determine how many subjects of various types with structural imaging data are available across the publicly accessible datasets would allow researchers to determine whether they already have access to data to answer their specific questions about brain structure, or whether they need to collect their own targeted data. Different sites may specialize in collecting and providing data for particular kinds of experiments and/or subsets of populations, investigators on the other hand would typically want access to information on multiple experiment kinds and a wider subject pool.

A common approach to information integration, when data exists across multiple sources, is to create a single data *warehouse* to centralize data access (Keator et al., 2009) and each warehouse takes a different approach, with its own strengths and limitations. *BrainMap*, for example[2], is a single database that stores results from the published functional neuroimaging literature. *PubBrain*[3] is a search interface that uses PubMed to identify neuroimaging papers and visualize the results. The warehousing approach has created a plethora of databases focused on different aspects of biomedical

[1] www.birncommunity.org

[2] www.brainmap.org
[3] www.pubbrain.org

research. For example the database issue of the 2010 Nucleic Acids Research journal lists 1230 available databases (Galperin and Cochrane, 2009). The neuroscience information framework (NIF) (Gupta et al., 2008) has registered over fifty different available data resources (as of April 2010) which include data or information regarding genes, cells, neuroanatomy, and neuroimaging studies, to name a few. The data warehousing approach, while effective in many environments, also has limitations. A major source of complexity in warehouses is that a copy of the data from each source has to be maintained in the warehouse and must be kept consistent in the face of updates at the original data source. Due to such complexity often the data in the warehouse is not the most recent data in the sources, but is only as current as the latest update cycle.

Data registries and catalogs, such as the NIF (Bug et al., 2008; Gardner et al., 2008; Gupta et al., 2008), point users to sources of information but do not actually provide the integrated data access that is eventually required. *Federation* based solutions, as we describe here, on the other hand do not disturb the autonomy of individual data sources and provide a *virtual* information source that end users can then use for integrated query to the information. On the other hand federated solutions also have some of their own disadvantages, such as being reliant on available networks (to sources) and network bandwidth for data transfer, performance issues, and changes at the original sources which in turn require changes at the overall federation level – all of which are absent of lesser consequence in a fully materialized or warehoused solution. The choice of the solution is influenced heavily by the environment in which data integration is being done and the application requirements for that integrated solution.

The BIRN project has at the core of its mission facilitating large-scale data sharing within various research communities. It published the BIRN Data Repository in 2009 (Fennema-Notestine, 2009), a public repository of neuroimaging and related metadata. The BIRN enabled data sharing via two database schemata: The human imaging database (HID, 2007) schema, as developed by functional imaging BIRN (FBIRN; Keator et al., 2008; Keator, 2009), and the XNAT database as developed by the Morphometry Imaging BIRN (Marcus et al., 2005). We present here the technology BIRN has developed to provide query access to these two data sources as though they were a single data source. Our approach is a *mediation* based approach (Wiederhold, 1992) where information mediator technology is used to provide integrated access to the individual data sources without requiring any modifications on the part of individual information sources and providers. Although in this paper we focus on neuroimaging sources, the approach is general and it can be applied to other sources and domains (cf. Section Results).

MATERIALS AND METHODS

In this section we first present our general data integration architecture. Then, we describe in detail the application of this architecture and integration methodology to the FBIRN use case, as well as more briefly to other use cases.

CORE DATA INTEGRATION ARCHITECTURE

Our BIRN mediator follows the classical virtual data integration architecture (Wiederhold, 1992; Ullman, 1997; Florescu et al., 1998; Halevy, 2001; Lenzerini, 2002), as shown in **Figure 1**. The salient features of the *mediation* architecture are:

(i) Data resides at and is maintained at the original data sources, and no changes are required to any of the data sources.
(ii) All that is required (for integration) is knowledge of the information content and access to the data source via an API, for example a JDBC connection or a web service.
(iii) Integrated data access is provided via a mediator, which is software that can reside on any server and which contacts the various data sources at query time to obtain the information requested by the user/client.

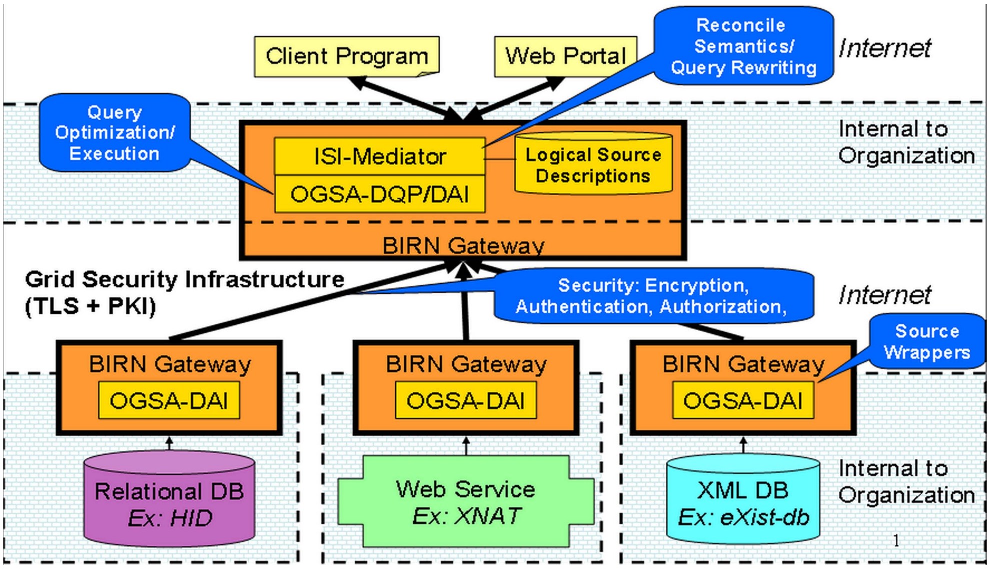


FIGURE 1 | Mediator architecture.

Two critical concepts in our approach to data integration are virtual organization and domain schema. A *virtual organization* is a community of data providers and data users that want to share, access, analyze data for some specific purposes. A *domain model* is the view of the data in the application domain that the virtual organization agrees upon as useful for the purposes of data sharing. The concept of domain modeling is closely related to that of ontologies. In fact, domain modeling can be seen as an incremental, data-driven, pragmatic approach to domain ontology development. A novel feature is that our architecture is built upon grid computing technologies (Foster et al., 2001) in order to leverage their scalability and strong security model, as we describe below.

