MetaDL: A Digital Library of Metadata for Sensitive or Complex Research Data

Fillia Makedon[1], James Ford[1], Li Shen[1], Tilmann Steinberg[1], Andrew Saykin[2], Heather Wishart[2], and Sarantos Kapidakis[3]

[1] The Dartmouth Experimental Visualization Laboratory, Department of Computer Science, Dartmouth College, Hanover, NH 03755, USA
{makedon,jford,li,tilmann}@cs.dartmouth.edu
[2] Brain Imaging Laboratory, Dartmouth Medical School, Lebanon, NH 03756, USA
{saykin,wishart}@dartmouth.edu
[3] Department of Archive and Library Sciences, Ionian University, Greece
sarantos@ionio.gr

Abstract. Traditional digital library systems have certain limitations when dealing with complex or sensitive (e.g. proprietary) data. Collections of digital libraries have to be accessed individually and through non-uniform interfaces. By introducing a level of abstraction, a Meta-Digital Library or MetaDL, users gain a central access portal that allows for prioritized queries, evaluation and rating of the results, and secure negotiations to obtain primary data. This paper demonstrates the MetaDL architecture with an application in brain imaging research, BrassDL, the Brain Support Access System Digital Library. BrassDL is currently under development. This paper describes a theoretical framework behind it, addressing aspects from metadata extraction and system-supported negotiations to legal, ethical and sustainability issues.

1 Introduction

Traditional digital library systems have certain limitations when dealing with complex or sensitive (e.g. proprietary) data. This is true especially in cases where this data may be very useful to a large group of users but the owners have certain valid restrictions in allowing ubiquitous sharing of the data. Examples of such data can be found in medical, scientific, commercial, entertainment, security and other applications. Distributed access to this information necessitates alternative mechanisms of sharing. This paper describes a new type of digital library (DL) model framework called **MetaDL** that allows information sharing even when distribution limitations are present.

A MetaDL contains only data about data, referred to as *metadata*, and not the data themselves. Through a standard of metadata representation, sensitive objects can be securely and efficiently accessed, traded, and evaluated in a summary form. This not only protects the original data from malicious abuse but

M. Agosti and C. Thanos (Eds.): ECDL 2002, LNCS 2458, pp. 374–389, 2002.

also makes highly heterogeneous objects interoperable and amenable to being pooled for various purposes. MetaDLs are user-centered because they provide the user with a one-stop interactive interface that is personalizable (user defines priority and mode of data access), focused on satisfying user needs (built-in evaluation [10] operates on this basis) and lightweight (not dealing with cumbersome primary data).

This paper describes a MetaDL implementation in the area of neuroscience where there is a perceived need [6,21] for sharing valuable human brain data to facilitate meta-analysis, integrative investigations and data mining, and thus make progress towards the larger goal of understanding the brain. Proponents of data access have called for public dissemination of scientific data sets for a decade [33,15]. This kind of access has been attempted in certain fields, for example genomics [31], while it is still largely under discussion in other fields like neuroscience [35]. Among the concerns researchers face is that unfettered access to raw data may work against the interests of the data suppliers, as when their own data is used to anticipate their research [26]. The MetaDL approach proposed here can be applied to protect the interests of the data suppliers but still allow information sharing at several different levels.

The example MetaDL implementation is called **BrassDL**, the **Brain Access Support System Digital Library**. In it, different types of data are represented by metadata: multiple types of scans and datasets, subject data, experiments, methods and results. BrassDL is intended to provide the members of the brain imaging community with a resource that addresses many needs. It allows them to gain an overview of each other's work, search a metadata library of this work, and formulate requests for data sets that augment their available data. It is also designed to provide a negotiation and feedback system between data owners and data users that benefits the users (by making more data available) and the owners (by evaluating the data and thus making it more valuable). The philosophy of the design aims at providing user flexibility (e.g., user can revise or withdraw metadata once posted), simplicity (e.g., simple and uniform method of data entry and simplicity in data sharing), security and ethics in data sharing, and automation.

The rest of paper is organized as follows. Section 2 describes related work. Section 3 presents the MetaDL architecture. Section 4 describes the BrassDL system. Section 5 provides an incentive model. Section 6 concludes the paper.

2 Related Work

The concept of using metadata in place of desired data for indexing and searching is not a new one. The earliest use of the idea may have been in public records offices employing *collection level description* over a century ago [34], motivated by the desire to have remote (albeit limited) access to voluminous records data by way of summaries of holdings. Museums and libraries continue to use collection level description to allow indexing and searching of materials that have not yet been extensively examined or annotated, such as the archives of a famous person

[29], which might be described in a catalog by the number and dates of letters, diaries, and photographs that the collection contains.

