Institut Supérieur de l'Electronique
et du Numérique
Tél. : +33 (0)2.98.03.84.00
Fax : +33 (0)2.98.03.84.10
CS 42807 - 29228 BREST Cedex 2 - FRANCE

N5

School Year - 2008/2009
Training Period

NIMS: Neurobiological Image Management System

Improving the storage, sharing, mining and analysis of neuroimaging data using specific data management system

---

From April 6th to August 31st

at
Vision, Imaging Sciences and Technology Activities (VISTA) Laboratory
Department of Psychology Stanford University

Jordan Hall - 450 Serra Mall - Stanford University
CA 94305 Stanford - USA
Phone: +01 650-725-2466

Supervisors: Brian WANDELL, Vista Laboratory - Stanford University, wandell@stanford.edu
Jorge PHILLIPS, ConjureLLC, jp@alum.mit.edu

Referring Tearchers: Eric STINDEL - Master SIBM
Dominique MARATRAY - ISEN Brest

by Vincent SIMOES

Acknowledgment

First of all, I would like to thank Brian WANDELL who gave me the opportunity to realize this internship at the Stanford University and who has always been here to assist me in my work. I would also like to thank Pamela WIDRIN for her help on the administrative side, especially to obtain the visa.

Of course, my thoughts also go to my team of work and specially Jorge PHILLIPS, Robert F. DOUGHERTY and Anthony SHERBONDY for allowing me to work on such an interesting project, for their assistance when I needed, their time to answer to my questions and their understanding about my bad control of English. I learned a lot with them on very different topics. Indeed, I have worked on the creation of a complete MRI images management system, which involved two main fields: the computer science and the medical imaging science. Moreover, working in a such prestigious place is a good way to discover the research's world and to have the possibility to assist to a lot of talks given by researchers from across the world.

Then, I would like to thank Eric STINDEL to have accepted to supervise this internship and Dominique MARATRAY for her assistance.

Finally, I would like to thank all the people working in the VISTA lab for their reception and their friendliness. Thanks to them, my time in the laboratory has been a real pleasure and they made me feel good by including me in the team very easily.

Contents

List of Figures

Abstract

The use of systems to manage large amounts of data becomes increasingly important in the field of neuroscience as researchers work on neuroimaging methods such as positron emission tomography (TEP) and functional magnetic resonance imaging (fMRI). Indeed, these techniques have produced an explosion of new findings in human neuroscience associated with an explosion of the amount and the size of the data involved in the researches. Moreover, the necessity to share both data and results of processing with others researchers and laboratories is very important nowadays in order to create efficient collaborations and thus to help to make new findings.

In this context, this document propose an interesting solution through the creation of the Neurobiological Image Management System (NIMS) allowing the management of neurological data. Thus, this project consists on an experimental software platform designed to ease and to improve the sharing, storage, mining and analysis of data coming from scanners in different groups of research. This has been made in a way that can evolve depending on the needs of each thanks to an open-source and flexible approach. The document presents similar projects existing in other research centers, a first version with a standard neurobiological imaging workflow and gives some ideas about the creation of an ontology for a such system.

Keywords: data sharing, database, neurobiological images, open-source, metadata, workflow, ontology

Résumé

L'utilisation de systèmes pour prendre en charge d'importantes quantités de données devient de plus en plus fréquente dans le domaine des neurosciences alors que les chercheurs travaillent sur des méthodes de neuroimagerie telles que la tomographie par émission de positions (TEP) et l'imagerie par résonance magnétique fonctionnelle (IRMf). En effet, ces techniques ont produit une explosion de nouvelles découvertes en neuroscience associée à une explosion de la quantité et de la taille des données impliquées dans les recherches. De plus, la nécessité de partager à la fois les données et les résultats des traitements avec les autres (chercheurs et laboratoires) est très importante de nos jours afin de créer des collaborations efficaces et ainsi d'aider à faire de nouvelles découvertes.