The data integration system has three major components:

- Mediator. This component presents a uniform semantically-consistent schema, the domain model, to clients/users of the system. To users the system looks like a single database. However, this is a virtual database as there is no data stored at the mediator. The data remains at the sources. The mediator reconciles the semantic discrepancies among the sources, by using a set of declarative logical descriptions of the contents of the sources. The user poses queries to the mediator using terms from the domain model. Then the mediator uses the source descriptions to identify the sources relevant to the user query and to rewrite the domain-level user query, expressed in terms of the domain model, into a source-level query, expressed on terms of the source schemas. Since there are no changes to the source schemas, only source-level queries can actually be evaluated at the sources. In the next section we present detailed examples of the domain model and integration rules.

  Our architecture is designed in a modular fashion, so that any mediation approach that produces source queries in a language supported by the query evaluation engine can be plugged into the system. As we describe below, the mediator for the FBIRN use case is based on the global-as-view (GAV) model (Adali et al., 1996; Florescu et al., 1996; Garcia-Molina et al., 1997; Ullman, 1997). The query language supported is SQL-92. However, we plan to support other mediation approaches such as local-as-view (LAV) (Halevy, 2001; Lenzerini, 2002) and other languages such as OWL2-QL (Calvanese et al., 2007).

- Distributed Query Evaluation Engine: This component evaluates source-level relational queries after they are generated by the Mediator. This component is based on the open grid services architecture (OGSA) distributed access and integration (DAI), and distributed query processing (DQP) projects (OGSA, 2010). OGSA-DAI (Anjomshoaa et al., 2003; Grant et al., 2008) 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 (2010) grid web service (Sotomayor and Childers, 2005; Foster, 2006). OGSA-DQP (Lynden et al., 2008, 2009) 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 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-92 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 into the system.
- Source Wrappers: The actual data sources are wrapped as OGSA-DAI resources. OGSA-DAI includes a library of connectors to common data sources such as relational databases, and provides a common extensible framework to add new types of data sources.

Security is also a critical design requirement for biomedical applications. Our data integration system leverages the grid security infrastructure (GSI) (Lang et al., 2006) which provides encryption of transmitted data using industry standard TLS/SSL protocol and public key infrastructure to authenticate users, sources and servers. In additional to the standard GSI security, we have instrumented the system with logging and auditing mechanisms, so that administrators at both the mediator node and the data sources nodes can know which user executed which query. Finally, we are developing an expressive user data access control approach, whose description is beyond the scope of this paper.

This data integration infrastructure is the same for all application domains. As we describe in the next section, when we need to integrate data sources in a new application domain, the developer just needs to define a declarative domain model and source descriptions. Occasionally, if the domain includes novel types of sources not previously encountered, then the developer needs to define a wrapper for the new source type, which is then added to the library of wrappers and can be reused in future applications.

THE FBIRN DATA INTEGRATION USE CASE

General purpose data integration technology (Adali et al., 1996; Arens et al., 1996; Florescu et al., 1996; Garcia-Molina et al., 1997; Ullman, 1997; Halevy, 2001; Lenzerini, 2002; Thakkar et al., 2005) is a powerful tool that is applicable in building a variety of information integration applications. However the development of any new integration application is a reasonably complex process requiring the time and effort of personnel who are specialized to some degree. **Figure 2** illustrates the development "lifecycle" of a new data integration application showing the main processes and the key kinds of personnel required (thin arrows indicate the involvement of a class of personnel in a step, and block arrows indicate the ordering among steps). The development is done entirely by one or more *application developers*, who (at the least) have a working knowledge of how to *apply* the mediator technology in building data integration applications, have basic data modeling expertise and have basic programming expertise.

The first step is requirements gathering. The application developer(s) meet with *domain experts* to obtain an understanding of the need for data integration in their domain, the data sources they require integrated access over, and the kinds of capabilities they would expect from integrated data access.

