

Multi-Functional Data Collection Interfaces for Biomedical Research Collaboration

*Fillia Makedon¹, Tilmann Steinberg¹, Laurence G. Rahme²,
Aria Tzika³, Heather Wishart⁴, Yuhang Wang¹*

Abstract

This paper describes data collection interfaces for research collaboration in biomedical applications where there is need for secure sharing of sensitive data. These interfaces are multi-functional because they (a) are structured templates for entering experiments and tools within a given domain; (b) provide hierarchical entry/presentation of data; (c) offer security via control of access by user type for each level of hierarchy; (d) provide tracking of data usage; (e) combine workflow management of tasks, thus enabling collaboration; (f) facilitate interoperability for heterogeneous data via common metadata representations; (g) enable advanced searching capabilities allowing multiple parameters; (h) offer assessment services evaluating the quality of entered data; and (i) support intelligent data management and services (*e.g.* notification and risk analysis).

1 Introduction

A fundamental challenge in biomedical research is providing access to data collections to facilitate early detection or discovery. One major technical barrier is the non-interoperability of data due to different methods (formats) of collection or representation (Roland, Svensson *et al.* 2001). This paper describes the Catalog system, an interface built to be open, flexible, user-centered and modular. Catalog is self-standing but can also serve as the front end of any collaboration system where different researchers must collect and share data within a given domain. The system is currently being developed at the Dartmouth Experimental Visualization Laboratory (DEVLAB) for different target applications in neuroscience, molecular biology and heart dynamics research. It collects data for use in *Negotiation Based Sharing (NBS)* (Ye, Makedon *et al.* 2003), and is extensible and amenable to intra- and inter-domain data sharing.

NBS is based on two simple connected principles: (a) the use of extracted metadata to represent primary data and publicize ongoing research and (b) the use of a negotiation mechanism (SCENS) to provide incentives for users to share their data. The Catalog system collects metadata rather than primary data. The Catalog system must provide ease of use, security, scalability and sustainability (Makedon, Ford *et al.* 2002). Since maintaining the rights of information owners and providing incentives for participation are key to the SCENS framework (**Figure 1**), the data owners (A, B, C) submit metadata of their datasets and policies for their use via the Catalog system. Catalog enables the owner to define the conditions of sharing: when, how, by whom, and for how long. A subscriber-user (A) can query the metadata for qualifying sets, and then enter negotiation mode for access to the original data. Once the data owners' requirements have been met, SCENS

¹ Dartmouth College, 6211 Sudikoff Laboratory, Hanover, NH 03755, makedon@cs.dartmouth.edu

² Department of Surgery, 50 Blossom Street, Massachusetts General Hospital, Boston, MA 02114

³ NMR Surgical Laboratory, 51 Blossom Street, Massachusetts General Hospital, Boston, MA 02114

⁴ Dartmouth Medical School, Dartmouth-Hitchcock Medical Center, Lebanon, NH 03756

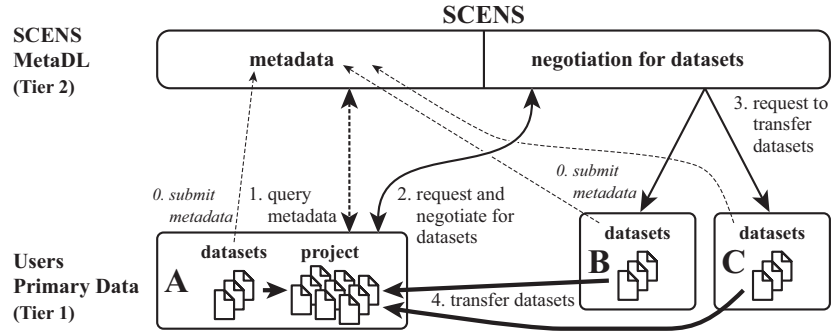


Figure 1. The SCENS framework. *Dashed lines: metadata transfer; thin solid lines: negotiation traffic; thick solid lines: actual data transfer. The **Catalog System** resides in Tier 2, negotiation.*