Dans ce contexte, ce document propose une solution intéressante à travers la création d'un système de management d'images neurobiologiques. Ce projet consiste en une plateforme logicielle expérimentale construite pour faciliter et améliorer le partage, le stockage, l'extraction et l'analyse de données provenant de scanners dans différents groupes de recherche. Ceci a été fait d'une manière qui peut évoluer en fonction des besoins de chacun grâce à une approche open-source et flexible. Le document présente des projets similaires dans d'autres centres de recherches, ainsi qu'une première version de celui-ci avec un flux de travail (workflow) standardisé et donne quelques idées concernant la création d'une ontologie pour un tel système.

Mots clés : partage de données, base de données, images neurobiologiques, open-source, metadonnées, flux de travail, ontologie

Introduction

My training period took place in the Vision, Imaging Sciences and Technology Activities (VISTA) laboratory, one of the research centers of the Stanford University in Palo Alto. During approximately 5 months, from April 6th to August 31th, I worked in collaboration with an outside company, ConjureLLC represented by Jorge PHILLIPS and Anthony SHERBONDY. Brian WANDELL and Bob DOUGHERTY, part of the laboratory were the supervisors of the project by checking the evolution of the work.

Nowadays, the increasingly complex research questions addressed by neuroimaging research impose substantial demands on computational infrastructures. These infrastructures need to support management of massive amounts of data in a way that affords rapid and precise data analysis, to allow collaborative research, and to achieve these aims securely and with minimum management overhead. This is particularly true in the field of functional magnetic imaging (fMRI), which has produced an explosion of the amount and the size of data involved. For example, a single research study may require the repeated processing, using computationally demanding and complex applications, of thousand of files corresponding to hundreds of functional MRI studies.

Thus, the main goal of the project is to create an open-source and flexible management prototype system for neurobiological images in order to improve their sharing, storage, mining and analysis between several researchers groups. This is designed to improve researcher and team productivity by supporting neurobiological imaging workflow through management of images and data classification and storage. This prototype should be sufficiently modular and extensible to be able to accommodate near future changes in functionality, implementation technologies and deployment modalities, and scalable to enterprise quality in future projects.

In this report, the system shown is a result of the work made on the project during the 5 months I was in the laboratory and is due to be used in a first step by several laboratories of the psychology department of Stanford University. First of all, after a quick presentation of Stanford, of the Vista laboratory and of the overall project, the first part will be concentrated on the similar projects existing in other research centers as the Computational Neuroscience Applications Research Infrastructure (CNARI) or the Extensible Neuroimaging Archive Toolkit (XNAT). Afterward, a second part will explain the technological choices made for our system, the standard workflow used for the data coming from a scanner and the structure of the overall system. In a third part, I will expose some ideas extracted from the literature to create an ontology for our system. Finally, I will present the next steps that the team will achieve in the future to answer to the initial requirements of the project.

Chapter 1

Presentation

1.1 Stanford University, Palo Alto

Figure 1.1: Stanford University, Palo Alto

Leland Stanford Junior University (commonly referred to as Stanford University) is a private research university located in the San Francisco Peninsula, more exactly in Palo Alto, California, United States. Stanford was founded in 1885 by former California governor and Senator Leland Stanford as a memorial to his son Leland Stanford Jr. The Stanfords used their farm lands to establish the university hoping to create a large institution in California. The campus is the largest campus in the world, in terms of contiguous acreage with 33.1 km^2.

Today, Stanford enrolls about 6,500 undergraduate and about 10,000 graduate students from the United States and around the world every year.

The university is divided into seven schools: School of Humanities and Sciences, School of Engineering, School of Earth Sciences, School of Education, Graduate School of Business, Stanford Law School and the Stanford University School of Medicine.

In the world, Stanford is one of the most highly regarded schools. Thus, in 2008, it was ranked 2th in the world by the annual list, Top 500 World Universities, published by the Institute of Higher Education at Shanghai Jiao Tong University, China.

The University has a significant impact in the world of research with a lot of research centers and with one of the most important research spending in the world. Moreover, current community of scholars includes a large number of high awarded researchers among them, 18 Nobel Prize and 135 members of the National Academy of Sciences. As part of the Silicon Valley, its alumni have founded companies like Hewlett-Packard, Sun Microsystems, Nvdia, Yahoo!, Cisco Systems and Google.

1.2 Department of Psychology

The department is organized into five different areas of study within the field of psychology. They are: Cognitive, Developmental, Neuroscience, Personality and Social Psychology.