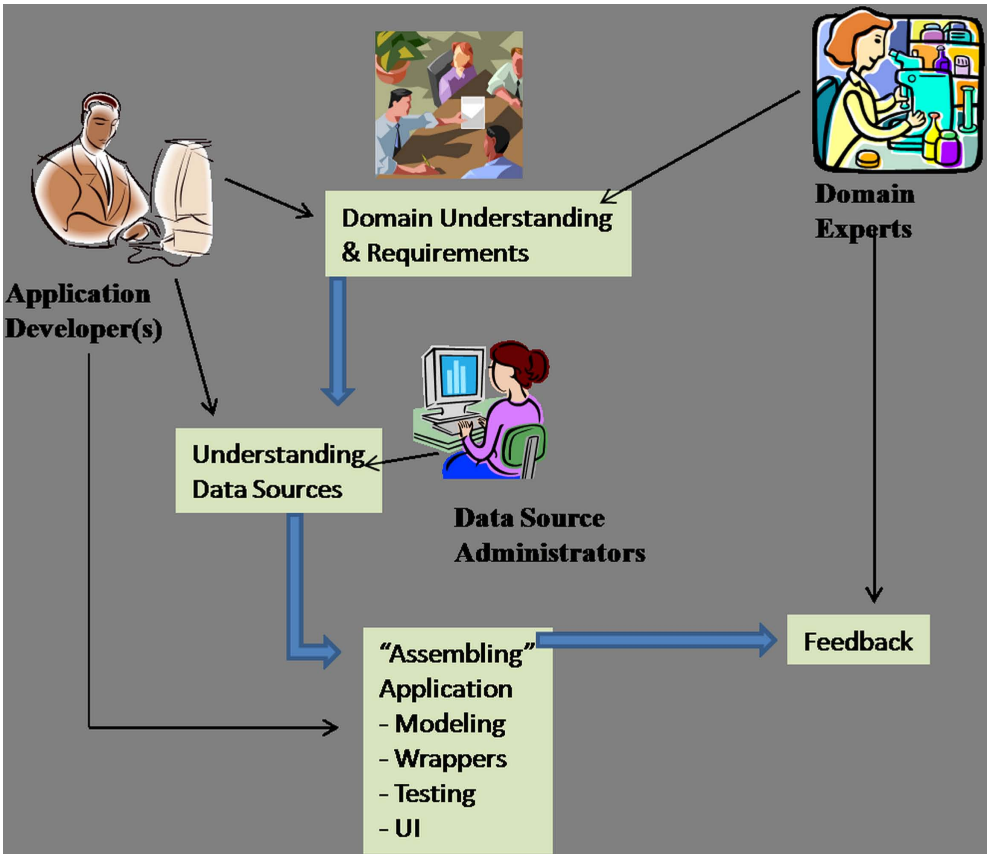


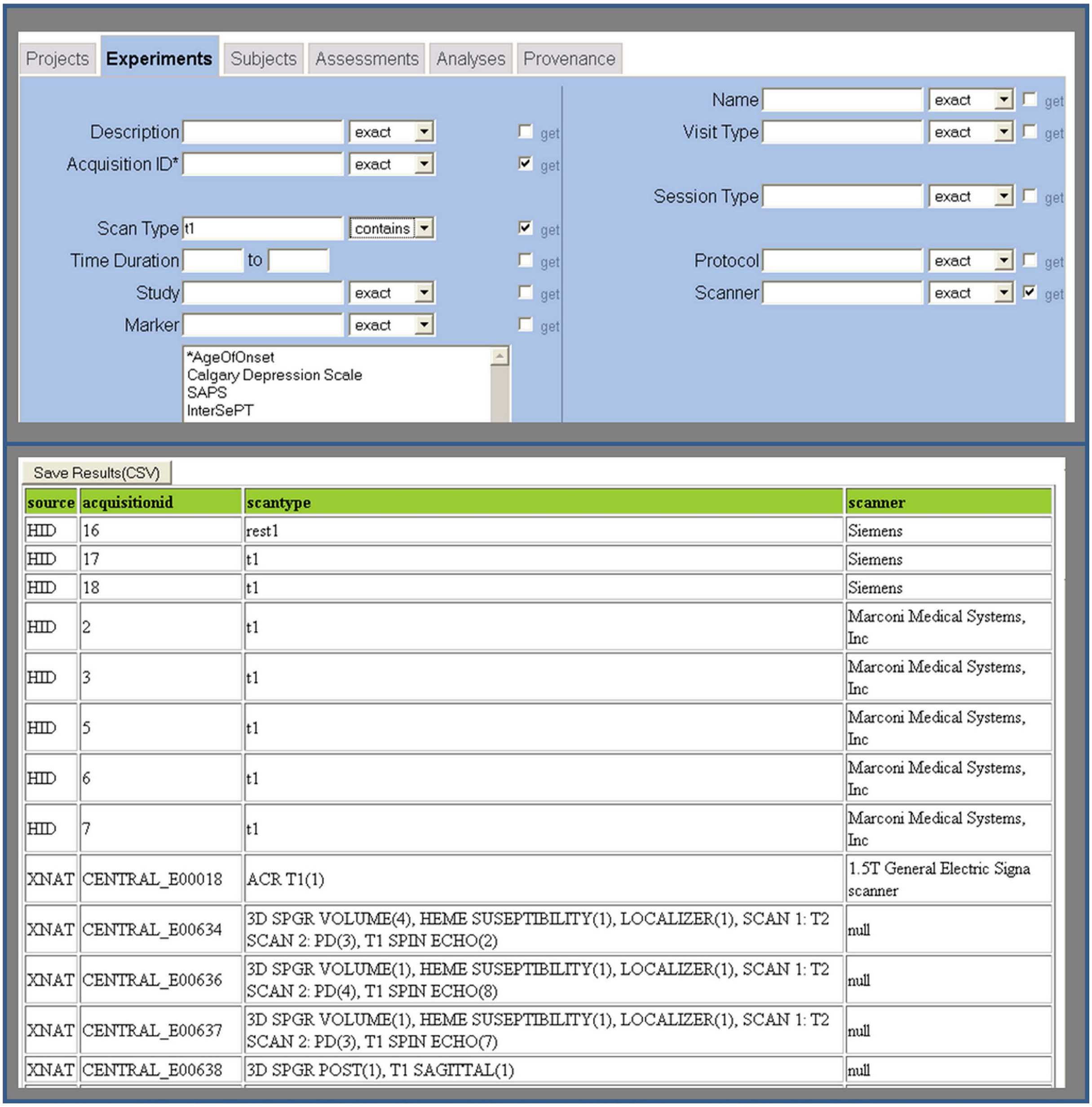
FIGURE 2 | Developing a data integration application.

The "Domain Understanding & Requirements" process box in **Figure 2** illustrates this step. Note that the set of data sources is but one parameter in defining a data integration application. Equally important parameters are the particular portions or aspects of information that are of interest in the particular application and what kinds of questions the eventual users expect the integrated system to be able to answer. Therefore, soliciting the domain understanding and requirements from domain experts is critical in the success of a new application. The application developer acquires an understanding of the particular data sources to be accessed (the "Understanding Data Sources" box). This involves understanding the type, interfaces (if any), content, and access information for that data source. It involves looking at the data source documentation and also interviews with the data administrators if possible.

The developer then proceeds to the actual development phase (**Figure 2** "Assembling Application"). Based on the requirements and understanding of data sources, the developer configures the mediator to meet the particular application needs. The term configuration here should not be interpreted as a simple step of setting some parameters, rather it is a complex process involving detailed data modeling and possibly the development of specific components (such as "wrappers," which we elaborate on shortly) to meet the application needs. The developer then has to evaluate and ensure that the application is indeed able to answer the kinds of queries that users are ultimately interested in. In many cases a graphical user interface (GUI) is also provided to non-expert users to facilitate query formulation. Finally, developers seek feedback from the intended user community steering development of improved versions of the application.

Steps and Processes

We elaborate below on these key steps and processes. We do this in the context of what we call the "FBIRN Data Integration System" that is an actual, functioning data integration application that we have built using the above described data integration technology and methodology. The FBIRN data integration case involves the integration (for a start) of two distributed and heterogeneous sources in the neuroimaging domain. These are (i) The HID (Keator et al., 2009) which is a relational database of experimental information, and (ii) The eXtensible Neuroimaging Archive Toolkit (XNAT) (Marcus et al., 2005), which is an XML web-service based repository of experimental information. Specifically, we integrated HID instances at UC Irvine and the Mind Research Institute, and XNAT Central at Washington University in Saint Louis. Two data sources are related if they both contain experimental information about subjects, i.e., about various kinds of experiments done for different studies and visits on subjects, the particular data collected such as scans in such experiments, etc. The sets of subjects (individuals) in the two data sources can be and are usually disjoint. Additionally,



Description [ ] exact □ get
Acquisition ID* [ ] exact ☑ get
Scan Type [t1] contains ☑ get
Time Duration [ ] to [ ] □ get
Study [ ] exact □ get
Marker [ ] exact □ get
*AgeOfOnset
Calgary Depression Scale
SAPS
InterSePT

Name [ ] exact □ get
Visit Type [ ] exact □ get
Session Type [ ] exact □ get
Protocol [ ] exact □ get
Scanner [ ] exact ☑ get

source	acquisitionid	scantype	scanner
HID	16	rest1	Siemens
HID	17	t1	Siemens
HID	18	t1	Siemens
HID	2	t1	Marconi Medical Systems, Inc
HID	3	t1	Marconi Medical Systems, Inc
HID	5	t1	Marconi Medical Systems, Inc
HID	6	t1	Marconi Medical Systems, Inc
HID	7	t1	Marconi Medical Systems, Inc
XNAT	CENTRAL_E00018	ACR T1(1)	1.5T General Electric Signa scanner
XNAT	CENTRAL_E00634	3D SPGR VOLUME(4), HEME SUSEPTIBILITY(1), LOCALIZER(1), SCAN 1: T2 SCAN 2: PD(3), T1 SPIN ECHO(2)	null
XNAT	CENTRAL_E00636	3D SPGR VOLUME(1), HEME SUSEPTIBILITY(1), LOCALIZER(1), SCAN 1: T2 SCAN 2: PD(4), T1 SPIN ECHO(8)	null
XNAT	CENTRAL_E00637	3D SPGR VOLUME(1), HEME SUSEPTIBILITY(1), LOCALIZER(1), SCAN 1: T2 SCAN 2: PD(3), T1 SPIN ECHO(7)	null
XNAT	CENTRAL_E00638	3D SPGR POST(1), T1 SAGITTAL(1)	null

FIGURE 3 | FBIRN query interface and results.

not all kinds of (experimental) information are present in both sources. For instance, some psychometrics data is present in XNAT but not in HID.

