Towards A Virtual Laboratory for fMRI Data Management and Analysis

Silvia D. Olabarriaga [a,1], Aart J. Nederveen [b], Jeroen G. Snel [b] and
Robert G. Belleman [a]
[a] *Informatics Institute, University of Amsterdam*
[b] *Academic Medical Center of the University of Amsterdam*

Abstract. Functional Magnetic Resonance Imaging (fMRI) is a popular tool used in neuroscience research to study brain activation due to motor or cognitive stimulation. In fMRI studies, large amounts of data are acquired, processed, compared, annotated, shared by many users and archived for future reference. As such, fMRI studies have characteristics of applications that can benefit from grid computation approaches, in which users associated with virtual organizations can share high performance and large capacity computational resources. In the Virtual Laboratory for e-Science (VL-e) Project, initial steps have been taken to build a grid-enabled infrastructure to facilitate data management and analysis for fMRI. This article presents our current efforts for the construction of this infrastructure. We start with a brief overview of fMRI, and proceed with an analysis of the existing problems from a data management perspective. A description of the proposed infrastructure is presented, and the current status of the implementation is described with a few preliminary conclusions.

Keywords. medical image analysis, grid computing, functional MRI, virtual organizations IT for large population studies

1. Introduction

Functional magnetic resonance imaging (fMRI) is a popular tool used in neuroscience research to study brain function. In fMRI studies, large amounts of data are acquired, processed, compared, annotated and stored for future reference. The users of fMRI (in particular psychologists, psychiatrists, radiologists etc.) typically have limited technical background in computing and, as such, face several difficulties to organize their workflow comfortably and efficiently based on individual solutions available for personal computers. Computational resources with higher capacity and performance are needed to properly address the needs of these users.

The Virtual Laboratory for e-Science (VL-e) Project[1] has taken initial steps to build a grid-enabled infrastructure to facilitate data management and analysis

---

[1] Correspondence to: Silvia D. Olabarriaga, Kruislaan 403, 1068 SJ, Amsterdam. Tel.: +31 20 525 7549; E-mail: silvia@science.uva.nl

[1] http://www.vl-e.nl/

for fMRI studies. This article presents our current efforts for the construction of this infrastructure. We start with a brief overview of fMRI (section 2), and proceed with an analysis of the existing problems from a data perspective (section 3). The proposed infrastructure is described in section 4, followed by a brief discussion and preliminary conclusions in section 5.

2. fMRI at a Glance

fMRI enables the study of brain activation in a non-invasive manner. The basic idea is to scan a subject while he/she is submitted to brain *stimulation* through a physical or cognitive activity. Depending on the type of sensory stimulus (visual, auditory, motor, etc.) or cognitive task, the neuronal activity increases in different parts of the brain, and a heamodynamic response occurs. In simple terms, the active region receives oxygen-rich blood, and the changes in the oxygenation level can be measured with MRI.

An fMRI scanning session produces a series of 3D datasets (volumetric images) containing measurements along time, some obtained during stimulation and some at “rest”. These images are subsequently analysed to determine the location of activated areas – refer to [1] for details. First the 3D volumes in the time series are aligned to each other to compensate for artefacts introduced by temporal sampling and motion. Additionally, filters are applied to reduce noise and normalise the measurements. Next, statistical analysis is performed to correlate the measured signal with the stimulation pattern. This step generates a *statistical map*, which is again submitted to statistical analysis to detect activation based on an adaptive threshold. The final result of the analysis is an *activation map* that can be further analysed to determine the location and size of activation clusters or *activated regions.* Activation maps are overlaid to an additional high-resolution structural scan for visual inspection of activation with respect to the anatomy. Instead of a structural scan of the same subject, a reference brain can be used, e.g. the Montreal Neurological Institute average brain[2].

fMRI is largely used in neuroscience studies, for example, to characterize brain function in populations. Often the activated regions detected in different scans are compared, which requires additional image registration steps for their alignment to a common coordinate system. A future perspective is that fMRI will also be used in a broad range of situations in diagnosis, prognosis and treatment planning. Examples include aids to detect anomaly, prediction of functional damage due to trauma, or planning of neurosurgery [3]. And finally, the acquired data and metadata (results, annotations) are typically shared in large multi-centre studies, which are becoming increasingly popular for the characterization of large populations in neurosciences or for the evaluation of new healthcare procedures.

