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

Fillia Makedon, James Ford, Li Shen, Tilmann Steinberg
The Dartmouth Experimental Visualization Laboratory
Department of Computer Science
{makedon, jford, li, tilmann}@cs.dartmouth.edu

Andrew Saykin, Heather Wishart
Brain Imaging Laboratory
Dartmouth Medical School
{saykin, wishart}@dartmouth.edu

Dartmouth College
Hanover, NH 03755, USA

Sarantos Kapidakis
Department of Archive and Library Sciences
Ionian University, Greece
sarantos@ionio.gr

Abstract: Traditional digital library systems have difficulties when managing heterogeneous datasets that have limitations on their distribution. 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 transactions to obtain primary data. This paper demonstrates the MetaDL architecture with an application from human brain neuroimaging research, BrassDL, the Brain Support Access System Digital Library. This is the first such system that covers all aspects of a digital library for sensitive and complex human brain data, from secure acquisition and access, user to user system-supported transactions, to legal, ethical and sustainability issues.

I. 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 of this data have certain valid restrictions in allowing ubiquitous sharing of the original information. Examples of such data can be found in medical, scientific, commercial, entertainment, security and other applications. Access to this information necessitates alternative mechanisms of sharing, especially where public funding has been involved. This paper describes a digital library (DL) system called MetaDL to provide such a mechanism to allow information sharing even when these limitations present.

A MetaDL contains only data about data, referred to as metadata, and not the data themselves. Through a standard of metadata generation, sensitive objects are securely and efficiently accessed, traded, and evaluated. This not only protects the original data from malicious abuse but also makes highly heterogeneous objects interoperable and amenable to being pooled for large analyses. Also, 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 operates on this basis) and lightweight (not dealing with cumbersome primary data).

This paper demonstrates a MetaDL implementation in the area of neuroscience where there is a perceived need [Chi2000, Kos2000] for sharing valuable human brain data to facilitate meta-analysis, integrative investigations and data mining, and 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 [Sko1992, Gel1992]. This kind of access is the norm in certain fields, for example genomics [RDPM2001], while it is still under discussion in other fields like neuroscience [HGKWA2001]. 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 undermine their research [Mar2002]. 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.

This MetaDL is called BrassDL, the Brain Access Support System Digital Library, where different types of data become interoperable: multiple types of scans and datasets, subject data, experiments, methods and results. BrassDL is intended to provide the members of 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 transaction and feedback system between data owners and data users that benefits the user (by making more data available) and the owner (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 II describes related work. Section III presents the MetaDL architecture. Section IV describes BrassDL system. Section V provides a discussion on several issues. Section VI concludes the paper.

II. 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 [ST2000], when it was motivated by the desire to have remote (albeit limited) access to voluminous records data by way of summaries of holdings. Museums and libraries have used this kind of metadata 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 [Mil2000], which might be described 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. Systems like GenBank [GB], the European Computerized Human Brain Database [ECHBD], and the fMRI Data Center [GHWIK2001] 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 the GenBank Locus Number, 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. Another goal is that the database may allow for the mining of highly heterogeneous and voluminous fMRI data. Our proposed MetaDL architecture is also a digital collection. However, the difference from the above projects is that we collect only metadata and links to the 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 levels such as journal publications, scientific meeting presentations, formal funded studies, and thus 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.

A similar idea is presented in the Stanford Digital Library metadata architecture [BCGP1997], the Alexandria Digital Library architecture [FFFHJ2000], and the Common Data Model [GKAEW2001, GAKDE2001], and in the BrainMap [FL2002], BioImage Database Project [CS1999], MARIAN

[MF2001], ARION [HL2001], SenseLab [HSSN1997, SHSNM1999, SMNMS2000, CMMS2002], and GEREQ [AMOZ2000] systems. In all of these, metadata is used to link to data from external sources that is never itself 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 image-derived research. BrainMap relates brain locations to human behavioral functions and data sources: brain regions are mapped to clinical conditions associated with them, and behavioral function are mapped to the brain regions supporting that behavior. The BioImage Database Project stores biological 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 [CS1999], 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 is used as a means to organize, index, and query medical or similarly sensitive data in the NeuroGenerator project [RSLRB2001], the fMRI Data Center [GHWIK2001], mentioned above, current pharmaceutical data warehouses [BK2002], and a system proposed by researchers at Rutgers in 2000 [Mar2000b]. 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 requiring 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 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 [DCMI] is an organization dedicated to promoting the widespread adoption of interoperable metadata standards and developing specialized metadata vocabularies for describing resources that enable more intelligent information discovery systems. The so-called "Dublin Core" of metadata elements is used as a basis for many metadata schemas. The METAe project [METAE] 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 [MW]—also known as MetaWeb—is to develop indexing services, user tools, and metadata element sets in order to promote the use of, and exploitation of metadata on the Internet. The MetaDL model adapts