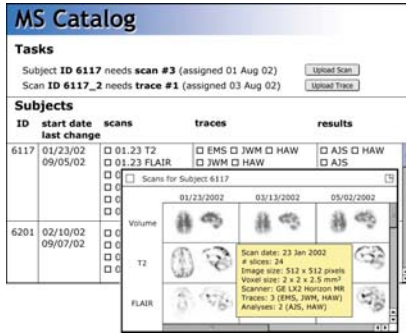


Figure 2. Catalog Interface for MS

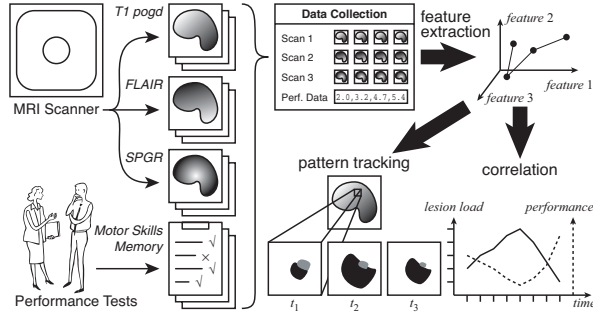


Figure 3. Data acquisition, integration, & analysis

sends a request to the primary data owners, B and C, to transfer the datasets A has requested.

This paper reports on the development of the Catalog system for spatiotemporal multi-modal data streams (**Figure 2** shows an example for MS). The Catalog system helps to "fuse" these data streams and arrive at high-level patterns for early diagnosis. The initial goal is to create local metadata that are later uploaded to a central SCENS server for sharing with other Catalog users.

2 Related Work

Metadata-based collections are used, among others, in the **BIRN** project (Marx 2002), **GenBank** (National Institutes of Health), the **European Computerized Human Brain Database (ECHBD)** (ECHBD), the **fMRI Data Center** (Grethe, Van Horn et al. 2001), and the **Open Archives Initiative** (Open Archives Initiative). Similar ideas also exist in the **Common Data Model** (Gardner, Knuth et al. 2001); **BrainMap** (Fox and Lancaster 2002); and the **BioImage Database Project** (Carazo and Stelzer 1999). In all these, metadata are used to link to data that have not been themselves integrated into the system and do not contain Catalog facilities (Makedon, Ford et al. 2002) but are centralized indexes of distributed data sources. Two cataloging tools are **Axiopie** (Axiopie project), which combines independent local metadata into a federated system; and **NeuroSys** (Pittendrigh and Jacobs), which provides metadata based on domain ontologies.

The Catalog system is suited for federated systems where there is need to integrate different sche-

mas from different domains and data models using a common metadata standard. Our system builds on existing metadata standards (OAI, **Dublin Core Metadata Initiative** (DCMI)). Other influential metadata projects include the **METAe** project (METAe) and **BrainML** (Weill Medical College of Cornell University Laboratory of Neuroinformatics).

3 System Features

3.1 Multi-Functionality of Data Collection

The first challenge is to enhance the data collection process so that it combines **collection, integration, and analysis** effectively (see **Figure 3**), by integrating multi-modal data, workflow management, tracking data origins, and analysis tools (*e.g.* consistency checking) at multiple time points. This requires the development of an effective common representation of imaging, clinical, and treatment data. To date, no generally available system exists for the simultaneous consideration of multiple imaging modalities and none that is systematic, scalable and automated.

A second challenge is to automate metadata extraction for multidimensional data. This facilitates stream fusion and pattern detection, which enables researchers to identify (and be notified of) key events in the course of the disease progress (*e.g.*, when the ratio of lesion volume exceeds a certain percentage). This process is based on expert knowledge and patterns learned from training data.

A third challenge is how to integrate data sharing in the data collection process (Ye, Makedon et al. 2002). To address a seamless integration at an early stage in the data collection stage, the Catalog system must (a) prompt for and encode particular data sharing or usage conditions set by data owners, and conditions imposed by laws and institutional policies; and (b) offer multi-level access and security features. Presenting high-level information of the usage of the local data by others can aid in policy making.