Currently there are many different kinds of research studies being conducted in this department, on such topics as aggression, social behavior, competitiveness, dreaming, color perception, spatial relations, learning and memory. Researchers rely on participants to keep this research going.

Neuroscience investigates the human brain, from the functional organization of large scale cerebral systems to microscopic neurochemical processes. Topics include the neural substrates of perception, attention, memory, language, learning, neurological disorders, affect, stress and motivation. A variety of experimental techniques are used, including functional magnetic resonance imaging (fMRI), electro/magneto-encephalogry (EEG/MEG), and transcranial magnetic stimulation (TMS).

1.3 VISTA laboratory

Stanfords Vision, Imaging Sciences and Technology Activities (VISTA) laboratory is a major research center focused on the use of neuroimaging methods (fMRI, DTI) and behavior to study the human visual system. It is part of the department of Psychology of Stanford university.

The centers current research portfolio encompasses projects across several technologies and applications areas, examples of which include:

- Color vision: the team uses both neuroimaging and behavior testing to understand the action of the visual portions of the brain. They have developed a set of methods for identifying and measuring distinct and specialized regions of human visual cortex, including regions that respond powerful to motion and color. Image systems: this work is to support multidisciplinary training, research and collaboration on technologies leading to novel imaging systems that include the capture, processing, transmission and rendering of visual information.

- Reading development: the laboratory is applying a powerful set of measurement methodologies to study human brain development. In one group of studies, they are measuring the signals and growth of visual cortex in children, aged 8-12, during the period children become skilled readers. Using very high spatial resolution and neuroimaging techniques, including some methods developed by this group, they are hoping to understand how visual signals contribute to the neural pathways of reading.

For their researches, they have developed several tools regrouped in an overall software tool called mrVista (Mister Vista). This software includes methods for processing anatomical, functional and diffusion tensor imaging data. It is written mainly in Matlab. Other softwares have been developed by the team. Among them, ITKGray is a tool for segmentation and CINCH is a tool for visualization and segmentation of tractography results into meaningful fiber bundles.

More information about the researches conducted by the team and their softwares is given on the website: http://white.stanford.edu.

1.4 Presentation of the overall project

1.4.1 Problematic

As explain before, the main goal of the project is to answer to the problems of data management met by a lot of research centers around the world, especially in the field of neurosciences, where the amount of data is huge. These problems are similar in other fields of research such as genomics with for example, the Human Genome Project database and the Protein Data Bank or in astronomy where a lot of work are made around the management of data. For example, the Sloan Digital Sky Survey is making an increasing use of DBMS technology to describe millions of celestial objects, and to enable searches across that data (Nieto-Santisteban et al, 2005).

In the field of neuroscience, these problems have already been studied by several groups of researchers and some of them have produced interesting works as the CNARI project: The Computational Neuroscience Applications Research Infrastructure, fMRIDC: The Functional Magnetic Resonance Imaging Data Center or XNAT: The Extensible Neuroimaging Archive Toolkit. Details about these projects and some others will be given further in this document as an inventory on the domain.

1.4.2 Objectives and hindrances

Nevertheless, the need and the possibility to improve considerably the storage, sharing, mining and analysis of the neuroimaging data still exists and significant improvements will come with the multiplication of system of data management. In this purpose, we would like to create a new data system management called: NIMS: Neurobiological Image Management System, which will be used and shared in a first step by all the laboratories of Stanford University working on the neuroscience, most of them part of the Psychology Department. As last step, the system might be extended to other laboratories. Indeed, we use an open-source approach thanks to a common language (Python) to allow an easy deployment and an easy arrangement in contradiction with current project as CNARI, which use their own language (Swift).

Thus, with this work, the Vista laboratory would like to improve considerably the researcher and team productivity by supporting a new neurobiological imaging workflow and by allowing an expansion in the types and the complexity of questions we are able to ask to the system. Another goal is to allow a collaborative research through several research centers. Indeed, the members of the laboratory believe that sharing information (images, results of analysis) between laboratories will permit to facilitate and increase the number of new findings in neuroscience. The figure 1.2 shows an overall view of NIMS. As explain, data from scanners, processed images and related measures should be able to be loaded in NIMS. Local user and collaborators should be able to retrieve these data.