3. Data-related Issues in fMRI Studies

From a data perspective, fMRI studies involve data *acquisition*, *storage*, *analysis* and *shared access.* Note that “data” here refers to a large variety of information and measurements acquired or generated during an fMRI study, being heterogeneous by nature. Examples of data are scanned images (functional, structural), information about the applied stimuli, parameters of the acquisition protocol, subject characterization (e.g., age, gender, pathology), results of the analysis (e.g., statistical and activation maps, locations of activated regions) and interpretation (e.g., annotations). These data are generated by different types of physically dispersed equipments and image analysis utilities, requiring a significant amount of time and effort for adequate management. Such effort is likely to increase as the amount of data grows in response to developments in scanning techniques, analysis methods, and collaborative and multi-centre research. Below we discuss a few of the problems encountered.

Data acquisition in this discussion is restricted to the images and associated signals recorded during an fMRI scanning session, involving a collection of equipments in complex experiments[2]. Experiment design is based on prior knowledge about brain function and imaging protocols, which is typically accumulated and shared informally by researchers and practitioners in the field. One of the difficulties is to have access to resources with structured information about experimental design (e.g., databases of acquisition protocols or stimuli). By keeping documentation of experimental procedures, such resources could facilitate the validation of experiments, the design of new experiments and the standardization of existing ones. Experiment control involves synchronizing image acquisition (by a *scanner*) with stimulation (by a *stimulus computer*), as well as recording images and other signals (e.g. electroencephalography - EEG). At the end of an fMRI session, all data (images, stimuli, signals, etc.) are gathered from the acquisition equipments and exported to a remote storage resource. The acquisition equipments are often dispersed, heterogeneous, and located behind a hospital firewall. In a clinical setting, the images are stored in a Picture Archival and Communication System (PACS), while other data are manually transported using some physical medium (DVD, memory stick) or `sftp`. In a typical research setting, also the images are manually transported to the external storage resource for further analysis. Data gathering could be facilitated by connecting all data acquisition equipments in the scanning site directly and securely to a (remote) storage resource that can store all the collected data.

Data storage for fMRI studies present three main difficulties. First, large storage capacity is needed, since studies involve many instances (typically above 20) of large datasets (500MB to 1GB per scanning session). Second, the storage system should be flexible to accommodate heterogeneous data types such as images, signals, and metadata. Although the adoption of a PACS would be the natural choice in a medical environment, current systems are still limited with respect to data capacity, image format, and storage of non-pixel data. And finally, when patient data is used in these studies, high demands are imposed on data confidentiality. Not only secure connection is required, but also all identity information must be striped from the data before they leave the scanning site.

Data analysis in fMRI involves applying complex and computation-intensive image processing methods to large amounts of data. It can take more than one

---

[2] In fact, data and metadata are collected during the whole lifetime of a study.

hour to complete the analysis of one fMRI session on workstations that are typically available for researchers in their home institutions. Normally the analysis is executed as a post-processing step, and the results become available a long time after the scanning session. If a problem is detected during the analysis (e.g., motion artefacts), a new scanning session must be scheduled. The usage of high performance computing (HPC) resources could be helpful to reduce latency and enable interactive inspection of data quality while the subject is still at the scanning site. Additional problems are faced for running image analysis in large scale, for example, when the study includes a large number of subjects or for performing parameter optimisation. The complete analysis in these cases can take days to be completed on typical workstations. HPC capacity would be beneficial here for achieving higher throughput by parallel execution of independent tasks (e.g., analysis of individual scans). Moreover, the logistics of data and computational resources require much effort to guarantee proper error handling, enough storage space for intermediate results, proper data conversion and transfer, proper parameter settings, etc. The researchers involved in fMRI usually do not have enough technical knowledge to set-up an infrastructure to perform reliable, efficient and secure image analysis. These users could benefit from sharing a common IT infrastructure for large-scale fMRI studies, in which the workflow can be automated.