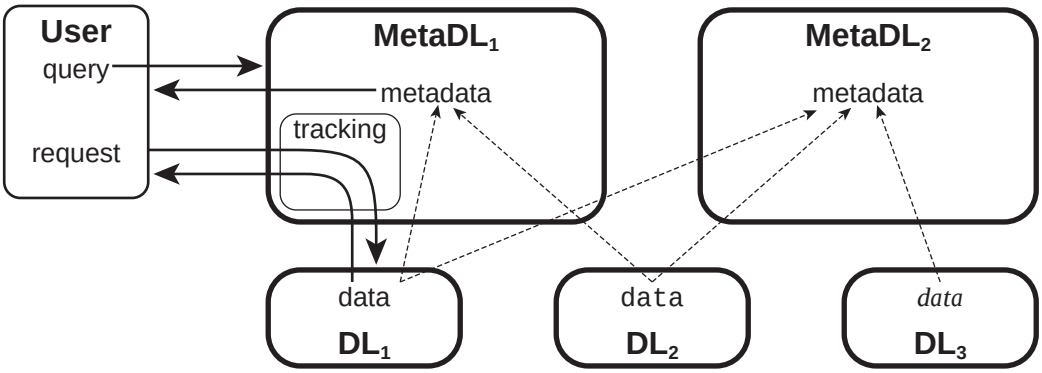

Figure 1: How MetaDLs improve use of digital libraries (DLs). Rather than accessing every DL separately using each DL's individual interface (and possibly finding no matches or accessing a DL that is not appropriate), a user can query multiple DLs via a MetaDL that has collected metadata from each DL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

techniques from these previous studies and systems to the task of providing a comprehensive digital information service for domains with complex or sensitive information.

III. MetaDLs

A MetaDL exists within a two-tier architecture (figure 1) that supports two endeavors: searching for data (and methods) via metadata, and sharing this data in a secure fashion. Tier 1 consists of autonomous DLs containing data, each with an interface allowing it to specify access conditions. Tier 2 systems contain data about the Tier 1 Digital Libraries and permit browsing and searching for data (including methods) 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.

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 such 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) that may prevent this to happen. 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 a particular interface to Tier 1 systems that provides an overview of data resources to the user. 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 connects to Tier 1 to authenticate herself, negotiate conditions and access the data. Tier 1 DLs adhere to specific rules, to ensure maximum functionality on Tier 2 collaboration. In what follows, a more detailed description of the tier functions is provided.

III.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 their requests; (d) validates and serves the re-quests 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), user and object usage information to Tier 2. Every data provider is encouraged to provide metadata by a mechanism of incentives, value added rewards, and a built-in support system for data exchange in Tier 1. Local users, or other authorized ones, can connect directly to a Tier 1 system through a command line or a graphical interface, and locate (browse or search) and access the data in it. General users always connect to Tier 2, which provides a friendly one-stop (graphical) interface and transparently forwards their requests to the appropriate Tier 1 DLs. All object handling is done in Tier 1, and the object accessibility is actually a property of the object (object-oriented design) and can be different in different objects in the same collection.

Tier 1 negotiations include conditions on sharing the data, Tier 2 provides a front-end interface to these conditions, as well as support for policy and decision-making through a notification system connected to relevant legislation. 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 as a software tool distributed from the MetaDL website. This software can help data providers build their own autonomous DLs while also formatting metadata for Tier 2 systems in a uniform way.

III.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), produced by the data that Tier 1 provides; and (d) other public data such as generic demographic information on neuroscience 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 to have, he enters a request through Tier 1. A Tier 2 system assists with the user-to-user data interchange or transaction by authenticating the user once and acting as a mediator between Tier 1 components involved, managing negotiations and passing them the user authentication automatically.

The two-tier architecture permits better scalability, allows autonomous operations on each data provider, easier adoption of the system and is able to combine metadata and data from different providers while not sacrificing data ownership and access rights.

III.3. Transaction Model

Transactions 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 transaction 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 make conclusions as to the nature of the 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. The goal is to form a "web of trust". 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 has to be a doctor, not a nurse).
- *Conditions of use:* allow the owner to make distinctions between users or their use of the data (e.g. doctor vs. researcher: is the goal to solve a specific patient's problem, or is it a publication of a novel idea). Without this flexibility, the data's 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 label 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 receives a copy of this review to add 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 as proof that his data was used as a basis for the paper.