The FBIRN data integration use case is well suited as a representative application for our approach because of the following key aspects:

(1) *The heterogeneity of the information sources being integrated.* The two information sources HID and XNAT are very heterogeneous mutually. This heterogeneity stems from:

Data model heterogeneity – The two information sources, while containing similar information, represent it very differently. HID uses the relational model whereas XNAT uses the (fundamentally different) XML data model. Further, the manner in which the two sources have represented many of the same concepts is quite different. For instance information about projects is stored in XNAT in an *XML element* called `Project` (**Figure 4**), whereas in HID such information is represented a *relational table* called `nc_Experiment` (XML elements and tables are only partially shown in **Figure 4** for readability). The XML element and relational table names as well as names of some corresponding field names are also different across the two sources. As a more complex example consider information about experimental scans that is stored in XNAT in an XML element `SessionData` (second row in **Figure 4**) whereas the same information is stored in HID across *three* tables – `nc_ExpSegment`, `nc_ExpStudy`, and `nc_Protocol`. Note that

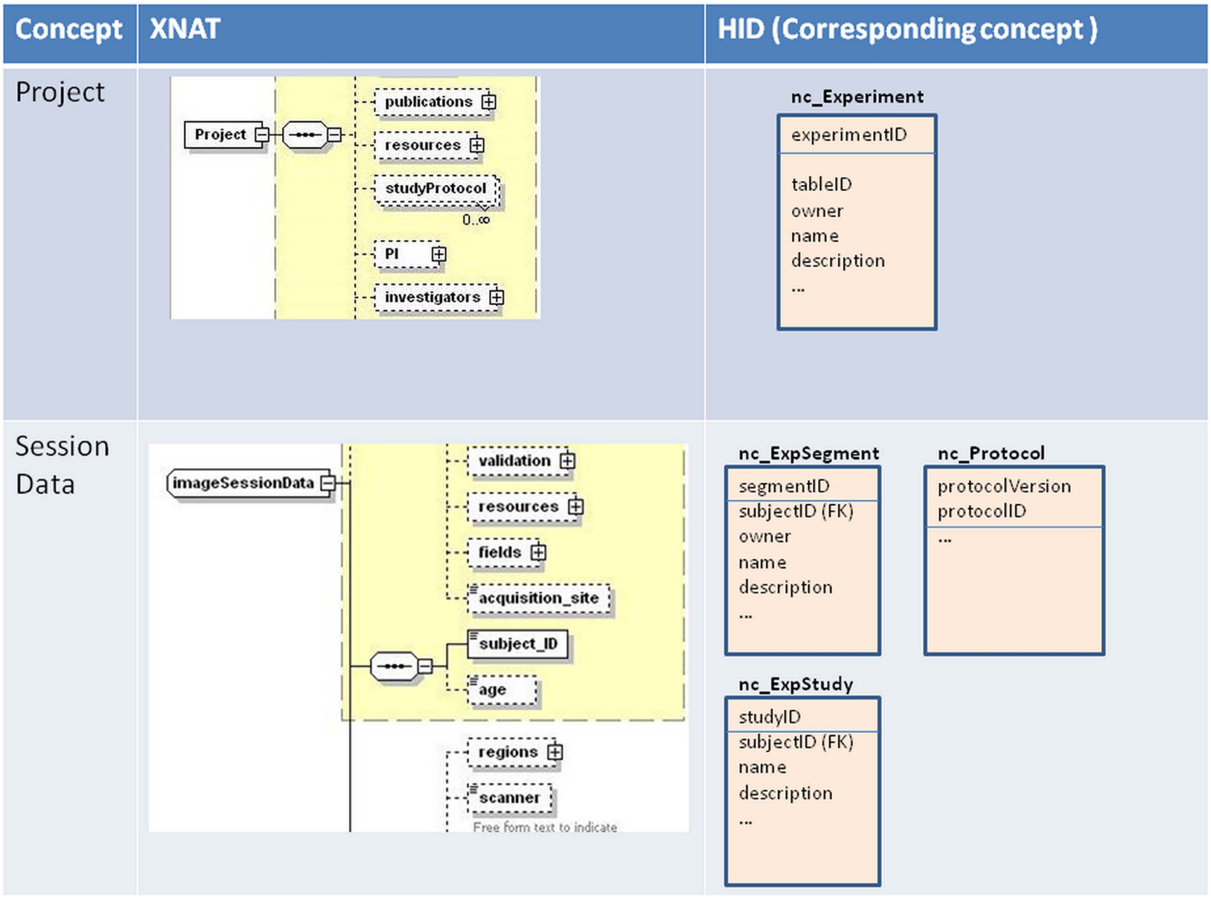


FIGURE 4 | (Model) Heterogeneity Across HID and XNAT.

in this case there is a *structural* difference as well in the information representation (three tables in HID vs a single XML element in XNAT).

Integrated data access to these two sources requires the resolution of many such model heterogeneities. Note that while both data sources have to some extent followed the recommendations of XCEDE (Keator et al., 2009), which is a data modeling and definition recommendation for the biomedical experimental informatics domain. However, XCEDE is not a data model *per se* for adoption. Thus the data models of XNAT and HID are quite different as shown in **Figure 4**.

Query language and API heterogeneity – HID data (being relational) is accessed using SQL whereas XNAT (being XML based) is queried using XPath (1999). Besides, HID offers connectivity using relational database access primitives such as JDBC whereas XNAT offers a "RESTful" API using a Web services framework (XNAT, 2010).

(2) *Value to investigators* – There is significant merit of integrated access to these two data sources to scientific investigators (acknowledged by them) as both sources contain rich experimental information about different subjects and their integration provides investigators with access to information on a wider pool of subjects.

(3) *Highlighting all multi information source aspects* – An application integrating two sources is not, as we shall see, a pathologically simple case in data integration, i.e., as compared to applications with more than two sources. The solution, even with two sources, requires addressing the entire gamut of sub problems that must be addressed even with a larger number of information sources to be integrated.