Data access and sharing are challenging issues because fMRI studies are performed by groups of users associated to the same institution or to multiple centres. Moreover, a growing trend in neuroscience is to share the acquired data and generated metadata (results) with other researchers after the study is completed [4]. In this manner data can be reused, experiments can be reproduced or repeated with different settings, or results can be used for meta-analysis. The following issues characterize the demands for shared data access in this context. First, multiple and physically dispersed sites are involved, requiring remote and secure access to (distributed) data. Second, it is necessary to control and monitor access to data, respecting strict data privacy policies that may be different per site or study. Third, large amounts of data are involved, requiring mechanisms for efficient retrieval such as query based on metadata. And finally, data should be archived for long periods of time, requiring extremely large and permanent storage capacity.

The scenario described above indicates that a proper IT infrastructure is fundamental to accomplish fMRI studies successfully. Table 1 summarizes the different problems faced for data management in fMRI studies and the challenges posed to the construction of an adequate infrastructure.

4. A Virtual Laboratory for fMRI: VL-f

The construction of an IT infrastructure addressing the challenges in Table 1 obviously requires technical knowledge that is beyond the scope of neuroscience, and perhaps also beyond what could be accomplished with traditional computing paradigms. The characteristics of this application indicate, on the other hand, that it could benefit from grid computing approaches [5] for the following reasons.

Table 1. Difficulties for data management in fMRI and associated IT challenges.

Characteristics	Challenges
Acquisition:	
Complex experiments	Share and reuse experiment design
Multiple equipments	Access dispersed and heterogeneous systems
Storage:	
Many large data instances	Large storage capacity
Heterogeneous data types	Flexible storage system
Patient data	Data confidentiality
Analysis:	
Computation-intensive analysis	HPC for throughput
Interactive response	HPC for real-time computation
Large scale processing	Logistics of data and resources
Shared Access:	
Multiple centres	Remote access to distributed data
Many users	Controlled access to (confidential) data
Large amounts of data	Query based on metadata
Data Archival	Long term storage/retrieval

First, fMRI studies are *data intensive*, since large amounts of data are stored, analysed and manipulated. Second, they require *high throughput computation on demand* for real-time image analysis and for large scale studies. Finally, *collaboration and distributed computing* are essential, in particular for multi-centre studies, in which data is acquired, analysed and shared from different locations. This application therefore requires coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations (VOs), which is the goal of grid technologies [6].

A virtual laboratory for fMRI (VL-f) is under development in the scope of the VL-e project to address some of the challenges listed in Table 1. The goal is to construct a shared computational infrastructure with hardware, software and services to efficiently, reliably and securely perform (large scale) fMRI studies. The following specific goals will be pursued:

1. Facilitate data gathering at the scanning site, by providing homogeneous access to the acquisition equipments.
2. Facilitate data storage and archival, by providing access to large capacity and long-term storage resources.
3. Enable high data analysis throughput, by providing access to HPC resources to perform parallel analysis of mutually independent data.
4. Facilitate the data logistics in (large scale) fMRI studies, by providing tools to automate the workflow (data gathering, removal of subject identity, data format conversion, and image analysis).
5. Provide remote data access via interactive interface to the storage resource from workstations located anywhere.

6. Enable secure data sharing, by providing mechanisms for controlling access to the data for users and groups.
7. Facilitate data retrieval, by providing infrastructure for generation of metadata and query mechanisms based on metadata.

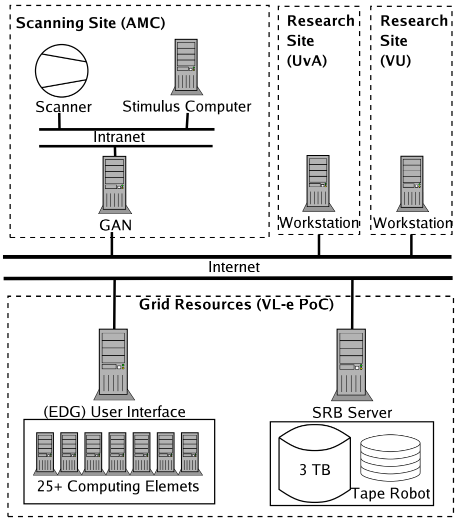

Figure 1. Computational resources of the VL-f.