There is a stage at which a digital library would like to receive some compensation for the use of their data. There are several possibilities, most of which are likely to exclude a number of users. One way 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 education and potential as future members of the community). Another possibility used to varying degrees of success is the sponsorship of certain datasets, i.e., a data supplier pays the MetaDL to present his or her datasets along with normal query results if the sponsored dataset relates to the query (similar to Google's "Sponsored Links"). In the absence of money-based solutions, a good currency is reputation: data owners gain reputation by producing good datasets, while users gain by providing accurate and helpful feedback. Good datasets benefit the users directly; quality feedback improves the accuracy of queries.

IV. 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 and is the focus of the Organization for Human Brain Mapping [OHBM]. Through a massive combination and correlation of 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 [Wag2000]. For example, from a top-down perspective, with large, multisensor datasets, study-specific methodological variation is likely to cancel out, permitting an improved signal-to-noise ratio in research applications. From a bottom-up perspective, there is the potential to enhance clinical prediction for individual cases where a suffi-

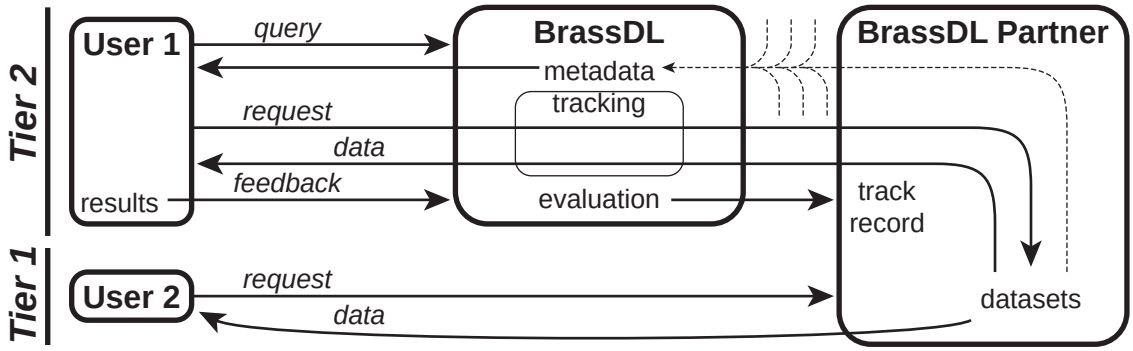

Figure 2: User-BrassDL interactions. Tier 1 interactions are between the user and a BrassDL partner only: the user requests data; the BrassDL partner verifies the request and returns data if the request is granted. In Tier 2, the user interacts with BrassDL: initially, to query content of all BrassDL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

cient normative database is available. BrassDL aims to provide researchers with a moderated forum for exchanging data, ideas, and commentary in the field of neuroscience.

Primary neuroscience data is very expensive and difficult to share. Most remain inaccessible to researchers outside of the project [Mar2000a, Kos2000, Kos2002]. 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 [FL2002, GHWIK2001, HGKWA2001, GKAEW2001, GAKDE2001, Mar2000b], 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 [Mar2000b]. Recent advances in non-invasive neuroimaging technologies have compounded the problem of access as they have resulted in an explosion of new research findings. A huge number of diverse 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 [Kro2001]). 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 very difficult to combine processed data or results [RSLRB2001]. 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. The second obstacle is non-technical and includes 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, we observe that 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 metadata from different providers while not sacrificing data ownership and access rights.