In more recent digital collections, where the pooling of large amounts of digital information might seem to obviate the need of using metadata descriptions as a "stand-in" for data, metadata remains valuable for its abstraction of data. Scientific data repositories like GenBank [16], the European Computerized Human Brain Database (ECHBD) [9], and the fMRI Data Center [18] typify this new kind of system. GenBank is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. Records in GenBank contain sequences and data such as sequence description, source organism, sequence length, and references. ECHBD is a 3D computerized database for relating function to microstructure of the cerebral cortex of humans. ECHBD collects homogeneously processed data, and then distributes these back to the submitters. The fMRI Data Center is a database established specifically for the storage and sharing of fMRI data from cognitive studies published in specific journals. It also allows for the mining of highly heterogeneous and voluminous fMRI data. Our proposed MetaDL architecture is also a digital collection; however, it differs from the above projects is that it collects only metadata and links to actual data resources. In the case of BrassDL, the advantages are twofold. On one hand, due to the compactness of metadata, the system is more scalable, can cover different brain science areas such as studies using fMRI, MRI and PET, and stores metadata from studies of different types such as journal publications, scientific meeting presentations, and formal but unpublished studies, and thus can provide a more complete research information repository for neuroscience study. On the other hand, since the structured metadata capture the important features of the raw data, our system does not lose functionality in terms of finding information. Actually, it collects exactly all the information that is expected to help users find what they look for.

Similar ideas are present in the Stanford Digital Library metadata architecture [2], the Alexandria Digital Library architecture [12], and the Common Data Model [13,14], and in the BrainMap [11], BioImage Database Project [5], MARIAN [17], ARION [19], SenseLab [7], and GEREQ [1] systems. In all of these, metadata descriptions are used to link to data from external sources that are never themselves integrated into the system. The Stanford architecture was designed with traditional text-based library documents in mind, and sets up a multi-source metadata index that allows users to search many repositories with a single query. The Alexandria architecture was demonstrated with earth science data, and features metadata-based indexing and searches on a centralized server with ties back to data repositories. The Common Data Model is a framework for integrating neuroscience data from the perspective of mediating data exchange between independent and heterogeneous data resources. BrainMap is a software environment for meta-analysis of the human functional brain-mapping literature. Its purpose is to help researchers understand the functional anatomy of the human brain and the studies that have been done on it through access to current imaging-derived research. The BioImage Database Project stores biolog-

ical data obtained using various microscopic techniques. Data are stored along with metadata describing sample preparation, related work, and keywords. Issues of data access and ownership are discussed in [5], but no specific system of access and usage control is offered. MARIAN is a system for integrating data from various repositories of theses and dissertations into a single view. ARION facilitates better searching and retrieval of digital scientific collections containing data sets, simulation models and tools in various scientific areas. SenseLab is a repository of chemosensory receptor protein and gene sequence data that integrates sequences from 100 laboratories studying 30 species. GEREQ (GEography REsource discovery and Querying management project) is a system for indexing and searching large geographic databases using metadata representations. All the systems mentioned here can be considered "special cases" of MetaDLs, although they lack some proposed MetaDL features. All are centralized indexes of distributed data sources, as with a MetaDL, but a MetaDL adds sophisticated access control and control of metadata indexing by data source owners.

Metadata descriptions are used as a means to organize, index, and query medical or similarly sensitive data in the NeuroGenerator project [32], the fMRI Data Center [18], mentioned above, current pharmaceutical data warehouses [3], and a system proposed by researchers at Rutgers in 2000 [24]. The NeuroGenerator database system is based on the concept of storing raw data at a central site and making processed versions of it available. Researchers submit raw PET and fMRI data, along with detailed metadata describing its collection, and the central site uses current methods for data processing to integrate it into homogeneous collections. Users can then access collections that correspond to the data and processing type they are interested in. The fMRI data center receives data in concert with publications in certain journals that require a contribution of data to the center as a condition of publication. Data formatting and metadata tagging is done by the originating sites. Pharmaceutical data warehouses use metadata records to integrate various existing stores of data and allow for data indexing and advertisement for sale. The proposed Rutgers system aims to facilitate peer-to-peer sharing of datasets by using a centralized site to allow researchers to register what data is available, and under what conditions. The central site would use cryptographically signed exchanges between data providers and users to create a binding agreement before data is released. Although described in news format in 2000 in *Science*, publications have not yet been made on the proposed system.

Considerable work has been done on the development and promotion of metadata standards for creation and dissemination of metadata. The Dublin Core Metadata Initiative [8] is an organization dedicated to promoting the widespread adoption of interoperable metadata standards and to developing specialized metadata vocabularies for describing resources. The so-called "Dublin core" of metadata elements is used as a basis for many metadata schemas. The METAe project [27] aims to ease the automated creation of (technical, descriptive, and structural) metadata during capture or digitization. The aim of the Metadata Tools and Services project [28] — also known as MetaWeb — is to develop in-

dexing services, user tools, and metadata element sets in order to promote the metadata on the Internet. The BrainML [4] project is creating a language and organizational ontology for metadata for neuroscience. The MetaDL model adapts techniques from these previous studies and systems to the task of providing a comprehensive digital information service for domains with complex or sensitive information.

3 MetaDLs

A MetaDL exists within a two-tier architecture (Figure 1) that supports two endeavors: searching for data via metadata, and sharing these data in a secure fashion. Tier 1 consists of autonomous DLs containing data, while Tier 2 systems contain data about the Tier 1 DLs and permit browsing and searching for primary data that are contained in Tier 1 DLs. Tier 1 DL systems by definition must contain actual data, while Tier 2 MetaDL systems by definition must contain only metadata. For this reason, the two tiers contain non-overlapping sets of systems.