Below we present a description of the resources (section 4.1), the ideal use scenario pursued by VL-f (section 4.2), the plans for the first pilot implementation (section 4.3) and its current status (section 4.4).

4.1. VL-f Resources

The simplified scheme in Figure 1 presents the hardware and software resources of VL-f, which are distributed among scanning, research and services sites.

Scanning sites are the locations where a scanner and other acquisition devices are installed, normally in the radiology department of a hospital. These equipments are connected to each other directly, and possibly also to others, via an internal network (intranet) protected by a firewall. They are accessible only from workstations located in their physical vicinity (e.g., examination room). Some workstations have access to public networks (e.g., internet), being called *grid access nodes* (GANs) in the proposed scheme. In the first phase of VL-f, scanning is performed at the radiology department of the Amsterdam Medical Center (AMC), which includes a Philips 3T Intera MRI scanner, a stimulus computer, and a GAN.

Research sites are locations where the (neuro)scientist interacts with the data from a workstation. In the first phase of VL-f, the research sites are located at several departments of the University of Amsterdam (UvA), the Free University (VU), and private computing facilities (e.g. at home). Workstations based on Windows or Linux platforms will be supported.

Services sites provide compute and storage resources. Although an open grid-oriented architecture will be adopted, in the first phase of VL-f only the grid resources provided by the VL-e Proof-of-Concept Environment (VL-e PoC) are used. These resources are provided by SARA Computing and Networking Services[3] and consist of computing elements and on- and near-line storage elements. Access to the hardware facilities and software services will be granted to authorized users associated to one or more VOs.

The computing elements are accessed via an *user interface* machine linked to the European DataGrid (EDG). In a first phase, only the Matrix cluster at SARA will be available for the fMRI VO. This cluster consists of 36 IBM x335 nodes equipped with dual Xeon 3.06GHz processors, 2GB memory, and 2 local EIDE disks of 120GB. The nodes are connected with 1 Gb/s Ethernet. Other computers and clusters located at other VL-e partners will be added in the future (e.g., National Institute for Nuclear Physics and High Energy Physics, NIKHEF).

Storage resources include on-line storage and tape silos for off-line, permanent and unlimited storage space. The data is transported automatically between the on- and off-line storage systems based on usage patterns. Data integrity and accessibility is provided by automatic and periodic back-ups. The on-line storage resource consists of the Storage Resource Broker (SRB[4]) system, which provides a seamless interface to store and retrieve data and metadata across a wide area network [7].

4.2. Ideal Scenario

The ideal functional scenario pursued by VL-f is illustrated in Figure 2.

When a scanning session is completed, data from the scanner and stimulus computer are gathered into a single workstation at the scanning site. Identity information is removed from the images with an application that additionally provides aids to control pseudonyms and real identities in the context of individual or multiple studies. The user schedules the transfer of identity-free data to the SRB, performing SRB authentication with a Grid Security Infrastructure (GSI[5]) certification protocol. Metadata encapsulated by the file format (e.g., DICOM) is automatically associated with the data upon upload. The data are transferred to the SRB, and the user is notified via e-mail when and where (`uri`) the data have been successfully stored.

The user schedules data conversion (images, stimuli data) from a workstation at the research site, indicating that the source and destination files are stored in the SRB. Grid and SRB authentication are used to enable access to the storage and the VL-e PoC computational resources, where the data conversion job is performed. The user is notified via e-mail when the conversion is complete and where the results have been stored.