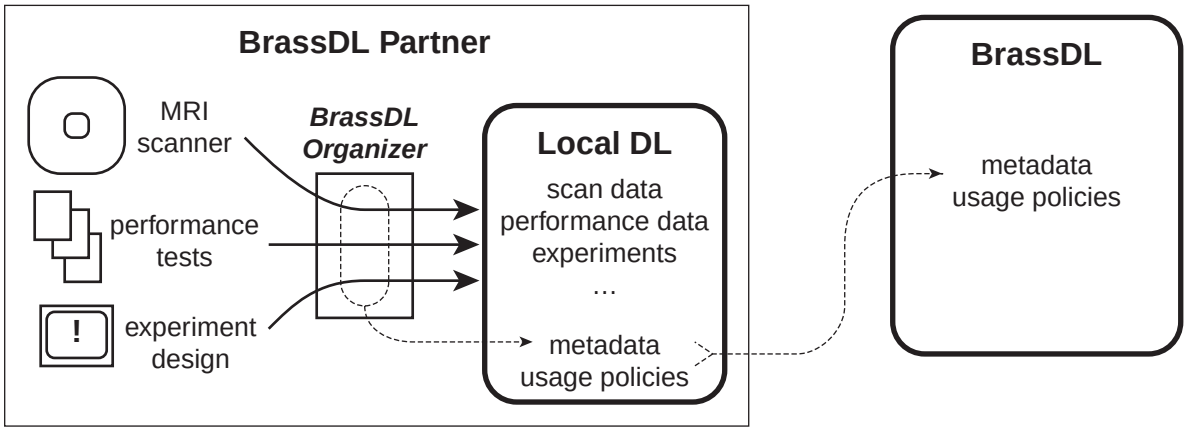

Figure 3: Contributing to BrassDL. A BrassDL partner site uses an organizer tool provided by BrassDL to manage its data streams (e.g. scans, performance data, experiment descriptions) in a local digital library. As data are added, metadata is created (e.g. scan kind and size, type of performance test, high-level description of experiment) and stored together with the actual data. To contribute to BrassDL, the metadata is submitted to the main BrassDL site, together with preliminary restrictions of use (e.g. "academic use only").

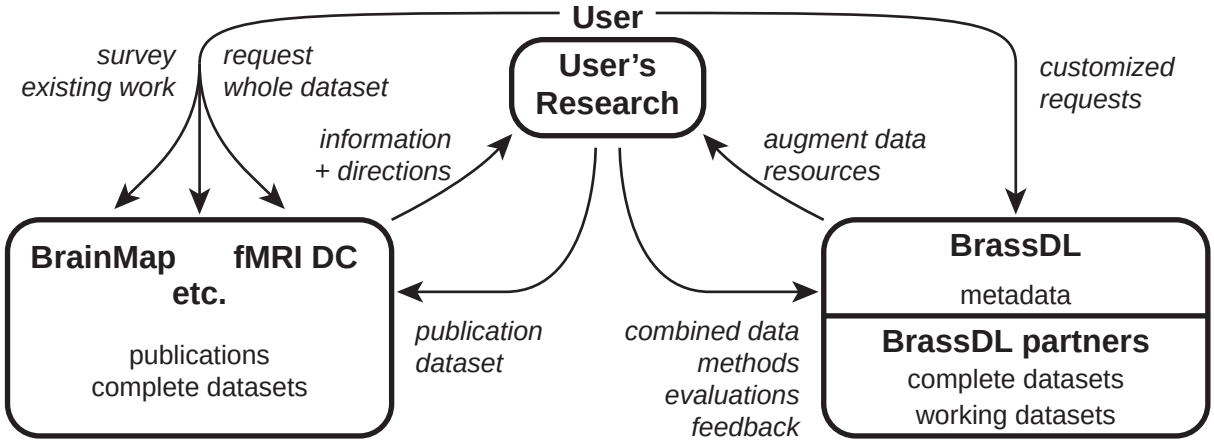

Figure 4: Information flow in brain imaging research. A user researching a given topic can currently look at BrainMap, the fMRI Data Center, and other online resources to get an overview of existing, published work. BrassDL is a synergistic superset to existing brain imaging resources as it (a) links 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 partners to help the user produce a publication which in turn is added to the pool of the traditional online resources. The metadata of the user's results are added to BrassDL to benefit future research.

IV.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 allows different types of data to be interoperable, but also provides a transaction and feedback system to facilitate multi-level data sharing and data evaluation. Unlike existing data sharing models [FL2002, GHWIK2001], BrassDL does not restrict itself

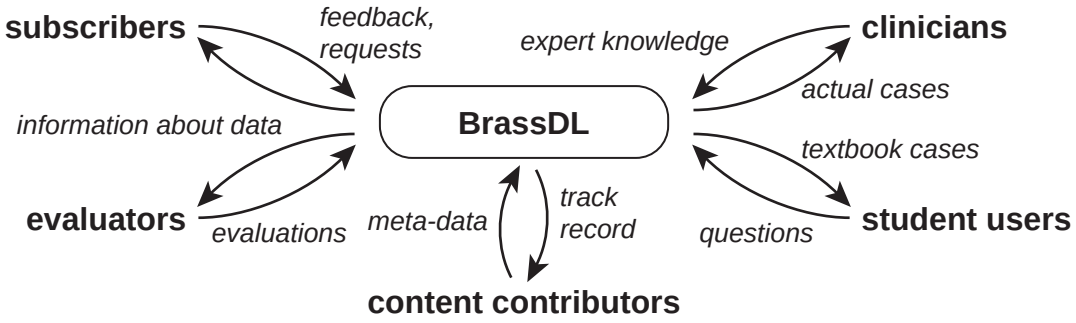

Figure 5: Usage of BrassDL by different users and partners. Each type of user benefits differently from BrassDL and contributes parts than are useful for other users. For example, a case contributed by a partner is evaluated as useful for a certain condition. A clinician then uses this case to treat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

to published-only data but starts bottom up, from the datasets that exist or are being developed in a laboratory to what becomes published. 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 transactions 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 contains 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 (including methods) 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 BrassDL partner sites by searching the metadata that BrassDL has collected from each partner; subsequently, to request data sets from a particular partner; and finally, to submit feedback about the data sets that were received from the partner as additions to the partner's track record.