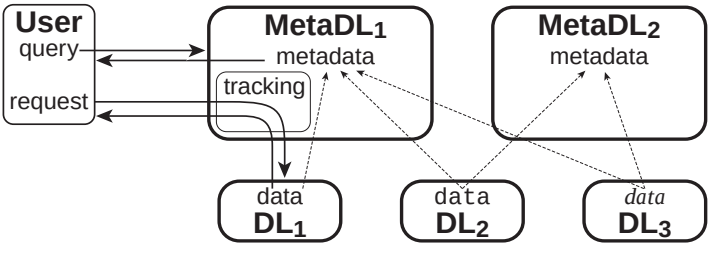

Fig. 1. How MetaDLs improve use of digital libraries (DLs). Rather than accessing every DL separately using each DL's individual interface (and possibly finding no matches or accessing a DL that is not appropriate), a user can query multiple DLs via a MetaDL that has collected metadata from each DL (dashed arrows), giving the user a homogeneous interface. Requests for actual data are facilitated and tracked by the MetaDL. Multiple MetaDLs can exist with individual priorities and interfaces, reflecting different purposes of each MetaDL.

The naming of the tiers reflects the distance from the primary data, Tier 1 being the closest. Any provider can set up and operate her own Tier 1 system, and a wrapper can be used to make existing systems conform to the model. Each system is independent and autonomous, allowing for flexibility in organization and configuration. As an example, a group of Tier 1 providers (e.g. some hospitals in the same country) may create their own Tier 2 system (with information for their Tier 1 DLs — and possibly some other important Tier 1 DLs that they also want to access). In practice, a single universal usage system may not be as efficient because of bandwidth or policy reasons (e.g. national laws). Therefore,

the MetaDL model allows for any number of MetaDL systems providing coverage for possibly overlapping or redundant sets of digital libraries.

Tier 2 systems use metadata submitted by Tier 1 systems to provide an overview of data resources. General users access Tier 2 systems. Local users of a Tier 1 system have the option to query their Tier 1 system directly (for local operations only) or to access a Tier 2 system (for accessing data from many providers). If the user needs to access any of Tier 1 DLs, then she can use a Tier 2 system to authenticate herself, negotiate conditions and access the data. In what follows, a more detailed description of the tier functions is provided.

3.1 Tier 1 – Primary Data Management

A Tier 1 system does the following: (a) keeps track of the ownership, status and access rights of its objects; (b) authenticates its users, to verify and determine their access rights; (c) records user requests; (d) validates and serves the requests of its interactive users and of Tier 2; (e) can store alert conditions, for notifying the users on specific conditions — like insertions of new data sets; (f) supports different modes of data interchange between the data requester and the data owner; and (g) provides object information (usually public metadata), and optionally user and object usage information to Tier 2. Every data provider is encouraged to provide metadata by a mechanism of incentives and a built-in negotiation system that supports data exchanges in Tier 2. General users of a MetaDL always connect first to Tier 2, which provides a friendly one-stop (graphical) interface and transparently forwards requests to the appropriate Tier 1 DLs. All object handling is done in Tier 1, and the object accessibility is actually a property of each object (object-oriented design) and can be different in different objects in the same collection.

Tier 1 systems set conditions on sharing the data, and Tier 2 provides a front-end interface to these conditions, as well as support for a notification system connected to relevant legislation and other appropriate resources. The amount of the information, including both metadata and data, which are given to a user will depend on the specific Tier 1 DLs that contain the information, and their configuration and even on the data and metadata themselves. For example, a Tier 1 DL may contain some public-access sets for promotional purposes or an educational Tier 1 DL may contain and provide only public-access sets of "clean datasets" or benchmarks donated for educational purposes.

Tier 1 functionalities are implemented through a software tool distributed from the MetaDL website. This software can help data owners index their autonomous DLs while formatting metadata for Tier 2 systems in a uniform way.

3.2 Tier 2 – Meta-data Information Service

The contents in a Tier 2 system are (a) static descriptions of Tier 1 DLs — what Tier 1 servers exist, which collections they contain, information (in structured metadata and free text descriptions) about them, etc.; (b) dynamically generated descriptions for Tier 1 DLs and their objects, as they are produced by the public

metadata that Tier 1 systems provide; (c) dynamically generated object usage and user information (such as object tracking, data use statistics, user alert data, etc.), both provided by Tier 1 systems and obtained from Tier 2 usage monitoring; and (d) other public data such as generic demographic information on relevant research activities. All these are actually metadata that relate user requests to Tier 1 DLs and objects in them. In Tier 2, a user searches, browses and manipulates information through a common interface, not accessing the original data directly. Once a user has identified a dataset he would like, he enters a request through Tier 2. A Tier 2 negotiation component (see Section 3.3) supports user-to-user data interchange: authenticate users, mediate between Tier 1 and Tier 2 components, manage transactions, and track condition.