Several software packages can be chosen for image analysis, for example `FSL` (fMRIB Software Library [8]) and `SPM` (Statistical Parametric Mapping [9]). The user configures the image analysis parameters from a workstation at the research

---

[3] http://www.sara.nl/

[4] http://www.sdsc.edu/srb

[5] http://www.globus.org/security/overview.html

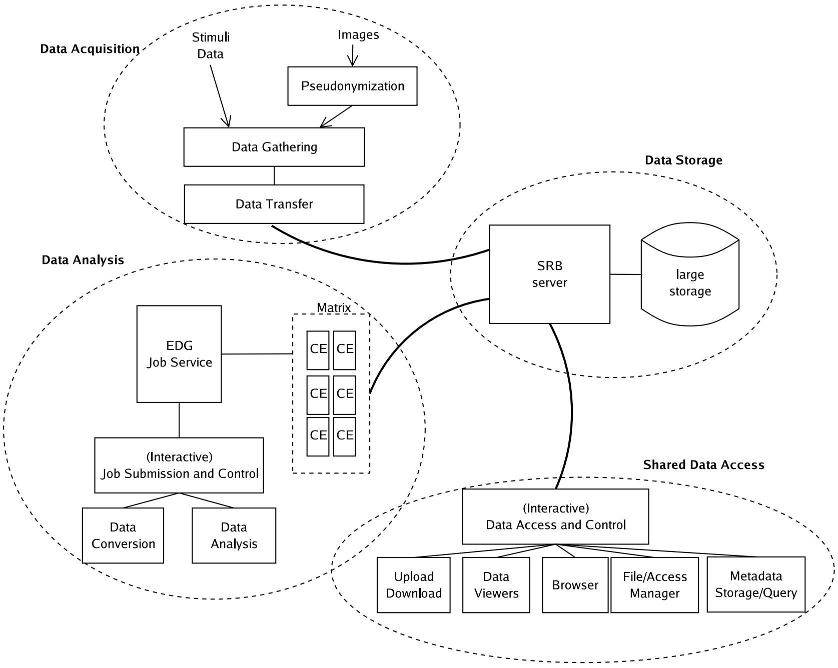

Figure 2. Overview of VL-f functional components.

site, indicating that the input and output files are stored in the SRB. Authentication (SRB, Grid) takes place, and the analysis job is scheduled to run at the VL-e PoC computing resources. The user is notified when the analysis is completed, with a link to the results in the SRB.

The LCG-2 grid middleware[6] is used for running jobs to perform data conversion and image analysis on the VL-e PoC. Job submission is performed via a web service that encapsulates the functionality of EDG command-line utilities. Jobs are retried a given number of times, and permanent faulty conditions are notified both to the user (researcher) and to the technical support (image analysis or grid specialist, depending on the problem).

At any time, and from any workstation at the research sites, the user can perform the following operations via interactive and intuitive clients and/or web portals:

- monitor and control the status of scheduled jobs for data transfer, data conversion and image analysis,
- browse data and inspect their content with `html`, text, image, and other specialized viewers,
- transfer data between the SRB and the local workstation,
- manage files and control data access permissions for individuals and groups, on a permanent or temporary basis,
- edit metadata stored in the SRB metadata catalogue, and

---

[6] http://lcg.web.cern.ch/LCG/activities/middleware.html

- retrieve data with queries based on metadata

And finally, tools for workflow automation enable the user to combine and schedule at once tasks such as data pseudonymisation, data transfer to the SRB, data conversion, image analysis and metadata generation.

4.3. First Pilot

A minimum but complete subset of the ideal functionality described in section 4.2 was selected for implementation in the first pilot. Existing software is used as much as possible to enable rapid development. Issues such as optimisation, development of intuitive GUIs, cross-platform functionality and workflow automation were left for a later phase.

Data gathering is performed in a workstation that has access to the file system of the scanner and the stimulus computer as a remote drive or directory. Pseudonymisation, if necessary, is performed by an application that simply removes identity information from 4-D images stored in the PAR/REC Philips Medical Systems proprietary file format. The data is transferred into the SRB using `inQ`[7], a browser for Windows platforms, which only supports password authentication. `inQ` is also used for browsing data on the SRB, controlling user access to groups and individuals, transferring data between the SRB and the workstation, general file management, metadata management and query.