The collection of metadata for BrassDL is shown in figure 3. Software at the BrassDL Partner's is used to organize the different data streams in a local digital library 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. Each BrassDL Partner can formulate policies to restrict usage to acceptable terms; specific agreements are made when a user requests a data set.

From the user's point of view, there are several options for accessing information about a given topic in brain imaging (see figure 4). Established sites offer either an overview of the field or complete sets of published data, and completed research can be submitted to these sites. On the other hand, BrassDL is more useful during the development and verification of a hypothesis by augmenting the available data resources. Furthermore, the user's results are incorporated individually, rather as a package.

BrassDL acts as a mediator between different types of users, each of whom contributes different services to the whole community (figure 5). As the foundation of the system, content contributors provide metadata (and keep the actual data) to the system. Basic subscribers use these metadata in their research and return feedback in form of results, or requests for specific data types. Evaluators classify the 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 the content contributors' track record.

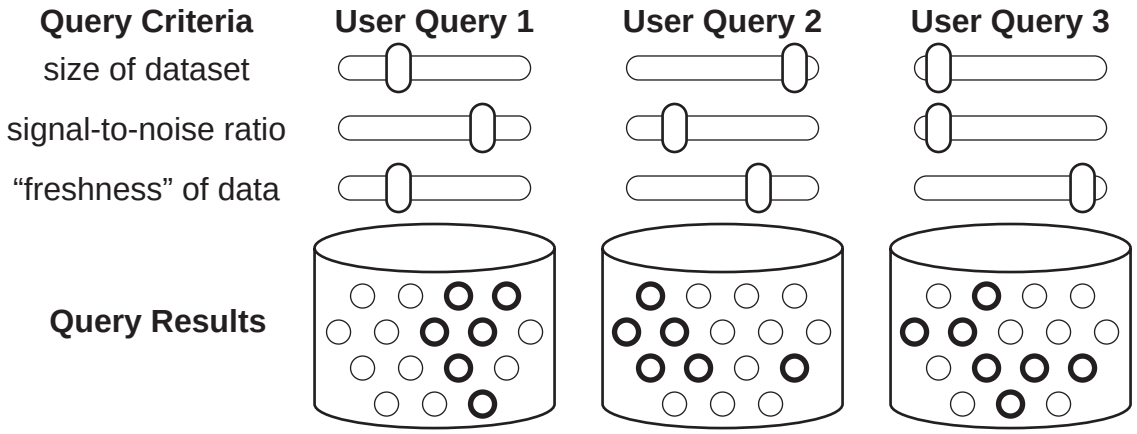

Figure 6: User-weighted queries. By giving search criteria different weights, the user can customize a query to match the priorities of a research topic. For example, query 1 prefers datasets with a very high signal-to-noise ratio; query 2 selects large new datasets, with size more important than age;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

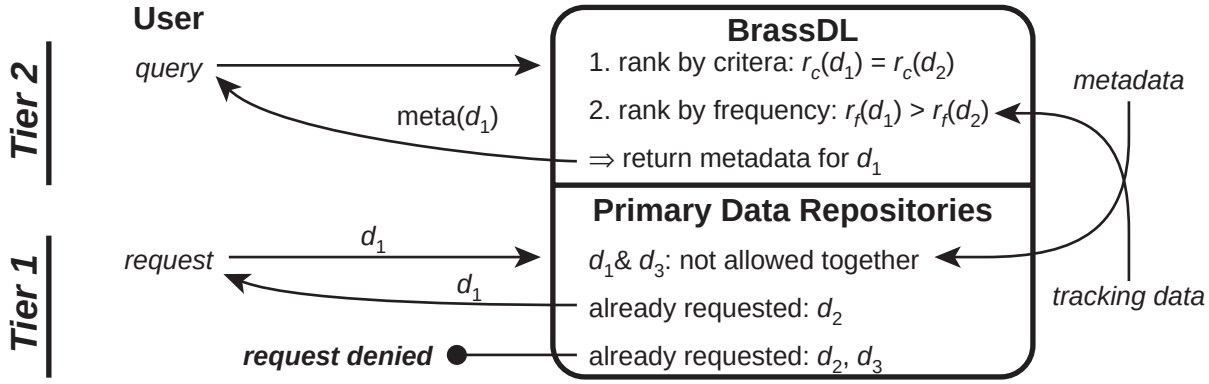

Figure 7: Evaluation on each tier influences user queries and requests. Results to user queries are primarily ranked by search criteria, but in case of ties, the ranking can include the frequency of use, determined by the tracking data from all previous accesses. Similarly, restrictions of use (e.g.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