3.3 Negotiation Model

The negotiation system [20] is composed of a set of transactions, see Section 4.3 for an example. The Negotiations between users and Tier 1 DLs need to fulfill the following requirements:

- *Proof of completion*: provide both parties with proof that each stage of the negotiation has completed.
- *Privacy of communication and security of transmission*: keep users' searches of and requests from DLs confidential, since this information may allow outsiders to draw conclusions as to the nature of this research. It is also obvious that data transmitted must not be usable by outsiders.
- *Verification of identity*: allow a data provider to verify a recipient, and allow a recipient to verify a data source. Especially over the Internet, both parties need to be certain about the other's identity. On the machine level, this is an old problem; for specific MetaDLs (such as BrassDL), there is the additional layer of being certain about the other side's credentials (e.g. recipient is a PhD).
- *Conditions of use*: allow the owner to make distinctions between users or their use of the data (e.g. a doctor may be allowed access to help solve a particular patient's problem, but access for researchers in general may be denied). Without this flexibility, the data owner would be forced to the lowest common denominator: deny access to everybody.
- *Comment on data quality*: protect the user from being given inadequate or substandard data by allowing the user to provide feedback on the data.

For example, two clinicians want to test their experiment on a larger number of subjects, and query a MetaDL geared towards their research for similar subject data. Of the descriptions they receive back, they pick one dataset and request it. The MetaDL forwards this request to the dataset's owner, who demands that the clinicians sign a privacy and nondisclosure agreement, as well as list the owner in any publication resulting from their use of this dataset. The clinicians agree, and the dataset is sent to them. They work on their combined data, and find that the new data augments the old very nicely, so they send a favorable review

of the dataset back to the MetaDL. The owner of the data is notified that this review has been added onto the dataset's history. Later, the clinicians write a paper about their work but omit to list the owner of the dataset they requested. The owner spots the publication and complains, using the tracking information from the MetaDL and the digitally signed agreements as proof that his data was unproperly used.

There is a stage at which a public Tier 2 system would like some financial support in order to continue to offer services. One option is to charge a membership fee, which is variable according to entity (institution, lab, researcher, student) on an annual basis. A payment scheme would be beneficial for large data suppliers but detrimental to students who have very little to contribute to the system (aside from their potential as future members of the community). For users, there is some compensation in the form of reputation: data owners gain reputation by producing good datasets, while users gain by providing helpful feedback. Good datasets benefit the users directly; quality feedback improves the accuracy of queries.

4 BrassDL

Recent non-invasive scanning techniques in neuroscience have resulted in an explosive growth of research, the aim of which is to discover how the brain functions. Mapping structure/function relationships is a grand challenge problem in neuroscience [30]. Using brain multi-sensor technologies, such as Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), Magnetic Resonance Spectroscopy (MRS), functional MRI (fMRI), Single-Photon Computed Tomography (SPECT), etc., novel discoveries can emerge with appropriate access and analytic strategies [36]. BrassDL aims to provide researchers with a moderated forum for exchanging data, ideas, and commentary in neuroscience.

Primary neuroscience data is very expensive and difficult to share. Most remain inaccessible to researchers outside of a particular project [21,22,25]. Currently, a traditional digital library containing original brain data is not realistic or practical due to the issue of data ownership. While researchers may be willing to exchange data in some circumstances, many are not willing to embrace a system that will allow unfettered access to and use of their hard-won data. In spite of many efforts to share this data [11,13,14,18,24,35], the state of the art is that each laboratory follows its own methodology, collects its own data and shares this data only through publications or in a limited manner [24]. New technologies have also resulted in an explosion of new research findings. A huge number of diverse, and possibly non-interoperable datasets and methods are being accumulated in various laboratories, often not known to other labs. From the standpoint of efficiency, it would be good to share these datasets, especially due to the very expensive equipment required and the high cost of each scan (on the order of several hundred to thousand US dollars [23]). In addition, to increase statistical power of analyses, it may be useful to integrate existing data with newer data where possible. For most labs right now, one must find old datasets

on local disc archives. Although assumptions and technologies have changed, much of the collected information is in raw format and can be reinterpreted.

There are two reasons behind this lack of a comprehensive collection point. One is technical: due to data complexity, the diversity of formats, diverse experimental assumptions and scanner conditions, it is often requires considerable effort to combine data or results [32]. For example, if two studies aimed to measure the same phenomenon, but collected data from different scanners, it may be difficult or undesirable to combine the data depending on the research context. The second obstacle is non-technical and involves ownership, security, and privacy concerns that make direct data access non-feasible except in small-scale situations among a small number of users. To overcome these obstacles, the MetaDL concept is a good choice. Its two-tier organization permits better scalability, allows autonomous operations on each data provider, eases adoption of the system, and is able to integrate different providers while not sacrificing data ownership and access rights.

4.1 BrassDL Model

We propose a MetaDL implementation in the area of neuroscience that we call **BrassDL**, for the **BRain Access Support System Digital Library**. BrassDL not only catalogs data of different types and from different sources, but also provides a negotiation and feedback system to facilitate multi-level data sharing and data evaluation. Unlike existing data sharing models [11,18], BrassDL does not restrict itself to published data but starts bottom up, from the datasets that exist or are being developed in a laboratory. BrassDL can attempt to evaluate metadata according to different criteria, such as consistency and reasonableness. It also monitors user interaction with the system and mediates different types of negotiations among users.