We present the application development steps for this use case.

Domain Understanding and Requirements

We had access to a domain expert, a neuroscientist who provided us with knowledge of the information content in the two data sources, the relationship between the sources, and the eventual need for integrated access across such sources. The XCEDE recommendations also proved to be useful in understanding key concepts and relationships in the domain of experimental information in general. For data sources, for HID we had access to the original database design documentation (as entity–relationship diagrams) as well as the administrator (DBA) for the HID at UC Irvine. Note that neither of documentation or a database administrator can be presumed to be present or available always for a data source. In the case of XNAT there is documentation provided in the form of XML schema diagrams which is an excellent starting point for understanding data sources based on XNAT. The XNAT Central data is publicly available and we could browse and access this information to understand the data content. Also, the XNAT framework provides a "search API," available as a REST web service, that we installed and started using and configuring for our needs.

Assembling the Application

Modeling. The first key step in assembling a new data integration application is to develop an integration model for the application. The integration model consists of three primary components: the definition of source concepts, the definition of domain concepts, and the definition of integration rules.

(a) *Source Concepts*: The first step in domain modeling is to model the contents of each information source that is part of the integrated application. Source concepts are used to model a data source table, element, or object class. **Table 1A** illustrates a portion of the source model specifications for the FBIRN application domain model. The first source concept corresponds to an element in XNAT, and the second concept corresponds to a table in HID. Note that in order to make every source predicate unique we prefix the original source concept with the data source, for example, in `HIDResource_NC_EXPERIMENT` the database table is `NC_EXPERIMENT` and `HIDResource` is the DAI resource wrapping the HID database.

(b) *Domain Concepts*: The domain concepts define the global, integrated view of the information. In the FBIRN application both data sources contain essentially the same *kind* of information: experimental information about neuroimaging subjects. The XCEDE recommendation describes key concepts used in experimental informatics, so we used terms recommended in XCEDE to define the domain model where appropriate.

Our domain concepts are based on the following key terms:

Project. A project is the top-level division of experiment data, and represents a research project which collects and analyzes data from one or more subjects.

Experiment. An experiment is one of the central concepts in this domain and represents a coherent investigation unit. Experiments comprise of different *episodes* which in turn comprise of different experimental *acquisitions*. In more detail:

(a) *Experiment Episode*. An episode represents a unit of data collection by one or more instruments over a given time interval.

(b) *Experiment Acquisition*. Each set of data collected (perhaps by different instruments) over this time interval should be represented by an acquisition. Multiple acquisitions within an episode should be understood to occur simultaneously over the time interval represented by the episode.

Assessment. An assessment is a kind of data element that captures information related to experiments and subjects.

Analysis. Analysis encapsulates metadata about data that is derived from one or more inputs.

Provenance. Provenance captures "associated" information about experiments such as say the make and model of scanners used for the experiment, etc.

Protocol. Structures to describe the expected course of an experimental paradigm are provided in the protocol.

Visit. A visit may represent a subject's appearance at an experiment "site" (for collaborative projects, this could be the institution or lab at which the data is being collected or analyzed). A visit may be further subdivided into one or more studies, each of would consist of one or more data collection episodes.

Study. A study is a component of an experiment visit.

Subject. The individual on whom experiments are conducted and data is maintained for.

Table 1B illustrates two domain concepts in the domain model, the **g_Project**[4] concept that models information on Projects and the **g_Subject** concept that models Subjects. As we see the g_Project concept contains key fields such as the project ID (identifier), project name, keywords associated with the project etc. The FBIRN domain model consists of 33 such domain concepts modeling various kinds of information such as projects, experiments, subjects, assessments, etc.

(c) *Rules*: Domain concepts, which are really "virtual" concepts, must be associated with actual information sources, i.e., with corresponding concepts or sets of concepts in the source model. This is done using declarative logical rules based on the GAV approach to information integration, where a global model is essentially a view over the source data model. **Table 1C** illustrates a set of domain model rules relating the domain concept **g_Project** to concepts in XNAT and HID. The first rule relates the g_Project domain concept to the `XNATProjectResource_xnat__projectData` source concept in XNAT. The attributes in the domain concept such as the `projectID, projname,` etc. are related to the corresponding attributes in the source concept using common names in the head as well as body predicates of the rule. The second rule relates this domain concept to the corresponding source concept in HID, `HIDResource_NC_EXPERIMENT,` which corresponds to the NC_EXPERIMENT table in HID. The rules are expressed in Datalog (Ullman, 1998) as is often done in information integration.

The FBIRN domain model comprises of 35 HID source concepts, 25 XNAT source concepts, 33 domain concepts and 52 rules. Defining such integration rules requires a thorough understanding of the information content of the sources being integrated. The application developers must know how the different sources are related, where there is overlap in the (kinds of) information content, and also the semantic relationships in the information across the sources.

Wrapping the data sources. As described above, in the mediator based approach we require a wrapper for each information source that is to be accessed.

[4] By convention and for clarity we prefix the domain (global) concepts with g_, but this is not required.

(a) In the case of HID the underlying database originally was Oracle 10g, a relational database and for which a wrapper already exists with the mediator infrastructure. In this case we were able to use the existing wrapper "as is" for the HID data source. A more recent instantiation of HID (at UCI) is in Postgres which we were able to wrap with equal ease.

(b) XNAT is an XML based data source for which access is provided via the "search" REST web service API set of tools. We developed a wrapper for XNAT over the search REST web service which essentially takes a mediator query in SQL, translates it to an equivalent XML document that is transmitted to XNAT using the REST API, and translates back the XML results from the web service into a relational form.

To summarize, the integration application developer needs to understand the source schemas, define a domain schema, and define the integration rules. This knowledge is recorded in a declarative integration model file with the syntax shown in **Table 1**. In addition, if a data source is of a type not previously seen; he/she may need to define a new type of wrapper (DAI resource) to add to the library. Although this process may involve significant modeling work, this is all the work needed for a new application. In prosaic terms, the developer just needs to write one file. The data integration system itself is already available as a software package that just needs to installed at the appropriate host(s) and simply configured with the integration model file. Thus, the BIRN mediator allows for rapid integration of heterogeneous sources.

Table 1 | Source and domain model, and integration rules.