IV.2. BrassDL Query and Data Evaluation

One strength of the abstraction layer that a MetaDL (and thus BrassDL) provides is the ability to specify queries with different weights on properties of the metadata being searched (see figure 6 for a simplistic example). The user can pick and choose which attributes to prioritize in a query in order to match the dataset already available to the user.

Metadata queries (Tier 2) are evaluated separately from transactions of actual datasets (Tier 1), yet influence each other (figure 7). The metadata of frequently requested datasets will rank higher than unused datasets given equal values for the search parameters. A transaction 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 recreate a face and identify a patient on a photo on file at an insurance company, leading to higher insurance rates) — based on what data the user already has available (according to BrassDL), the transaction agreement may include additional requirements.

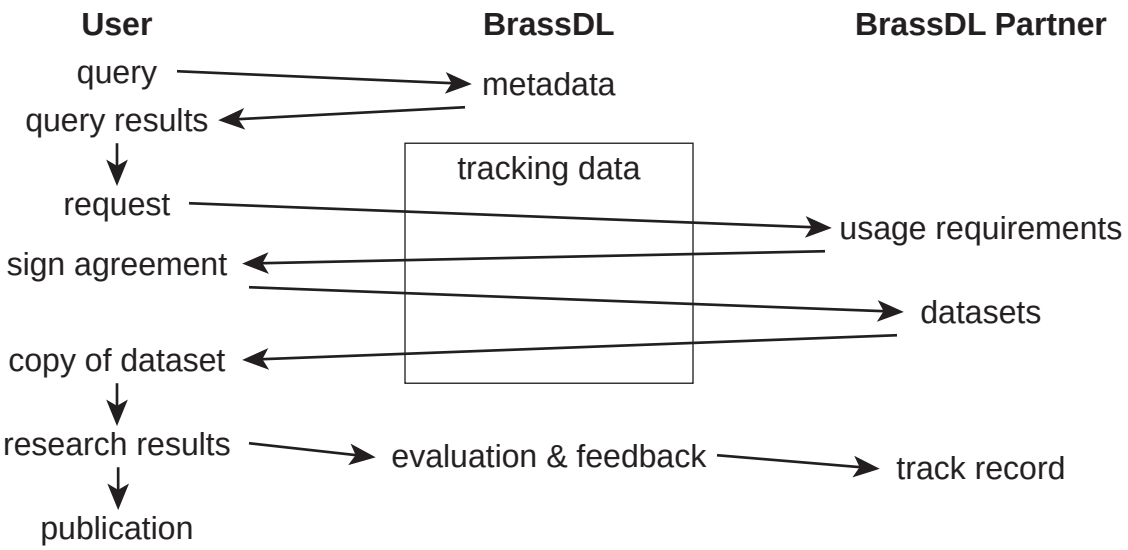

Figure 8: Protocol for data sharing and feedback. BrassDL acts as a mediator between user and data provider (BrassDL partner),initially to narrow the search for a good dataset, then to document the actual transaction, and finally to store comments about the dataset.

IV.3. Protocol for Primary Data Sharing

The following describes a transaction between a user, BrassDL, and a BrassDL partner supplying the dataset that the user would like to use (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 BrassDL partner. The partner replies with a set of usage requirements (e.g. nondisclosure, clearances, encryption). The user signs this binding agreement and returns it to the partner, who then releases the actual dataset to the user. This dialogue 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, and with the partner who can include them as part of a track record (incentive).

V. Discussion

BrassDL supports its 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 the user 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?" A more detailed study on this issue is presented in a separate paper [FMSS2002]. Here we briefly list several incentives for participation as follows.

- Visibility – The research activity of a data provider becomes visible to the community. This may help attract potential funding, patients or collaborators.
- Feedback — User feedback on the use of data may provide another way 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 and organize primary data in a standard fashion.