As a MetaDL, BrassDL exists within a two-tier architecture, as shown in Figure 2. Tier 1 consists of autonomous DLs controlled by data providers (or BrassDL Partners). These DLs contain primary data, metadata and information about their access conditions. BrassDL is a Tier 2 system and contains data about the Tier 1 DLs. Tier 2 provides functionalities including browsing and searching for data that are contained in Tier 1 DLs as well as tracking the primary data exchange between users and BrassDL Partners.

A user usually interacts with BrassDL in the following way: initially, to query content of all data providers by searching the metadata that BrassDL has collected from each provider; subsequently, to request data sets from a particular provider; and finally, to submit feedback about the data sets that were received from the provider as additions to the provider's track record.

The collection of metadata for BrassDL is shown in Figure 3. BrassDL cataloger is used to organize the different data in a local DL and at the same time extract metadata, both for the benefit of the local site and, optionally, for submission to BrassDL to benefit the community. A data provider can use the cataloger to enter experiments, methods, tasks and datasets locally and in a uniform way with other data providers. The cataloger then enables metadata

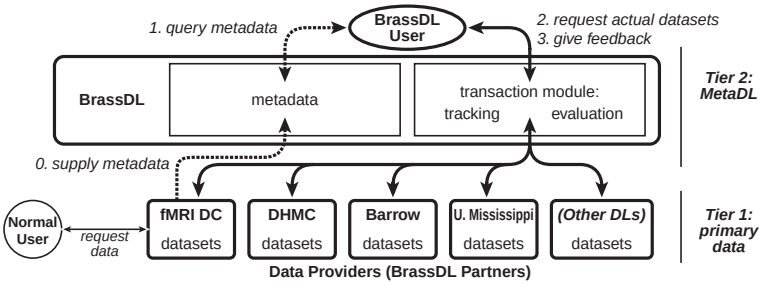

Fig. 2. User-BrassDL interactions. Tier 1 interactions are between the user and a data provider only: the user requests data; the data provider verifies the request and returns data if the request is granted. In Tier 2, the user interacts with BrassDL: initially, to query content of all data providers by searching the metadata that BrassDL has collected from each of its providers (dashed arrow); subsequently, to request datasets from a particular data provider; and finally, to submit feedback about the datasets which are passed on to the provider as additions to the provider's track record.

extraction from the local entries and with the consent of the owner. Each data provider can formulate policies to restrict usage to acceptable terms; specific agreements are made when a user requests a data set.

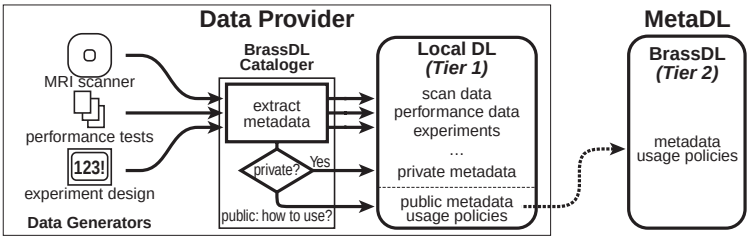

Fig. 3. Contributing to BrassDL. A BrassDL data provider uses an organizer tool provided by BrassDL to manage its data (e.g. scans, performance data, experiment descriptions) in a local DL. As data are added, metadata descriptions are created (e.g. scan kind and size, high-level descriptions of experiments) and stored together with the actual data. To contribute to BrassDL, metadata descriptions are submitted to the BrassDL system, together with any preliminary or general restrictions of use (e.g. "academic use only, except by arrangement").

From the user's point of view, there are two options for accessing information about a given topic in brain imaging (see Figure 4). On one hand, he can visit individual online resources, such as BrainMap or fMRI Data Center. On the other hand, he can visit BrassDL for a synergistic superset to existing brain imaging resources; since BrassDL (a) links to datasets, not all of which have been published, (b) makes links from data to publications and vice versa, and (c) adds

metadata descriptions of primary data that also become descriptions of the associated publications. BrassDL provides specific additional datasets via its data providers to help the user produce a publication which in turn is added to the pool of the traditional online resources. Established sites offer either an overview of the field or published datasets, and completed research can be submitted to these sites. BrassDL is more useful during the development and verification of a hypothesis by augmenting available data resources. Data providers' results are incorporated individually, rather than as a package.

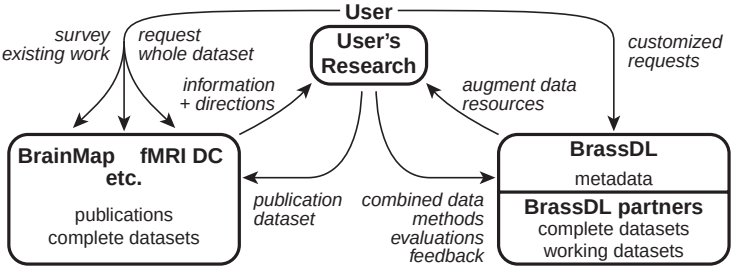

Fig. 4. Information flow in brain imaging research. A user researching a given topic can look at BrainMap, the fMRI Data Center, and other resources to get an overview of existing, published work. BrassDL is a synergistic superset to existing brain imaging resources as it (a) links to datasets, not all of which have been published; (b) makes links from data to publications and vice versa; and (c) adds metadata descriptions of primary data that also become descriptions of the associated publications, allowing meta-analysis. BrassDL provides specific additional datasets via its data providers to help the user produce a publication which in turn is added to the pool of the traditional online resources.