(A) Source concepts	**XNATProjectResource_xnat__ projecData**(insert_date:number, projectID:string, projname:string, … proj_ct_count:number, proj_ut_count:number)	**HIDResource_NC_ EXPERIMENT** (id:string … name:string, description:string, PIname:string)
(B) Domain concepts	**g_Project**(source,projectID, projname, stringprojdescription, PIname, projURI,keywords) **g_Subject**(source,subjectID, subjectname, subjectage, subjectrace, subjectgender,subjectethnicity)	
(C) Rules	**g_Project**("XNAT," projectID, projname, - projdescription, PIname, "NA," keywords): XNATProjectResource_xnat__projectData(insert_date, user, projectID, projname, projdescription PIname,…… ,….. proj_ct_count,proj_ut_count) **g_Project**("HID," projectID, projname, projdescription, PIname, projURI,"NA"):- HIDResource_NC_EXPERIMENT (projectID,tID, owner, modtime, moduser, projname, projdescription, PIname, projURI, isRegression)	

Not all details shown, for readability.

RESULTS

FBIRN

The intended use of the FBIRN data integration application is to provide domain scientists with an information-gathering tool that they can use to make their current data exploration and access tasks easier. One key scientific task is hypothesis postulation and validation – often with data analysis to validate (or otherwise) a proposed hypothesis. As an example consider a hypothesis such as *"Female Alzheimer's patients with at least 5 years' illness duration have less activation in their DLPFC during a working memory task in an fMRI scan on a 3T scanner than do healthy controls."* One can envision the use of experimental data analysis to assess whether the data we have does indeed support this hypothesis or not. The following are some key categories of information requests that are required in such cases:

(i) Subjects with particular demographic characteristics, for example gender, age, or socioeconomic status.

(ii) Assessments of particular kinds, for instance Alzheimer's patients can be determined by looking at the CDR and MMSE assessments of subjects.

(iii) Scans of particular kinds and taken with particular makes and models of scanners.

We thus evaluated the FBIRN data integration application by postulating multiple sets of meaningful queries of the above kinds, evaluating whether correct/expected results are obtained, and also assessing the query response time. **Table 2** illustrates the high-level information requests, actual mediator query, variants of the query, answer correctness, and average query response time associated. Our initial evaluations indicate that (i) We are indeed able to formulate meaningful data requests as formal mediator queries, (ii) The results returned are indeed from multiple data sources, aggregated, and integrated in a correct or expected manner, (iii) The query response times are reasonable given the fact that data is being accessed from multiple remote sites in real-time.

To facilitate the use of the FBIRN mediator by researchers, we also developed a web-based query formulation interface. **Figure 3** shows screenshots of the application query interface and the results of a query asking for all subjects in all data sources with "t1" scans (both query interface and results are only partially shown for brevity).

In **Table 3** we report on the personnel and effort it took to develop the FBIRN application as an illustration of the current complexity and skills required for assembling new applications. We must highlight that with the completion of this effort, it took us just about a few hours to successfully integrate another instantiation of the HID database, the HID installation at the Mind Research Network (MRN) into the integrated application. We were able to entirely reuse the models and rules from the first HID database we integrated. This is an illustration of *model reuse* as the MRN HID instantiation has exactly the same schema as the instantiation of HID (at UC Irvine) that we had modeled.

Table 2 | Mediator queries in FBIRN.

Information need	Mediator query	Variants	Correct answers (Size)	(Avg) response time (s)
Find all female subjects between the ages of 40 and 50.	Q(source, subjectid):- g_hasAge(source, subjectid, A) ^ g_hasGender(source, subjectid, G) and (40 < A < 50) ^ (G = "F")	Vary age parameters, and constraints. on other aspects such as handedness, race etc	Yes (20–100 tuples)	2.3
Find all subjects with indications of Alzheimer's	Q(source, subjectid, CDR, MMSE):- g_Assessment(source, subjectid, "CDR," C) ^ g_Assessment(source, subjectid, "MMSE," M) ^ C > 3 and M > 10	Vary (subject) conditions being searched for.	Yes (2–35 tuples)	1.9
Find all fMRI scans taken with a 3T scanner	Q(source, subjectid, scan):- g_ExperimentAcquisition() ^ (ST = fMRI) ? (scanner = 3T)	Vary types of scans and/or scanners	Yes (30–35 tuples)	12.1

Table 3 | FBIRN application development effort.

Task	Personnel	Time (person months)
Requirements understanding	Model developer, Domain expert, Data administrators	2
Data source understanding	Model developer	3
Developing domain model	Model developer	5
Wrapper development	Programmer	0.75
Query evaluation	All	2
GUI development	Programmer	1

OTHER DOMAINS

This integration approach and technology is also being applied successfully, and in a similar fashion, to other data integration cases within BIRN, namely the Non-Human Primate Research Consortium that integrates data from eight National Primate Research Centers, and to the Cardiovascular Research Grid (CVRG) which requires integrated access to distributed sources of patient data, ECG waveform data and analysis, and DICOM images of the heart; as well as non-BIRN use cases, such as a clinical trial of colon cancer. We describe the CVRG use case in some additional detail to compare with the FBIRN application of the BIRN mediator.

The CVRG Project (CVRG, 2010) has developed systems for storing genetic, transcriptional, proteomic, ECG, imaging, and clinical data, in addition to easy-to-use interfaces for querying, retrieving, and analyzing data. We applied the BIRN mediator to query across demographic, ECG and image metadata as well as raw waveform files and DICOM image files. The mediator integrated data from: (1) an open-source PACS dcm4che instance (the metadata is in a MySQL relational DB and the image files are in the file system), (2) a MySQL DB that contains ECG analysis from the MESA project, (3) an XML/XQuery database (eXistDB) that contains Chesnokov Analysis data, and (4) from another instance of eXistDB with waveform metadata in XML. The development of the CVRG application followed the same approach and steps as for FBIRN. The mediator code base is the same, but a different integration file was defined that modeled the CVRG domain. In addition we defined a new wrapper for the eXistDB, which is a XML/XQuery database. The wrapper flattened some of the tree-structured XML concepts so that they could be modeled as relations by the mediator and queried using SQL.

We expect that as we apply our mediator to more domains the wrapper library covers most of the common sources, and the existing domain concepts can be reused directly, or form the basis for extensions, in other applications. In this way we expect that creating integration systems for additional domains can be accomplished in an increasingly rapid fashion.

DISCUSSION

We have presented a general architecture for data integration in biomedical domains and described its application to a significant use case: FBIRN. Our goals in the Biomedical Information Network is to apply this architecture to additional domains of interest and exploit the synergies that integrated access to an increasing number of data sources can provide. We expect that our bottom-up virtual integration approach will provide an effective, scalable mechanism to achieve widespread sharing of biomedical data. Our architecture is not only extensible in terms of new application domains, such as genetics, oncology, or cardiology, but it is also modular in terms of its core components, so that the architecture will accommodate a *family* of mediation approaches.