- Value assessment — Built-in evaluation mechanisms [FMSSS2002] assess the value of a dataset based on user demand and user feedback that result in automatic dataset ranking.
- Security — BrassDL provides 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 on "similar" work, as defined by her profile, on new datasets inserted; alternatively, she may receive consultation on how to proceed with future queries. BrassDL facilitates collaboration between organizations who generate primary data and organizations who perform advanced data analysis.
- Data sharing management — BrassDL stores and manages each transaction of data exchange. This helps owners manage their data-sharing activities, since it records cases or evidence of possible misuse and protects owner's rights.

VI. Conclusions

We have presented a new framework for digital libraries managing data sets that have limitations for distribution, and a specific implementation. 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 one-stop description for data, methods, experiments, tasks and subject data, (d) maintaining statistics and demographics of data and methods usage and providing a built-in evaluation standard to base transactions and provide user incentives (e) providing support for meta-analysis of results and studies of research demographics, (f) providing secure user support.

VII. References

[AMOZ2000] Andres, F., Mouaddib, N., Ono, K., Zhang, A. (2000). *Metadata Model, Resource Discovery, and Querying on Large Scale Multidimensional Datasets: The GEREQ Project.* Kyoto International Conference on Digital Libraries 2000: 83-90.

[BCGP1997] Baldonado, M. , Chang , C.-C. K., Gravano, L., Paepcke, A. (1997). *The Stanford Digital Library metadata architecture.* International Journal on Digital Libraries, 1(2):108-121.

[BI] BrainInfo. http://braininfo.rprc.washington.edu/

[BK2002] Barrett, J. S., Koprowski, S. P. Jr. (2002). *The epiphany of data warehousing technologies in the pharmaceutical industry*. Int J Clin Pharmacol Ther, 40(3):S3-13.

[CMMS2002] Crasto, C., Marenco, L., Miller, P., Shepherd, G. (2002). *Olfactory Receptor Database: a metadata-driven automated population from sources of gene and protein sequences.* Nucleic Acids Res., 30(1):354-60.

[CS1999] Carazo, J. M., Stelzer, E. H. K. (1999). *The BioImage Database Project: Organizing Multidimensional Biological Images in an Object-Relational Database.* Journal of Structural Biology, 125:97-102.

[Chi2000] Chicurel, M. (2000) *Databasing the brain.* Nature 406, 822-825 (24 Aug 2000)

[DCMI] DCMI. The Dublin Core Metadata Initiative. http://dublincore.org/

[DLIR] Digital Library Information Resources. http://sunsite.berkeley.edu/Info/

[DLMR] Digital Library Metadata Resources. http://www.ifla.org/II/metadata.htm

[ECHBD] ECHBD. European Computerised Human Brain Database. http://fornix.neuro.ki.se/ECHBD/Database/

[FFFHJ2000] Frew, J., Freeston, M., Freitas, N., Hill, L. , Janee, G., Lovette, K, Nideffer, R., Smith, T., Zheng, Q . (2000). *The Alexandria Digital Library architecture.* International Journal on Digital Libraries 2(4):259-268.

[FL2002] Fox, P. T., Lancaster, J. L. (2002). *Mapping context and content: the BrainMap model.* Nature Reviews Neuroscience, 3(4):319-321.

[FMSSS2002] Ford, J., Makedon, F., Shen, L., Steinberg, T., Saykin, A., Wishart, H. (2002). *Evaluation Metrics for User-Centered Ranking of Content in METADLs.* Fourth DELOS Workshop on Evaluation of digital libraries: Testbeds, measurements, and metrics, Budapest June 2002.

[GAKDE2001] Gardner D., Abato, M., Knuth, K. H., DeBellis, R., Erde, S. M. (2001). *Dynamic publication model for neurophysiology databases.* Philos Trans R Soc Lond B Biol Sci, 356(1412):1229-47.

[GB] NIH. GenBank. http://www.ncbi.nlm.nih.gov/Genbank/

[GHWIK2001] Grethe, J. S., Van Horn, J. D., Woodward, J. B., Inati, S., Kostelec, P. J., Aslam, J. A., Rockmore, D., Rus, D., Gazzaniga, M. S. (2001). *The fMRI data center: An introduction.* NeuroImage, 13(6):S135.

[GKAEW2001] Gardner, D., Knuth K. H., Abato, M., Erde, S. M., White, T., DeBellis., R, Gardner, E. P. (2001). *Common data model for neuroscience data and data model exchange.* J Am Med Inform Assoc, 8(1):103-4.

[Gel1992] Gelobter, M. (1992). *Public Data-archiving: a Fair Return on Publicly Funded Research.* Psycoloquy: 3(56) Data Archive (3).