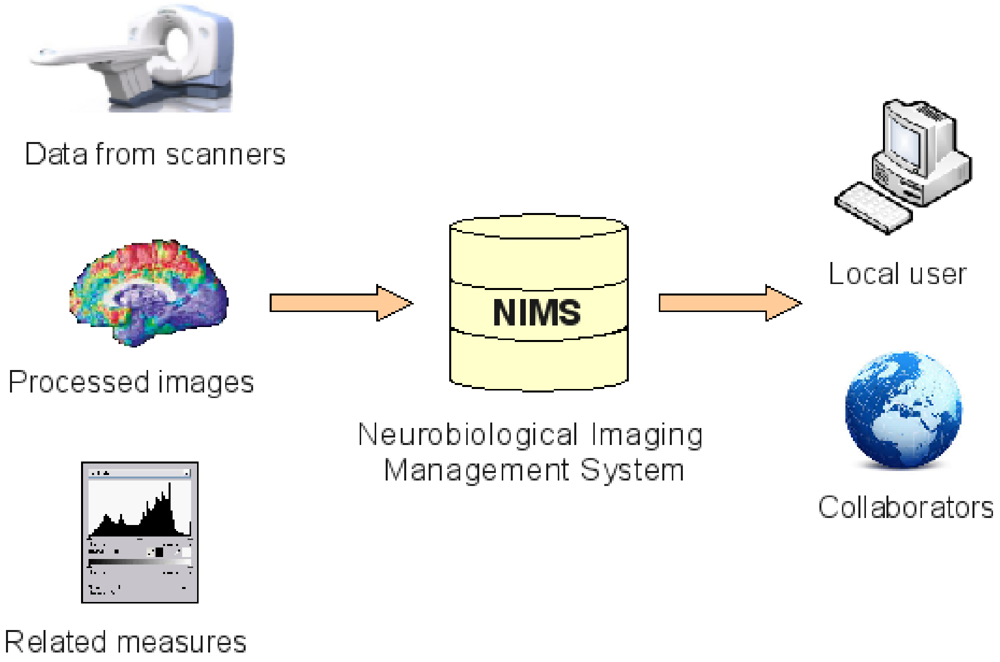


Figure 1.2: NIMS: overall view

However, with this project, we face several problems. First of all, the backup of images and research results are necessary for system like ours. Then, a work asking for a collaboration between research centers is often difficult to set in place because of the differences of point of view and of course the differences in the process, especially here in the image workflow (data format and conversion, image nomenclature, etc). Another thing is the problems of security and access control which are essential to protect the patients involved in the research and to respect the privacy and the work of each laboratory.

1.4.3 Initial road map

The initial road map consists in three main parts. First, we have to establish the requirements from major stakeholders by understanding the neuroimaging workflow in each laboratory (image format used, steps followed to store images, etc), by understanding pain points in each workflow and by capturing productivity improvement opportunities to create a new optimized workflow. The second part consists on the definition and the design of the NIMS architecture. The initial functionality of the system are an iterative refinement of raw scanner data into standardized images representations and the implementation of basic workflow analysis. At this step, we will have to create a new ontology of lab images and research results to allow to put user-defined attributes on images and thus, answer to the needs of each laboratory and increase the complexity of questions we will be able to ask. Then, the team will think for additional feature sets to add to the system such as the possibility to index images and research results by region. Finally, the NIMS system will be implemented in real condition for testing purposes. Of course, after this work, the team will be attached to allow the system to be shared with external research centers working on neuroscience especially in the field of fMRI.

Chapter 2

Background and similar projects

2.1 General background

Electronic data sharing has become an important tool in many scientific disciplines. This is especially true for those that work with large and complex data sets as. Thus, the value of data sharing has become increasingly apparent to scientists involved in neuroimaging for several reasons. The volume of data generated with brain imaging techniques is striking, and continues to be one of the most rapidly growing areas in this domain. Moreover, published findings reflect only a fraction of the data originally collected. The data themselves take a variety of forms and typically are not accessible for widespread sharing and use. Making neuroimaging data more accessible for sharing would facilitate the comparison of findings across laboratories.