Figure 4 shows that BrassDL is a mediator for data sharing among different types of repositories. Similarly, BrassDL also acts as a mediator between different types of users, each of whom may contribute different services to the whole community (Figure 5). As the foundation of the system, data providers contribute metadata to the system (but keep the actual data). System subscribers can use these metadata in their research and return feedback in form of results, or make requests for specific data types. Evaluators classify data according to quality with regards to different criteria. Clinicians use actual cases for comparison and add expert knowledge to the system. Student users learn from textbook cases and can post their questions. Finally, the comments about the datasets flow back into each data provider's track record.

4.2 BrassDL Query and Data Evaluation

One strength of the abstraction layer that a MetaDL (and thus BrassDL) provides is the ability of the user to specify queries with different weights on prop-

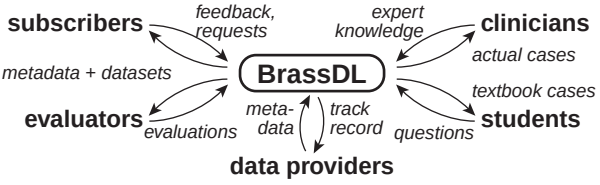

Fig. 5. Usage of BrassDL by different users and providers. Each type of user benefits differently from BrassDL and contributes parts that are useful for other users. For example, a case available from a data owner is evaluated as useful for a certain condition. A clinician then uses this case to treat a patient with the condition and adds her experiences, elevating the case to a textbook case used by medical students.

erties of the metadata being searched (see Figure 6 for a simplistic example). The user can choose which attributes to prioritize in a query in order to match his experiments. The system helps the user find the closest desired dataset based on the criteria he defines.

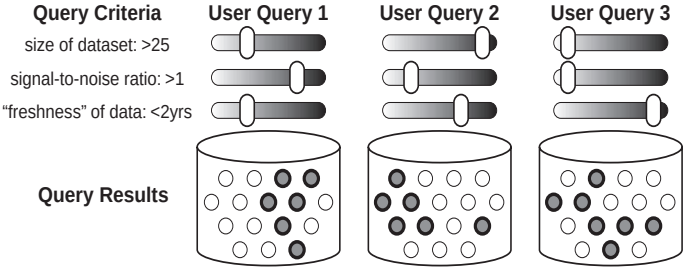

Fig. 6. User-weighted queries. By giving search criteria different weights, the user can customize a query to match the priorities of a research topic. For example, query 1 prefers datasets with a very high signal-to-noise ratio; query 2 selects large new datasets, with size more important than age; while the last query looks for datasets based on their age only.

Metadata queries (Tier 2) are evaluated separately from negotiations of actual datasets (Tier 1), yet can influence each other, as Figure 7 caption explains. Queries also influence the results of future queries: the metadata of frequently requested datasets will rank higher than unused datasets given equal values for the search parameters. A negotiation may take into account previous queries for datasets so as to not produce a compromising set in the hands of the user (e.g. sufficient data to identify a patient in conjunction with that on file at an insurance company, leading to higher insurance rates) — based on what data the user already has requested and obtained (through BrassDL), the negotiation offer from a data provider may include additional requirements. In [10], we de-

scribe a built-in evaluation metric that measures how closely BrassDL contents match the interests of users.

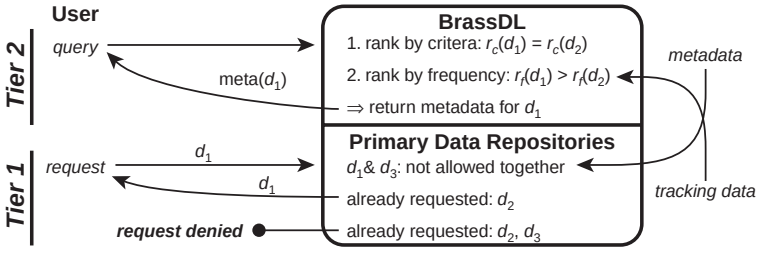

Fig. 7. Evaluation on each tier influences user queries and requests. Results to user queries are primarily ranked by search criteria, but in case of ties, the ranking can include the frequency of use, determined by the tracking data from all previous accesses. Similarly, restrictions of use (e.g. "datasets 1 and 3 constitute a compromising set") can be incorporated at the query level, hiding or flagging those datasets that would violate the restriction.