[HBP] Human Brain Project. http://www.nimh.nih.gov/neuroinformatics/index.cfm

[HGKWA2001] Van Horn, J. D., Grethe J. S., Kostelec, P., Woodward, J. B., Aslam, J. A., Rus, D., Rockmore, D., Gazzaniga, M. S. (2001). *The Functional Magnetic Resonance Imaging Data Center (fMRIDC): the challenges and rewards of large-scale databasing of neuroimaging studies.* Philos Trans R Soc Lond B Biol Sci, 356(1412):1323-39.

[HGP] Human Genome Project. http://www.ornl.gov/hgmis/

[HL2001] Houstis, C., Lalis, S. (2001). *ARION: An Advanced Lightweight Software System Architecture for accessing Scientific Collections.* Cultivate Interactive, issue 4, May 2001

[HSSNS1997] Healey, M. D., Smith, J. E., Singer, M. S., Nadkarni, P. M., Skoufos, E., Miller, P. L., Shepherd, G. M. (1997). *Olfactory receptor database (ORDB): a resource for sharing and analyzing published and unpublished data.* Chem. Senses, 22:321-326.

[ICBM] ICBM. International Consortium for Brain Mapping. http://www.loni.ucla.edu/ICBM/

[Kos2000] Koslow, S. H. (2000). *Should the neuroscience community make a paradigm shift to sharing primary data?* Nature Neuroscience 3, 863-865 (01 Sep 2000)

[Kos2002] Koslow, S. H. (2002). *OPINION: Sharing primary data: a threat or asset to discovery?* Nature Reviews Neuroscience 3, 311-313 (01 Apr 2002)

[Kro2001] Krotz, D. (2001). *PET and MRI race to detect early Alzheimer's.* 01/22/2001 http://www.auntminnie.com/articles/50098.asp

[MFSS2002] Makedon, F., Ford, J., Shen, L., Steinberg, T. Sustainability Models for MetaDL's. Technical Report, in preparation, 2002.

[METAE] METAe. The Metadata Engine Project. http://meta-e.uibk.ac.at/

[MF2001] Marcos, R. K. F., Fox, E. A. (2001). *MARIAN: Flexible Interoperability for Federated Digital Libraries.* Lecture Notes in Computer ScienceVolume, 2163:0173.

[MW] MetaWeb. http://www.dstc.edu.au/Research/Projects/metaweb/

[Mar2000a] Marshall, E. (2000). *A Ruckus Over Releasing Images of the Human Brain.* Science 2000 September 1; 289: 1458-1459.

[Mar2000b] Marshall, E. (2000). *Downloading the human brain with security*. Science 289, 2250.

[Mar2002] Marshall, E. (2002). *DNA Sequencer Protests Being Scooped With His Own Data.* Science, 295(5558):1206-1207.

[Mil2000] Miller, P. (2000). *Collected Wisdom.* D-Lib Magazine special issue on Collection Level Description, 6(9), Sep. 2000.

[OHBM] OHBM. The Organization for Human Brain Mapping. http://www.humanbrainmapping.org/

[RDPM2001] Roberts, L., Davenport, R. J., Pennisi, E., Marshall, E. (2001). *A history of the Human Genome Project.* Science, 291(5507):1195.

[RSLRB2001] Roland, P., Svensson, G., Lindeberg, T., Risch, T., Baumann, P., Dehmel, A., Frederiksson, J., Halldorson, H., Forsberg, L., Young, J., Zilles, K. (2001). *A database generator for human brain imaging.* Trends in Neuroscience, 24(10):562-564.

[SHSNM1999] Skoufos, E., Healey, M. D., Singer, M. S., Nadkarni, P. M., Miller, P. L., Shepherd, G. M. (1999). *Olfactory Receptor Database: a database of the largest eukaryotic gene family.* Nucleic Acids Res., 27:343-345.

[SMNMS2000] Skoufos, E., Marenco, L., Nadkarni, P. M., Miller, P., Shepherd, G. M. (2000). *Olfactory Receptor Database: a sensory chemoreceptor resource.* Nucleic Acids Res., 28, 341-343.

[ST2000] Sweet, M., Thomas, D. (2000). *Archives Described at Collection Level.* D-Lib Magazine special issue on Collection Level Description, 6(9), Sep. 2000.

[Sko1992] Skoyles, J. R. (1992). *FTP Internet Data Archiving: a Cousin for Psycoloquy.* Psycoloquy: 3(29) Data Archive (1).

[Wag2000] Wagner, D. A. (2000). *Early detection of Alzheimer's disease: An fMRI marker for people at risk?* Nature Neuroscience, 3(10): 973-974.