However, unlike fields in which databasing efforts have been successful, there are no universally accepted standards for the structure and content of neuroimaging data sets. Data formats vary widely across different laboratories and neuroimaging methods. This diversity of data and formats reflects several factors, including the rapid developments within the field and the rapid changes in our knowledge about brain. Indeed, neuroimaging methods are evolving quickly, and changes are sometimes accompanied by substantial changes in data content that directly affect the format. For example, with the transition from anatomical to functional MRI, data formats changed from three to four dimensional. More critically, imaging of brain function demands a clear specification of the behavioral conditions under which the data were acquired. This is linked to the scientific hypotheses being tested. The number of potentially important behavioral variables is large and poorly defined.

The figure 2.1 shows a part of the issues related to data sharing that need to be answered when researchers work on such projects.

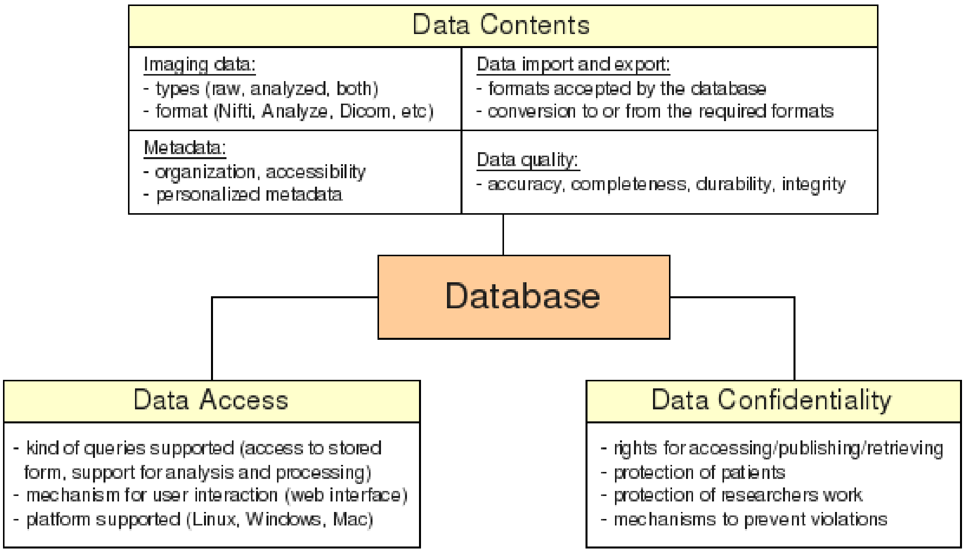

Data Contents
Imaging data:
- types (raw, analyzed, both)
- format (Nifti, Analyze, Dicom, etc)
Data import and export:
- formats accepted by the database
- conversion to or from the required formats
Metadata:
- organization, accessibility
- personalized metadata
Data quality:
- accuracy, completeness, durability, integrity
Database
Data Access
- kind of queries supported (access to stored form, support for analysis and processing)
- mechanism for user interaction (web interface)
- platform supported (Linux, Windows, Mac)
Data Confidentiality
- rights for accessing/publishing/retrieving
- protection of patients
- protection of researchers work
- mechanisms to prevent violations

Figure 2.1: Issues related to neuroimaging databases

Faced to these issues and recognizing the increasing need for data sharing within the neuroimaging community, several groups have conducted researches to develop neuroimaging databases. Different models have been explored. Some use a centralized model that directly manages the storage and distribution of data. Others use a distribution model, in which a centralized listing is maintained that describes the available data and their location, whereas the data themselves are stored locally, under the control of their owner. The diversity of databases projects is a good thing to explore various approaches and to value each of them. In the next pages, we will review some of the most interesting projects conducted by team across the world, which proves the importance given to this research since several years.

2.2 The Human Imaging Database, by the BIRN

The Biomedical Informatics Research Network (BIRN) is a research American project launched in 2001 and supported by the National Institutes of Health's National Center for Research Resources (NCRR). It is one of the best known and complete work related to the sharing of data in the field of biomedical. It is composed by around 35 laboratories across the United States of America. The main goal of the projects is to propose a geographically distributed virtual community of shared resources associated with a large range of tools.