Finally, a fourth challenge is to create a *domain-independent* system. By comparing system development in different target applications, one can determine common needs, critical information, and how to develop evaluation criteria, *e.g.* user satisfaction, that promote system use.

3.2 Data Collection, Analysis and Metadata Extraction in Biomedicine

The applications described below have different types of data, complexity, and clinical goals. Yet there are similarities: they both require effective summarization of changes with metadata; in both, early detection of critical events requires utilization of prior knowledge and analysis that extracts features of substance from the original data; and correlation with patient performance before or after (drug) treatment over time assumes objective quantification of these data as well. **Figure 4** gives an example of the type of data collection, integration, and analysis facilities needed: the original data and some “private” metadata are for research only and must be secured from outside access. Metadata extraction research involves (a) finding a minimal set of expressive features for further analysis, and (b) codifying security and usage policies to safeguard (manage) the public metadata and ensuring that it is followed.

We are studying and developing new algorithms and techniques for metadata extraction, including multiresolution description of the shape of various anatomical structures. An important benefit of these multiresolution deformable models is that they support both global deformation parameters, which efficiently represent the gross shape features of an object, and local parameters, which cap-

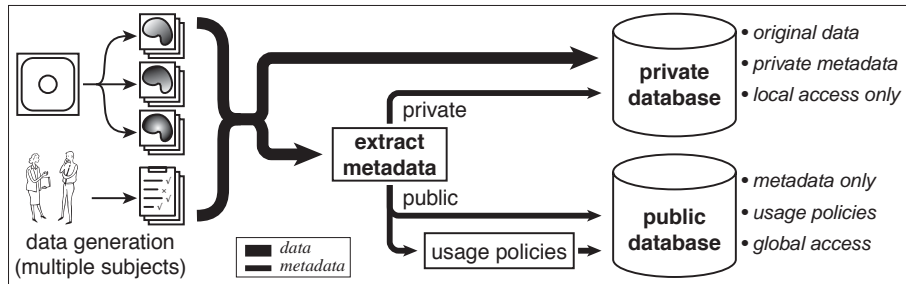


Figure 4. Data acquisition and dissemination.

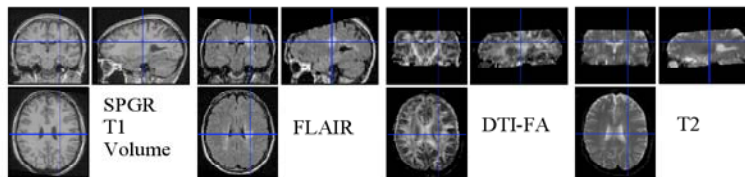


Figure 5. MR image sets of same brain locations taken using different modalities.

ture shape details. Identifying the dimensionality of the wavelet feature vector that best describes each structure, based on statistical shape variability, allows its use as an efficient and compact metadata descriptor.

3.3 Examples in Biomedical Research

Spatiotemporal Multimodal Streams (MS Lesion Analysis): Understanding MS activity requires monitoring neurobiological changes with magnetic resonance imaging (MRI) (Lee, Smith et al. 1999). Most MS-related research and clinical practice has depended on conventional MRI imaging techniques (T2, FLAIR; see **Figure 5**) to detect the lesions that are characteristic of the disease, but there is increasing MRI-based evidence (Filippi and Grossman 2002) of additional more diffuse pathology in normal-appearing brain tissue (NABT) that likely also contributes to symptoms. These diffuse changes in NABT are visible only on advanced types of MR scans, such as diffusion tensor imaging (DTI). However, no one scan type can image all the specific pathological processes or abnormalities. We are developing a data-collation paradigm useful for all modalities, and for any new imaging technology, that will make it easier to integrate data and examine the full extent of pathology in the central nervous system, allowing for more accurate monitoring of disease progression and more informed treatment decisions (Mainero, De Stefano et al. 2001).