Data sharing and integrated access are topics of ubiquitous interest in the biomedical informatics community, with multiple efforts in the area taking different approaches. The DXBrain system (Detwiler et al., 2009) offers a "light-weight" XML based approach to data integration and with a focus on the neuroscience domain. This work represents an interesting alternative approach to information integration, in which the light-weight nature of the integration framework makes it easy to add new sources. However, the cognitive burden of schema integration is shifted to the user. Our approach is to assemble a deeper and more semantic integration of sources up front, so that users are abstracted from this at query time.

caBIG (2010) is a National Cancer Institute initiated effort on data sharing, initially for the cancer research domain but now with applicability to many other medical and scientific domains. The caBIG approach is an "adapt or adopt" approach that requires data source providers that want to participate in a data sharing application to understand caBIG standards such as enterprise vocabulary service (EVS) and the common data elements and further use caBIG tools to adapt their sources to these standards. Our mediation approach does not require any additional schema modifications on behalf of data source providers, though the process of building the local domain model is a limited version of determining a standard vocabulary, and the common data elements may be considered similar to domain models.

The TraM system (Wang et al., 2009) provides a warehouse curation-based approach to biomedical data integration, offering a framework with data conversion and workflow tools with which a centralized data warehouse of data integrated from multiple sources can be realized. In comparison, the strengths of our approach are that it offers deep "semantic" integration of information from multiple sources as opposed to serving (just) a clearinghouse functionality. Further, it is based on mediation that does not demand any adoption, conversion, or cataloging on part of a data source provider – rather sources are integrated as is.

The Neuroscience Information Framework (Bug et al., 2008; Gardner et al., 2008; Gupta et al., 2008) is a web-based dynamic *inventory* of neuroscience resources such as data, materials, ontologies, and tools. It is a one-stop clearinghouse for resources obtained (originally) from different sources, with query capabilities that have focused on neurobiological ontology development and application. Users can automatically query for their chosen terms and semantically related terms, returning the data categorized by type (activation foci, clinical trials, grants, literature reviews) and by source. It provides access to individual information items or resources at one place. Our work is more focused on the deeper integration for seamless *structured querying* across multiple sources with the capabilities of an expressive query language (such as SQL). In this application as an example, what NIF can enable is a capability that first identifies sources and tables or elements within each source that are relevant to a query. Thus, with appropriate source registration done in advance using NIF tools such as DISCO (Marenco et al., 2010), a search for say "scans" will identify the relevant elements within XNAT containing scan data, and the relevant tables within HID containing that data. It is then for the user to specify appropriate constraints and conditions on individual fields within the tables or elements in each source, and then results are obtained *per source*. This is fundamentally different from our approach where the user is completely abstracted from the number and nature of sources containing relevant data, and the details of the relevant tables and elements within. Moreover, the BIRN mediator retrieves and integrates the actual individual data from multiple sources in response to complex structured (SQL) queries.

Our experience with the FBIRN application has identified several directions for further work including some that we have already initiated work on. There is room for improvement in getting the application to a point where investigators actually adopt such an application for their day-to-day use. We conducted evaluations of the FBIRN application in individual sessions with many senior FBIRN investigators at various institutions (including the Universities of California San Francisco, Iowa, Minnesota, UCLA and UC Irvine), and identified key components to increasing usability of the interface. The latter includes (1) Better documentation of information integration applications. We are in fact taking a knowledge engineering approach to such documentation with the goal of semantically annotating integration models in significant detail so that such knowledge can be easily reused *across* applications, (2) Developing better, i.e., more intuitive, and exploratory user interfaces for applications of the kind that intended users (scientific investigators) would be able to comprehend and effectively use. Here we have also identified directions for the automated model based *generation* of such interfaces that may be automatically generated given the application domain model and other configuration information. We must mention here that our newly initiated work on knowledge engineering for information integration documentation is also aimed at addressing practical problems that arise in a mediation environment when data sources, i.e., their actual data and/or schemas evolve with time. We are developing techniques that require a minimal effort (for model changes) from application builders in the face of such changes and evolution.

ACKNOWLEDGMENTS

This work would not be possible without the valuable guidance and technical help of several individuals including Dan Marcus and Tim Olsen at Washington University for help with the XNAT system, Gowi Kumaraguruparan, Craig Knoblock, and Craig Rogers at USC/ISI for data integration, wrapper technology and user interface development and Joseph Ames, David Keator, and Dingying Wei at UC Irvine for assistance with HID and overall BIRN strategic input, and finally Steven Potkin also at UCI for overall guidance in the FBIRN domain. This work was supported in part by the Biomedical Informatics Research Network (1 U24 RR025736-01) and by FBIRN (1 U24 RR021992).

REFERENCES

Adali, S., Candan, K. S., Papkonstantinou, Y., and Subrahmanian, V. S. (1996). Query caching and optimization in distributed mediator systems. *SIGMOD Record* 25, 137–148.

Anjomshoaa, A., Antonioletti, M., Atkinson, M. P., Baxter, R., Borley, A., Chue Hong, N. P., Collins, B., Hardman, N., Hicken, G., Hume, A., Knox, A., Jackson, M., Krause, A., Laws, S., Magowan, J., Palansuriya, C., Paton, N. W., Pearson, D., Sugden, T., Watson, P., and Westhead, M. (2003). The design and implementation of grid database services in OGSA-DAI. *Concurrency Comput. Pract. Ex.* 17, 357–376.

Arens, Y., Knoblock, C. A., and Shen, W.-M. (1996). Query reformulation for dynamic information integration. *J. Intell. Inform. Syst.* 6, 99–130.

Bug, W. J., Ascoli, G. A., Grethe, J. S., Gupta, A., Fennema-Notestine, C., Laird, A. R., Larson, S. D., Rubin, D., Shepherd, G. M., Turner, J. A., and Martone, M. E. (2008). The NIFSTD and BIRNLex vocabularies: building comprehensive ontologies for neuroscience. *Neuroinformatics* 6, 175–194.

caBIG. (2010). https://cabig.nci.nih.gov

Calvanese, D., Giacomo, G. D., Lembo, D., Lenzerini, M., and Rosati, R. (2007). Tractable reasoning and efficient query answering in description logics: the DL-Lite family. *J. Automat. Reas.* 9, 385–429.

CardioVascular Research Grid (CVRG). (2010). www.cvrgrid.org

Detwiler, L. T., Dan, S., Franklin, J. D., Moore, E. B., Poliakov, A. V., Lee, E. S., Corina, D. P., Ojemann, G. A., and Brinkley, J. F. (2009). Distributed XQuery-based integration and visualization of multimodality brain mapping data. *Front. Neuroinform.* 3:2. doi: 10.3389/neuro.11.002.2009.

FBIRN Human Imaging Database (HID). (2007). *User Manual 2007*. Available at: http://nbirn.net/research/function/hid.shtm

Fennema-Notestine, C. (2009). Enabling public data sharing: encouraging scientific discovery and education. *Methods Mol. Biol.* 569, 25–32.