As part of their huge work, the BIRN has been working on the Human Imaging Database (HID) which is an extensible database management system. It has been developed and implemented to address the problems associated with managing the increasingly large and diverse datasets collected as part of the morphometry and function BIRN (mBIRN, fBIRN) collaboratories. This system is comprised of three core components, among them, The Human Clinical and Imaging Database itself and an intuitive web based user interface.

The database is composed of an extensible schema and structured core. The core database contains a hierarchical description of an experiment and how experimental protocols relate to this hierarchy. Then, the database at a particular site can be extended to contain relevant information concerning the research subjects used in an experiment, subjects assessments, the experimental data collected, the experimental protocols used and any annotation or statistics (metadata) normally included with an experiment. The database can be extended utilizing extended tuples which can be re-used and/or modified for other experiments. The complete system is used to manage and query local data at various research sites within BIRN. In addition to local operations, the system allows for mediated queries across multiple federated databases allowing researchers to discover data across all relevant sites. There are currently 12 federated HID databases, 11 Oracle and 1 PostgreSQL versions, storing clinical information.

Regarding the user interface, the figure 2.2 shows his overall structure based on three tier Java 2 Enterprise Edition (J2EE) architecture. It consists of a client tier, a server/JSP based middle tier and a relational database based enterprise information source (EIS) tier. The EIS tier consists of the Human Clinical and Imaging Database and a collection of stored procedures/packages for low level data access functions. The middle tier currently consists of a web tier tier only. The underlying web application framework used for user interface is Jakarta Struts. It uses a controller servlet to intercept a web request and determine what to display next. The business logic layer is defined as the code manipulating business data (e.g. clinical assessments) relevant to the application. DAO is for Data Access Objects. Each software layer communicates with neighbor layers via well defined interfaces, which remain stable while the actual implementation can change drastically over time facilitating software maintenance and robustness.

Thus, the projects conducted by the BIRN are very interesting in our case, especially the Human Imaging Database which provide a web-based interface and a relational database back-end for storing and managing brain imaging data. I will come back later on BIRN's works regarding the ontologies with NeuroLex (formerly BIRNLex).

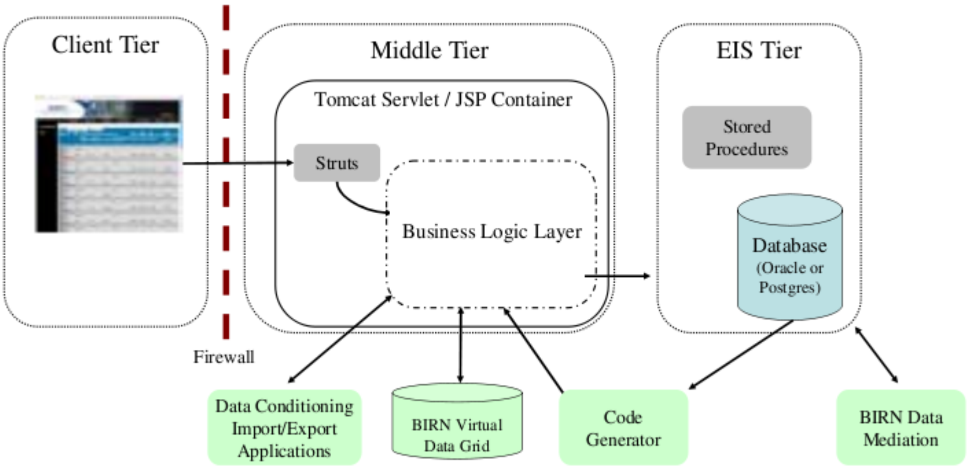
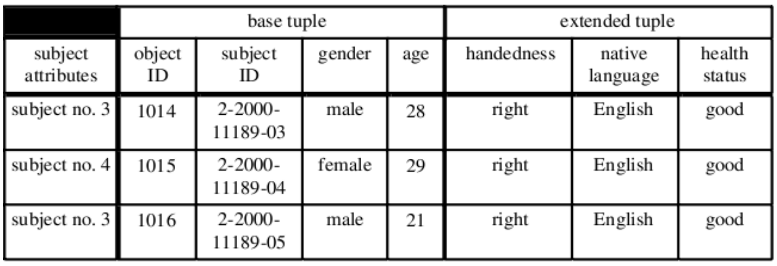


