Sharing data and image processing tools in neuroimaging.

Bernard Gibaud, Christian Barillot, Habib Benali, Michel Dojat, Alban Gaignard, Serge Kinkingnéhun, Jean-Pierre Matsumoto, Mélanie Pélégrini-Issac, Eric Simon, Lynda Temal

Sharing data and image processing tools in neuroimaging

B. Gibaud[a], C. Barillot[a], H. Benali[b], M. Dojat[c], A. Gaignard[a], S. Kinkingnéhun[b], J-P. Matsumoto[d], M. Pélégrini-Issac[b], E. Simon[d], L.Temal[a]

[a]Unit/Project VisAGeS U746, INSERM/INRIA/CNRS/U. of Rennes I, Rennes, France
[b]IFR 49, CHR La Pitié Salpetrière/CEA-SHFJ, Paris, Orsay, France
[c]Unité INSERM U594, Grenoble, France
[d]Medience SA, INRIA, Rocquencourt, France

Abstract. Research in neuroscience makes an extensive use of modern neuroimaging data. The ability to share such data as well as processing tools to analyze them becomes a key factor of success of future research. This paper reports about works carried out in the context of the Neurobase project, a collaborative exploratory action supported by the French Ministry of Research. The partners have studied the design of a system suitable for sharing heterogeneous data and image processing tools, according to a federated approach, based on a common ontology. A demonstrator has been implemented, providing basic services for querying data and launching processing tools distributed at several sites.

Keywords: data integration; distributed image processing; ontology; neuroimaging; GRIDs

1. Purpose

1.1. Context

Research about human brain has received considerable attention during the two last decades. The tremendous progresses of biomedical imaging, molecular biology, genomics, and the widespread availability of these techniques contribute to the creation of huge quantities of highly relevant data. In parallel, sophisticated imaging processing tools have been developed in several centres to cope with the specificities of neuroimaging data. Sharing and reuse of these data and tools have become a real challenge [1,2,3,4,5]. However, in spite of the wide scale deployment of the internet, few practical solutions are at hand to fulfil the needs of researchers and clinicians.

This paper specifically deals with the sharing and reuse of brain image data and image processing tools. A first objective is to federate distributed data repositories, containing the image data produced in various research contexts. From an application viewpoint, the goal of database federation is to overcome the limits of research on brain structure, function or pathology, presently restricted to quite small populations of subjects. A second objective is to be able to share image processing tools. Indeed, the classical way of disseminating such tools, through web-distribution of software packages (e.g. SPM, FSL or Brainvisa), is relevant but limited. Firstly, each package has to be installed in the user environment (requiring the relevant hardware, operating system and libraries), and secondly these components have to be manually integrated into dedicated problem-specific pipelines, a generally complex operation because such components have not been designed to interoperate.

Another motivation is reuse. Actually, research funding organizations and stakeholders encourage a wide dissemination of research data, for example, to reproduce the same results with alternative methods, or to reuse these data in the context of other research projects. Actually, only specific aspects of research data can be exploited at the level of each research centre due to limited skills, manpower and

resources. For image processing tools, dissemination can facilitate their comparison and their evaluation on various data sets. Moreover, it is highly relevant to combine them to produce new complex data processing pipelines, tailored to the specific needs of applications. Similarly, the added value of such new tools should be evaluated and diffused to the entire community.

1.2. Objective

The general goal of our research was to mature an up-to-date understanding of the problem, and to provide a demonstration of the feasibility of a federated system, capable of supporting the sharing and reuse of distributed neuroimaging data and tools.

2. Methods

Basic assumptions: a possible solution for resource sharing consists of gathering all data and processing tools in a central database and data processing resource. We consider this solution as unrealistic because of an inherent lack of flexibility and of a legitimate aspiration to autonomy of each research centre regarding the management of its own data repository and computer facilities. Then, we advocate for the use of a federated, rather than centralized, approach. A major difficulty arises from the fact that data collected in various independent repositories are usually heterogeneous. This means that the structure and semantics of the stored information are not the same, and must be reconciled. Our approach relies on a mediator/wrapper architecture [6]. Two possible strategies are then possible [7]: (1) a "global as view" approach, in which the common view is built from the concepts existing in the federated repositories; (2) a "local as view" approach, which maps the concepts of a given repository with those defined *a priori* in a common domain ontology.

We chose the latter approach. In the former, the insertion of new repositories to an existing federated system would generally lead to significant changes in the system. The latter approach assumes that a common ontology exists *a priori*. Then, the drawback of this approach is that the feasibility of building a common domain ontology in such a complex domain as "neuroimaging", is not yet demonstrated. Presently, we defined a domain ontology for the needs of our target applications.

2.1. Creation of a "neuroimaging" domain ontology

The scope of this ontology is image processing in neuroimaging, for two main applications in cognitive research (human vision) and clinical research (epilepsy).

Our work consisted first is identifying and modelling the relevant concepts, from existing knowledge sources such as the fMRIDC (fMRI Data Center) ontology, as well as the available standards (primarily DICOM), and from the experience of the project participants. In a second stage, we articulated those concepts using upper level concepts, according to the Ontoclean methodology [8]. Modelling was firstly achieved using UML class diagrams (Unified Modelling Language). A second more enhanced specification was then implemented using Protégé [9] (a frame-based language to represent ontologies, developed by the Stanford Medical Informatics group).