Florescu, D., Levy, A. Y., and Mendelzon, A. (1998). Database techniques for the world-wide web: a survey. *SIGMOD Record* 27, 59–74.

Florescu, D., Raschid, L., and Valduriez, P. (1996). "Answering queries using OQL view expressions," in *Workshop on Materialized Views: Techniques and Applications,* Montreal, Canada: SIGMOD, pp. 84–90.

Foster, I. (2006). Globus toolkit version 4: software for service-oriented systems. *IFIP International Conference on Network and Parallel Computing,* Springer-Verlag, LNCS 3779, 2–13.

Foster, I., Kesselman, C., and Tuecke, S. (2001). The anatomy of the grid: enabling scalable virtual organizations. *Int. J. Supercomput. Appl.* 15, 200–222.

Galperin, M. Y., and Cochrane, G. (2009). Nucleic acids research annual database issue and the NAR online molecular biology database collection in 2009. *Nucleic Acids Res.* 37, 1–4.

Garcia-Molina, H., Papakonstantinou, Y., Quass, D., Rajaraman, A., Sagiv, Y., Ullman, J. D., Vassalos, V., and Widom, J. (1997). The TSIMMIS approach to mediation: data models and languages. *J. Intell. Inform. Syst.* 8, 117–132.

Gardner, D., Akil, H., Ascoli, G. A., Bowden, D. M., Bug, W., Donohue, D. E., Goldberg, D. H., Grafstein, B., Grethe, J. S., Gupta, A., Halavi, M., Kennedy, D. N., Marenco, L., Martone, M. E., Miller, P. L., Müller, H., Robert, A., Shepherd, G. M., Sternberg, P. W., Van Essen, D. C., and Williams, R. W. (2008). The neuroscience information framework: a data and knowledge environment for neuroscience. *Neuroinformatics* 6, 149–160.

Globus. (2010). Web: http://www.globus.org

Grant, A., Antonioletti, M., Hume, A. C., Krause, A., Dobrzelecki, B., Jackson, M. J., Parsons, M., Atkinson, M. P., and Theocharopoulos, E. (2008). OGSA-DAI: middleware for data integration: selected applications. *eScience*, *IEEE Fourth International Conference on eScience* 343–343.

Gruenberger, M., Alberts, R., Smedley, D., Swertz, M., and Schofield, P. (2010). The CASIMIR consortium. *BMC Res. Notes* 3, 16. doi: 10.1186/1756-0500-3-16.

Gupta, A., Bug, W., Marenco, L., Qian, X., Condit, C., Rangarajan, A., Müller, H. M., Miller, P. L., Sanders, B., Grethe, J. S., Astakhov, V., Shepherd, G., Sternberg, P. W., and Martone, M. E. (2008). Federated access to heterogeneous information resources in the neuroscience information framework (NIF). *Neuroinformatics* 6, 205–217.

Halevy, A. Y. (2001). Answering queries using views: a survey. *VLDB J.* 10, 270–294.

Keator, D. B. (2009). Management of information in distributed biomedical collaboratories. *Methods Mol. Biol.* 569, 1–23.

Keator, D. B., Grethe, J. S., Marcus, D., Ozyurt, B., Gadde, S., Murphy, S., Pieper, S., Greve, D., Notestine, R., Bockholt, H. J., and Papadopoulos, P. (2008). A national human neuroimaging collaboratory enabled by the biomedical informatics research network (BIRN). *IEEE Trans. Inform. Technol. Biomed.* 12, 162–172.

Keator, D. B., Wei, D., Gadde, S., Bockholt, J., Grethe, J. S., Marcus, D., Aucoin, N., and Ozyurt, I. B. (2009). Derived data storage and exchange workflow for large-scale neuroimaging analyses on the BIRN grid. *Front. Neuroinform.* 3:30. doi: 10.3389/neuro.11.030.2009.

Lang, B., Foster, I., Siebenlist, F., Ananthakrishnan, R., and Freeman, T. (2006). "A multipolicy authorization framework for grid security," in *Fifth IEEE Symposium on Network Computing and Application*, Cambridge, USA, July 24–26.

Lenzerini, M. (2002). "Data integration: a theoretical perspective," in *Proceedings of ACM Symposium on Principles of Database Systems,* Madison, Wisconsin.

Lynden, S., Mukherjee, A., Hume, A. C., Fernandes, A. A., Paton, N. W., Sakellariou, R., and Watson, P. (2009). The design and implementation of OGSA-DQP: a service-based distributed query processor. *Future Generat. Comput. Syst.* 25, 224–236.

Lynden, S., Pahlevi, S. M., and Kojima, I. (2008). "Service-based data integration using OGSA-DQP and OGSA-WebDB," in *Proceedings of the 9th IEEE/ACM International Conference on Grid Computing* (Washington, DC: IEEE Computer Society), 160–167.

Marcus, D. S., Olsen, T., Ramaratnam, M., and Buckner, R. L. (2005). The extensible neuroimaging archive toolkit (XNAT): an informatics platform for managing, exploring, and sharing neuroimaging data. *Neuroinformatics* 5, 11–34.

Marenco, L., Wang, R., Shepherd, G. M., and Miller, P. L. (2010). The NIF DISCO framework: facilitating automated integration of neuroscience content on the web. *Neuroinformatics* 8, 101–112.

OGSA. (2010). Web: http://www.ogsadai.org.uk/

Sotomayor, B., Childers, L. (2005). *Globus Toolkit 4: Programming Java Services*. San Francisco: Morgan Kaufmann.

Thakkar, S., Ambite, J. L., and Knoblock, C. A. (2005). Composing, optimizing, and executing plans for bioinformatics web services. *VLDB J.* 14, 330–353.

Ullman, J. D. (1997). "Information integration using logical views," in *Proceedings of the Sixth International Conference on Database Theory,* Delphi, Greece, pp. 19–40.

Ullman, J. D. (1998). *Principles of Database and Knowledge-Base Systems,* Vol. I. New York: Computer Science Press.

Wang, X., Liu, L., Fackenthal, J., Cummings, S., Cook, M., Hope, K., Silverstein, J. C., and Olopade, O. (2009). Translational integrity and continuity: personalized biomedical data integration. *J. Biomed. Inform.* 42, 100–112.

Wiederhold, G. (1992). Mediators in the architecture of future information systems. *IEEE Comput.* 25, 38–49.

XNAT. (2010). Web: http://www.xnat.org/

XPath. (1999). Web: http://www.w3.org/TR/xpath/

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 16 July 2010; paper pending published: 07 September 2010; accepted: 05 December 2010; published online: 28 December 2010.

Citation: Ashish N, Ambite JL, Muslea M and Turner JA (2010) Neuroscience data integration through mediation: an (F)BIRN case study. Front. Neuroinform. ***4****:118. doi: 10.3389/fninf.2010.00118*