4.3 Protocol for Primary Data Sharing

The following describes a negotiation between a user, BrassDL, and a data provider (see Figure 8). The user queries BrassDL and receives query results in the form of metadata. Out of these, the user chooses which actual dataset(s) is most likely of benefit and formulates a request, which is forwarded to the data provider. The provider replies with a set of usage requirements (e.g. nondisclosure, clearances, co-authorship). The user signs this binding agreement and returns it to the partner, who then releases the actual dataset to the user. This dialogue of transaction is facilitated and tracked at each stage by BrassDL. Once the research has been concluded and is published by the user, the results are shared with BrassDL in form of evaluations and general feedback, giving the data provider a source of feedback and a record of collaboration.

5 Incentive Model

BrassDL supplements its use of the MetaDL model with a strong incentive model that overcomes the non-technical issues and can be adapted to other MetaDL applications. Since a MetaDL is based on metadata and the user's interaction with this data, a MetaDL would not survive without a good incentive for users to participate. The BrassDL incentive model is designed to answer the following question: "Why would a large neuroimaging laboratory go into the trouble of entering its metadata and why would a researcher want to access it (and enter her own)?" Below we briefly list several incentives for participation:

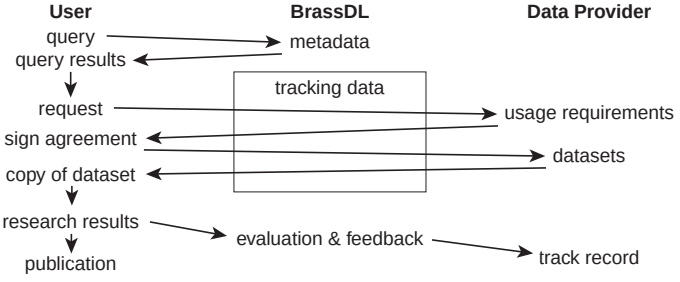

Fig. 8. Sketch of the protocol for data sharing and feedback. BrassDL acts as a mediator between user and data provider (BrassDL partner), initially to narrow the search for a good dataset, then to document the actual negotiation, and finally to store comments about the dataset.

- Visibility - The research activity of a data provider becomes visible to the community. This helps attract potential funding, patients or collaborators.
- Feedback - User feedback on the use of data may provide another means of data evaluation, as well as valuable advice on the improvement of data generation and data quality.
- Software support - BrassDL distributes a software tool to implement Tier 1 functionality. This software can help data providers build their own autonomous DLs, organizing data as they like, and create metadata in a standard fashion.
- Value assessment - Built-in evaluation mechanisms [10] assess the popularity of a dataset based on user demand and user feedback, resulting in automatic dataset ranking.
- Security - BrassDL helps provide for secure direct data exchanges by a mechanism of brokering, mediation, rights management and data tracking. This allows data sharing flexibility and the protection of data ownership and access rights.
- Notification and consultation - A user who places a query can receive future notifications of "similar" work, as defined by her profile, when new datasets are inserted; alternatively, she may receive consultation on how to proceed with future queries. BrassDL facilitates collaboration between organizations who create primary data and organizations who perform advanced data analysis.
- Data sharing management - BrassDL stores and manages each negotiation of data exchange. This helps owners manage their data-sharing activities, since it records cases or evidence of possible misuse and protects owners' rights.

6 Conclusion

We have presented a new framework for digital libraries managing sensitive datasets that have limitations on distribution, and a specific implementation.

For medical neuroimaging data, the MetaDL system extends previous notions of metadata-based digital libraries by (a) not including the original data; (b) supporting the data sharing process and recording the outcomes, (c) providing a uniform metadata description for data, methods, experiments, tasks and subject data, (d) maintaining statistics and demographics of data usage and providing a built-in evaluation standard to provide user incentives, and (e) providing support for meta-analysis of results and studies of research demographics.

References

[1] F. Andres, N. Mouaddib, K. Ono, and A. Zhang. Metadata model, resource discovery, and querying on large scale multidimensional datasets: The GEREQ project. In *Kyoto International Conference on Digital Libraries*, pages 83–90, 2000.

[2] M. Baldonado, C.-C. K. Chang, L. Gravano, and A. Paepcke. The Stanford Digital Library metadata architecture. *International Journal on Digital Libraries*, 1(2):108–121, 1997.

[3] J. S. Barrett and S. P. J. Koprowski. The epiphany of data warehousing technologies in the pharmaceutical industry. *International Journal of Clinical Pharmacology and Therapeutics*, 40(3):S3–13, March 2002.

[4] BrainML functional ontology for neuroscience. brainml.org.

[5] J. M. Carazo and E. H. K. Stelzer. The BioImage database project: Organizing multidimensional biological images in an object-relational database. *Journal of Structural Biology*, 125:97–102, 1999.

[6] M. Chicurel. Databasing the brain. *Nature*, 406:822–825, August 2000.

[7] C. Crasto, L. Marenco, P. Miller, and G. Shepherd. Olfactory Receptor Database: a metadata-driven automated population from sources of gene and protein sequences. *Nucleic Acids Research*, 30(1):354–360, 2002.

[8] The Dublin Core Metadata Initiative. dublincore.org.

[9] European Computerised Human Brain Database (ECHBD). fornix.neuro.ki.se/ECHBD/Database.