Image analysis is performed with FSL and consists of a sequence of customized steps implemented by command-line utilities (binary code and scripts). These steps have been encapsulated in FSL by a `tcl` script that takes parameters from a single configuration file. Some parameters are used to control image analysis, while others indicate the location of input data (complete file paths to images and stimuli data) and output results. The analysis results consist of several files (images, text) that are stored in a single directory with a given name. A summary report in `html` generated by FLS facilitates browsing the results.

For proper execution in the VL-e PoC, the FSL script needs to be wrapped into a higher-level component that handles files stored in the SRB and adequate user notification. The input files are automatically downloaded to the local file system of the computing node prior to running the original script, and the results are uploaded when it completes. Error handling, which is limited to displaying messages in the `stderr` in the original script, must be extended to also notify the user and the technical support. The analysis is started manually, using a job submission client running on the workstation at the research site. Lists of jobs (e.g., for several scans) can be submitted at once, in which case the analysis is performed in parallel in the available computing elements.

Data conversion facilities are limited to the formats required by the FSL utilities. For images, conversion is performed from PAR/REC to NIFTI-1[8] format.

---

[7] http://www.sdsc.edu/srb/inQ/inQ.html

[8] See http://www.fmrib.ox.ac.uk/fsl/fsl/formats.html.

4.4. Current Status

The implementation of this pilot is in its early stage. Images can be transferred to the SRB directly, but the data from the stimulus equipment still need manual intervention. Only MS Windows-based workstations are supported for interactive access to the SRB with `inQ`. `Scommands`[9] are used to upload and download data from the scripts that perform data manipulation autonomously. The data conversion and image analysis tasks have been combined into one script that is executed as a single job. Jobs are scheduled and monitored via existing utilities that offer an interactive GUI for EDG job submission. These utilities must be executed on the EDG user interface machine. Finally, no explicit metadata facilities nor specialized data viewers are present, except for those already implemented by the SRB and `inQ`.

5. Discussion and Conclusions

Other attempts have been reported to provide grid-enabled infrastructures for medical applications. Montagnat et al. present in [10] several medical applications that can benefit from grid technology by using parallel and distributed computing for higher throughput and lower latency. Rogulin et al. describe in [11] the Mammogrid project, which uses grid technology to integrate and provide access to databases of mammograms for computer-aided diagnosis. Barillot et al. describe in [12] the Neurobase project, which uses grid technology for the integration and sharing of heterogeneous sources of information in neuroimaging. Rex et al. present in [13] the LONI Pipeline Processing Environment, a cross-platform, distributed environment for the design, distribution, and execution of image analysis for neuroimaging applications. It has a visual programming interface where a large repertoire of components can be combined to perform the desired image analysis steps. A dataflow model is adopted to support a parallel processing architecture, which enables simultaneous execution of multiple tasks. These attempts have emphasised either information or computation aspects. Our efforts, however, are focused on the construction of an infrastructure that addresses both aspects transparently, efficiently and robustly.

The infrastructure proposed in section 4 has the potential to alleviate to a large extent the problems presented in section 3 because it provides large and long term storage capacity, remote and controlled access to distributed and heterogeneous data, facilities for metadata storage and query, access to HPC resources, and workflow automation. This potential remains to be confirmed when the implementation is completed and evaluated from the perspective of the end users.

The few experiences with the pilot already indicate that the implementation of VL-f will be a challenging task, and that several issues should be properly addressed before the infrastructure could be considered useful. First, error detection and notification is typically poor in legacy software, being typically limited to messages written to files (`stdout, stderr`) or a return code. It would be desir-

---

[9] http://www.sdsc.edu/srb/scommands/index.html

able to clearly notify failure and success in a compact manner to more efficiently guide the user in the inspection of results. The relevance of such a feature will increase proportionally with the scale and automation level of workflows. Second, more intuitive tools are needed to submit, monitor and control job execution, in particular for large numbers of jobs. We are currently investigating Nimrod-G [14] as an alternative for large scale job submission. Third, it is important to provide simple means to request grid services (for job submission) from any workstation at the research sites. We are planning to develop platform-independent clients that will communicate with the EDG web service implemented at the VL-e PoC to submit and control/monitor large numbers of jobs. This service is compliant with the Web Services Resource Framework (WSRF) architecture, which is under consideration also to implement functionality such as data conversion and image analysis. And finally, workflow automation must be improved, for example, by integrating the VL-f functionality into the Distributed Workflow Management System currently in use at the AMC [15].