Multiparametric Data Tracking (Brain Tumor Tracking): Multiparametric MR imaging using MRSI, HMRI and DWMRI can provide an enhanced assessment of neuroepithelial brain tumors, potentially allowing the distinction between higher-grade and lower-grade tumors, which becomes clinically important especially when the tumors are inoperable. An initial database infrastructure for better data collection and management allows for more power and flexibility in detailed data analysis. Using metadata and secure negotiation services (SCENS) will make it possible to carry out clinical research on a larger scale and share work.

4 Ongoing Work

The described system is currently under development at the DEVLAB. We are using Java for de-

ploying the Catalog Tool (**Figure 2**) for compatibility with most platforms, and the JDBC API to integrate this tool with any database solution (*e.g.* Oracle, Informix) that supports the Structured Query Language (SQL). The following locations will apply the system to their respective areas of research. Tumor data: Harvard-Mass General Hospital (Tzika & Astrakas) and Dartmouth-Advanced Imaging Center (Pearlman). Multiple sclerosis lesion data: Dartmouth Hospital, Brain Imaging Lab (Saykin and Wishart) and Harvard, Brigham and Women's Hospital (Guttmann). Heart data: Dartmouth-Advanced Imaging Center (Pearlman).

References

- Axiop project Data Sharing White Paper, <http://www.axiop.org/datasharingwhite.html>.
- Carazo, J. M. and E. H. K. Stelzer (1999). "The BioImage Database Project: Organizing Multidimensional Biological Images in an Object-Relational Database." *Journal of Structural Biology* **125**: 97–102.
- DCMI The Dublin Core Metadata Initiative, <http://dublincore.org/>.
- ECHBD European Computerised Human Brain Database, <http://fornix.neuro.ki.se/ECHBD/Database/>.
- Filippi, M. and R. I. Grossman (2002). "MRI techniques to monitor MS evolution." *Neurology* **58**: 1147–1153.
- Fox, P. T. and J. L. Lancaster (2002). "Mapping context and content: the BrainMap model." *Nature Reviews Neuroscience* **3**(4): 319–321.
- Gardner, D., K. H. Knuth, et al. (2001). "Common data model for neuroscience data and data model exchange." *Journal of the American Medical Informatics Association* **8**(1): 103–104.
- Grethe, J. S., J. D. Van Horn, et al. (2001). "The fMRI data center: An introduction." *NeuroImage* **13**(6): S135.
- Lee, M. A., S. Smith, et al. (1999). "Spatial mapping of T2 and gadolinium-enhancing T1 lesion volumes in multiple sclerosis: evidence for distinct mechanisms of lesion genesis? [see comments]." *Brain* **122**(Pt 7): 1261–70.
- Mainero, C., N. De Stefano, et al. (2001). "Correlates of MS disability assessed in vivo using aggregates of MR quantities." *Neurology*. **56**(10): 1331–4.
- Makedon, F., J. C. Ford, et al. (2002). MetaDL: A Digital Library of Metadata for Sensitive or Complex Research Data. European Conference on Digital Libraries (ECDL2002), Rome, Italy.
- Marx, V. (2002). "Beautiful Bioimages for the Eyes of Many Beholders." *Science* **297**(5578): 39–40.
- METAe The Metadata Engine Project, <http://meta-e.uibk.ac.at/>.
- National Institutes of Health GenBank, <http://www.ncbi.nlm.nih.gov/Genbank/>.
- Open Archives Initiative OAI Home Page, <http://www.openarchives.org/>.
- Pittendrigh, S. and G. Jacobs NeuroSys: An Electronic Laboratory Notebook and Semi-structured Database, <http://www.nervana.montana.edu/~sandy/paper.doc>.
- Roland, P., G. Svensson, et al. (2001). "A database generator for human brain imaging." *Trends in Neuroscience* **24**(10): 562–564.
- Weill Medical College of Cornell University Laboratory of Neuroinformatics BrainML functional ontology for neuroscience, <http://brainml.org/>.
- Ye, S., F. Makedon, et al. (2002). A Negotiation Framework for Secure Data Sharing. Hanover, NH, Dartmouth College Computer Science Department.
- Ye, S., F. Makedon, et al. (2003). SCENS: A system for the mediated sharing of sensitive information. In submission to the Third ACM/IEEE Joint Conference on Digital Libraries.