[10] J. Ford, F. Makedon, L. Shen, T. Steinberg, A. Saykin, and H. Wishart. Evaluation metrics for user-centered ranking of content in MetaDLs. In *Fourth DELOS Workshop on Evaluation of digital libraries: Testbeds, measurements, and metrics*, Budapest, Hungary, June 2002.

[11] P. T. Fox and J. L. Lancaster. Mapping context and content: the BrainMap model. *Nature Reviews Neuroscience*, 3(4):319–321, 2002.

[12] J. Frew, M. Freeston, N. Freitas, L. Hill, G. Janee, K. Lovette, R. Nideffer, T. Smith, and Q. Zheng. The Alexandria Digital Library architecture. *International Journal on Digital Libraries*, 2(4):259–268, 2000.

[13] D. Gardner, M. Abato, K. H. Knuth, R. DeBellis, and S. M. Erde. Dynamic publication model for neurophysiology databases. *Philosophical Transactions of the Royal Society of London: Biological Sciences*, 356(1412):1229–1247, 2001.

[14] D. Gardner, K. H. Knuth, M. Abato, S. M. Erde, T. White, R. DeBellis, and E. P. Gardner. Common data model for neuroscience data and data model exchange. *Journal of the American Medical Informatics Association*, 8(1):103–104, 2001.

[15] M. Gelobter. Public data-archiving: a fair return on publicly funded research. Psycoloquy: 3(56) Data Archive (3), 1992.

[16] GenBank genetic sequence database. www.ncbi.nlm.nih.gov/Genbank.

[17] M. A. Gonçalves, R. K. France, and E. A. Fox. MARIAN: Flexible interoperability for federated digital libraries. In *Proceedings of ECDL 2001, the 5th European Conference on Research and Advanced Technology for Digital Libraries*, pages 173–186, Darmstadt, Germany, September 2001. Springer.

[18] J. S. Grethe, J. D. Van Horn, J. B. Woodward, S. Inati, P. J. Kostelec, J. A. Aslam, D. Rockmore, D. Rus, and M. S. Gazzaniga. The fMRI Data Center: An introduction. *NeuroImage*, 13(6):S135, 2001.

[19] C. Houstis and S. Lalis. ARION: An advanced lightweight software system architecture for accessing scientific collections. *Cultivate Interactive*, 4, May 2001.

[20] S. Kapidakis, F. Makedon, L. Shen, T. Steinberg, and J. Ford. A negotiation model for mediated sharing of private digital data, 2002. In preparation.

[21] S. H. Koslow. Should the neuroscience community make a paradigm shift to sharing primary data? (editorial). *Nature Neuroscience*, 3:863–865, September 2000.

[22] S. H. Koslow. Sharing primary data: A threat or asset to discovery? (editorial). *Nature Reviews Neuroscience*, 3:311–313, April 2002.

[23] D. Krotz. PET and MRI race to detect early Alzheimer s. *AuntMinnie.com*, January 2001.
www.auntminnie.com/index.asp?Sec=nws&sub=rad&pag=dis&ItemId=50098.

[24] E. Marshall. Downloading the human brain with security. *Science*, 289(5488): 2250, September 2000.

[25] E. Marshall. A ruckus over releasing images of the human brain. *Science*, 289(5484):1458–1459, September 2000.

[26] E. Marshall. DNA sequencer protests being scooped with his own data. *Science*, 295(5558):1206–1207, February 2002.

[27] The MetaE metadata engine project. meta-e.uibk.ac.at.

[28] The Metadata Tools and Services Project (MetaWeb).
www.dstc.edu.au/Research/Projects/metaweb.

[29] P. Miller. Collected wisdom. *D-Lib Magazine*, 6(9), September 2000.

[30] The Organization for Human Brain Mapping (OHBM).
www.humanbrainmapping.org.

[31] L. Roberts, R. J. Davenport, E. Pennisi, and E. Marshall. A history of the Human Genome Project. *Science*, 291(5507):1195, 2001.

[32] P. Roland, G. Svensson, T. Lindeberg, T. Risch, P. Baumann, A. Dehmel, J. Frederiksson, H. Halldorson, L. Forsberg, J. Young, and K. Zilles. A database generator for human brain imaging. *Trends in Neuroscience*, 24(10):562–564, 2001.

[33] J. R. Skoyles. FTP internet data archiving: A cousin for psycoloquy. Psycoloquy: 3(29) Data Archive (1), 1992.

[34] M. Sweet and D. Thomas. Archives described at collection level. *D-Lib Magazine*, 6(9), September 2000.

[35] J. D. Van Horn, G. J. S., P. Kostelec, J. B. Woodward, J. A. Aslam, D. Rus, D. Rockmore, and M. S. Gazzaniga. The Functional Magnetic Resonance Imaging Data Center (fMRIDC): The challenges and rewards of large-scale databasing of neuroimaging studies. *Philosophical Transactions of the Royal Society of London: Biological Sciences*, 356(1412):1323–1339, 2001.

[36] D. A. Wagner. Early detection of Alzheimer s disease: An fMRI marker for people at risk? *Nature Neuroscience*, 3(10):973–974, 2000.