Note that the proposed infrastructure does not include explicit mechanisms for strict management of data confidentiality, besides removing identity information and controlling access to the data via SRB authentication. Although the current strategy may be insufficient for handling patient data, it seems adequate for a large number of research studies in which the subjects are volunteers. Constructing an adequate and useful IT infrastructure, even if for a limited scope of fMRI studies, is the goal of our current efforts in VL-f.

Acknowledgements

This work was carried out in the context of the VL-e project. This project is supported by a BSIK grant from the Dutch Ministry of Education, Culture and Science (OC&W) and is part of the ICT innovation program of the Ministry of Economic Affairs (EZ).

References

[1] S.M. Smith. Overview of fMRI analysis. *The British Journal of Radiology*, 77:S167–S175, 2004.

[2] A. C. Evans, D. L. Collins, , R. R. Mills, E. D. Brown, R. L. Kelly, and T. M. Peters. 3D statistical neuroanatomical models from 305 MRI volumes. In *Proceedings of the Nuclear Science Symposium and Medical Imaging Conference*, volume 3, pages 1813 – 1817. IEEE, 1993.

[3] E.J. Vlieger et al. Functional magnetic resonance imaging for neurosurgical planning in neuro-oncology. *Eur Radiology*, 14:1143–1153, 2004.

[4] J. D. Van Horn, S. T. Grafton, D. Rockmore, and M. S. Gazzaniga. Sharing neuroimaging studies of human cognition. *Nature Neuroscience*, 7(5):473–481, 2004.

[5] I. Foster and C. Kesselman. Computational grids. *Communications of the ACM*, 35(6):44 – 52, 1998.

[6] I. Foster, C. Kesselman, and S. Tuecke. The anatomy of the grid: Enabling scalable virtual organizations. *International J. Supercomputer Applications*, 15(3), 2001.

[7] C. Baru, R. Moore, A. Rajasekar, and M. Wan. The SDSC storage resource broker. In *CASCON'98 Conference*, 1998.

[8] S.M. Smith et al. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23:208–219, 2004.

[9] K.J. Friston. Statistical parametric mapping and other analysis of functional imaging data. In *Brain Mapping: The Methods*, pages 363 – 385. Academic Press, 1996.

[10] J. Montagnat et al. Medical Images Simulation, Storage, and Processing on the European DataGrid Testbed. *Journal of Grid Computing*, 2:387–400, 2004.

[11] Rogulin D. et al. A grid information infrastructure for medical image analysis. In *Proceedings of DiDaMIC Workshop (MICCAI'2004)*, 2004.

[12] C. Barillot et al. Neurobase: Management of distributed and heterogeneous information sources in neuroimaging. In *Proceedings of DiDaMIC Workshop (MICCAI'2004)*, 2004.

[13] D.E. Rex J.Q. Ma and A.W. Toga. The LONI pipeline processing environment. *NeuroImage*, 19:1033–1048, 2003.

[14] R. Buyya, D. Abramson, and J. Giddy. Nimrod-GResource Broker for Service-Oriented Grid Computing. *IEEE Distributed Systems Online*, 2(7), 2001.

[15] J. G. Snel, S. D. Olabarriaga, J. Alkemade, H. G. van Andel, A. J. Nederveen, C. B. Majoie, G. J. den Heeten, M. van Straten, and R. G. Belleman. A distributed workflow management system for automated medical image analysis and logistics. In *accepted to IEEE-CMBS, special track on Grids for Biomedical Informatics*, 2006.