Figure 2.2: Human Imaging Database Graphical User Interface structure[1]

2.3 The Functional Magnetic Resonance Imaging Data Center

The Functional Magnetic Resonance Imaging Data Center (fMRIDC) founded in 1999, has been thought to establish a facility for the sharing of functional neuroimaging data within the community of neuroscience. The Data Center has received during several years study data from fMRI analyses published in the Journal of Cognitive Neuroscience. Researchers exchange data with the Data Center through two simple means: contributing study data to the archive, in which researchers fill out detailed forms describing study protocols and individual subject information, and requesting study data from the record of studies in the repository.

To address the problems associated with managing the increasingly large and diverse datasets collected, an object-relational database schema called Neurocore was developed. By exploiting the object-oriented properties of object-relational database technologies, Neurocore is a database architecture that is completely modifiable while maintaining a standard core structure. Thus, this methodology can be used to extend the database to contain relevant information concerning the research subjects used in an experiment, the experimental data collected, the experimental protocols used and any annotations or statistics normally included with an experiment. The core database contains a hierarchical description of the experiment and how experimental protocols relate to this hierarchy. The ability to extend experimental descriptions in the database is accomplished through the use of object-relational technologies. This is very similar as the previous project (BIRN) exposed above.



	base tuple				extended tuple		
subject attributes	object ID	subject ID	gender	age	handedness	native language	health status
subject no. 3	1014	2-2000-11189-03	male	28	right	English	good
subject no. 4	1015	2-2000-11189-04	female	29	right	English	good
subject no. 3	1016	2-2000-11189-05	male	21	right	English	good

Figure 2.3: Descriptor "subject" in the database[2]

As shown in the figure 2.3, the inherent attributes for an entity constitute the "base tuple" that defines the minimum informational requirements for that entity. These base tuples can then

[1] Extracted from Ozyurt and al.,(2004) Web- accessible clinical data management within an extensible neuroimaging database. Society for Neuroscience, Washington, DC, 2005 Online

[2] Extracted from Van Horn, J.D., et al., The Functional Magnetic Resonance Imaging Data Center (fMRIDC): the challenges and rewards of large-scale databasing of neuroimaging studies, Science 356 (2001), 1323-1339

be extended with additional attributes through the definition of an extended tuple. Furthermore, the extended tuples can be reused and/or modified for other experiments, and be used to guide future interactive data entry forms.

Regarding the physical architecture of the fMRIDC computional resource, it is implemented in three tiers: the client Web browser (using HTML, XML, Java and Javascript), the web server running PHP and Python software and the database management system itself.

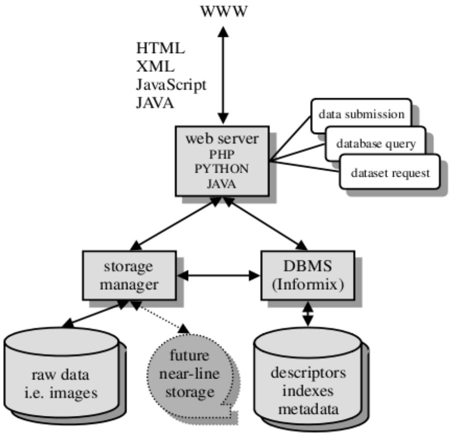


Figure 2.4: Physical architecture of the fMRIDC[3]

Even if this project is relatively old compared to others presented here, it still remains very interesting by giving a specific structure for the overall system and by having specifics goals. Indeed, one of their main goal is to allow small and/or limited funded laboratories to access to a large bank of data in order to be able to reliably reproduced the fMRI results. Of course, they have implemented a very specific workflow to add data in the system in order to respect the human subject as well as the rights of the original authors. Unfortunately, the fMRIDC doesn't accept any new datasets but the others datasets can still be found on their website.

2.4 The Computional Neuroscience Applications Research Infrastructure

2.5 The Extensible Neuroimaging Archive Toolkit

2.6 A work about other projects

[3] Extracted from Van Horn, J.D., et al., The Functional Magnetic Resonance Imaging Data Center (fMRIDC): the challenges and rewards of large-scale databasing of neuroimaging studies, Science 356 (2001), 1323-1339