2.2. Implementation of a demonstrator

The goal of the demonstrator was to achieve sharing, and processing of distributed heterogeneous data. The implementation is composed of two parts: an application (in

Java), based on a general Tomcat servlet container environment, and a mediation software called "Le Select" developed by one of the partners of the project. Upon request depending on the application, this mediator queries the data repositories using JDBC (Java Database Connectivity), an API for querying databases in SQL language. It means that appropriate data wrappers need to be implemented, in order that translate the specific file-based information into a relational structure. A second level of translation is implemented using a relational "view" mechanism, in order to translate the repository-specific entities into the common ontology entities.

3. Results

The project led to two major results: a draft domain ontology, and a demonstrator.

3.1. Draft domain ontology

The major concepts of our draft ontology concern : (1) Subjects, i.e. describing the various characteristics of healthy subjects or patients; this includes a description of the pathology and/or subject's cognitive state; (2) Exploration modalities: i.e. describing the various imaging techniques used to explore brain anatomy, structure (CT, MRI, MRA), or function (PET, fMRI, SPECT, etc.); (3) Data and data processing: i.e. representing neuroimaging data, as well as processing tools, and entities denoting actual execution of these tools.

In order to reinforce the ontology coherence, all these entities have been subsumed by a number of upper level entities (i.e. non domain-specific), such as Event, Process, Object, Artefact, Information, and Group of people. A re-engineering of this conceptual structure is in progress based on the Ontoclean methodology, and DOLCE ontology [10].
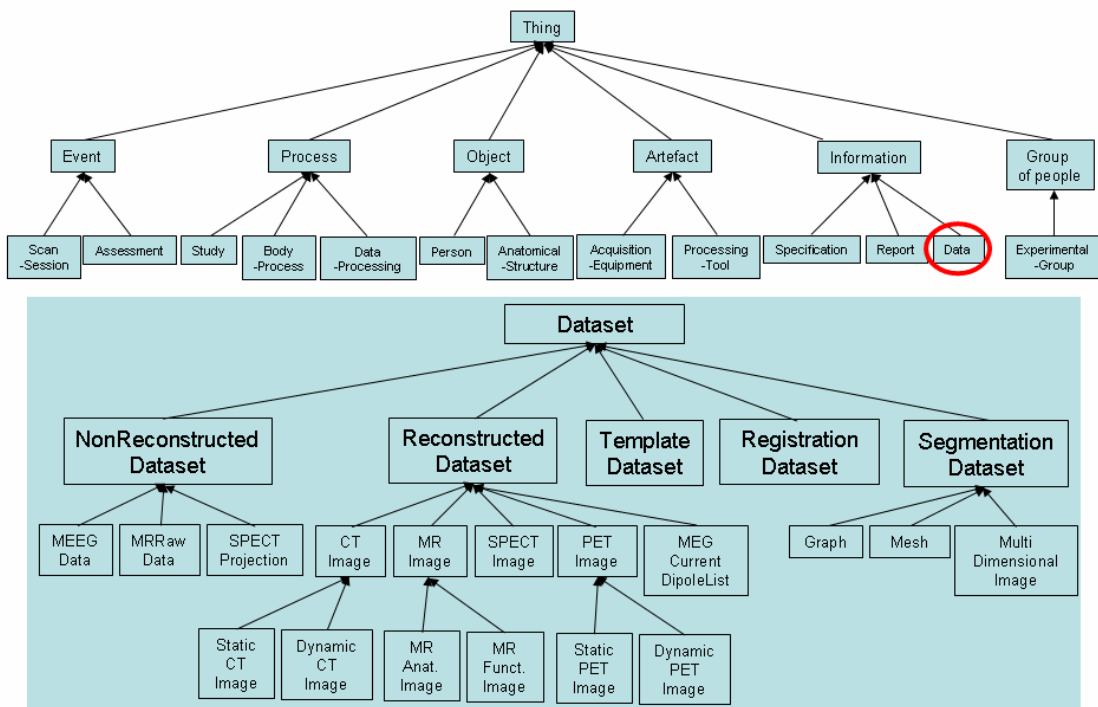

Fig. 1. A subset of our ontology: the lower part details the various kinds of Datasets.

3.2. Demonstrator

The demonstrator implemented the capability to query neuroimaging anonymized data available at four distributed repositories, using criteria about the subject, the study or the dataset. All these criteria have been defined from our draft domain ontology. The first data repository contains data about human vision obtained in both healthy subjects and patients. The data was organized as a file hierarchy, with both anatomical (original MR, segmented data showing e.g. brain, white matter and grey matter), and functional data (fMRI). The second repository contained similar data but organized according to another file hierarchy, with other conventions about file naming. The third repository was composed of a number of DICOM studies with MR, PET and SPECT images. The last one was implemented as a PostgreSQL database. It contained data about epileptic patients explored with MR and SPECT.

The demonstrator also implemented the capability to apply image processing pipelines to such data in a distributed way, thanks to a "data flow" model, in which the output of a particular processing tool can be re-directed to the input of another one, possibly located elsewhere. The image processing tools involved were as follows: image restoration, brain segmentation and brain tissue classification. Two different methods for brain segmentation and brain tissue classification were provided at different sites. The general scenario consisted in comparing the results of several pipelines, in order (1) to assess the influence of a prior restoration of MR images, (2) to compare the performance of two "brain tissue classification" tools.

It has to be noted that the demonstrator was not intended to provide a real service to users but simply to demonstrate the previous capabilities.
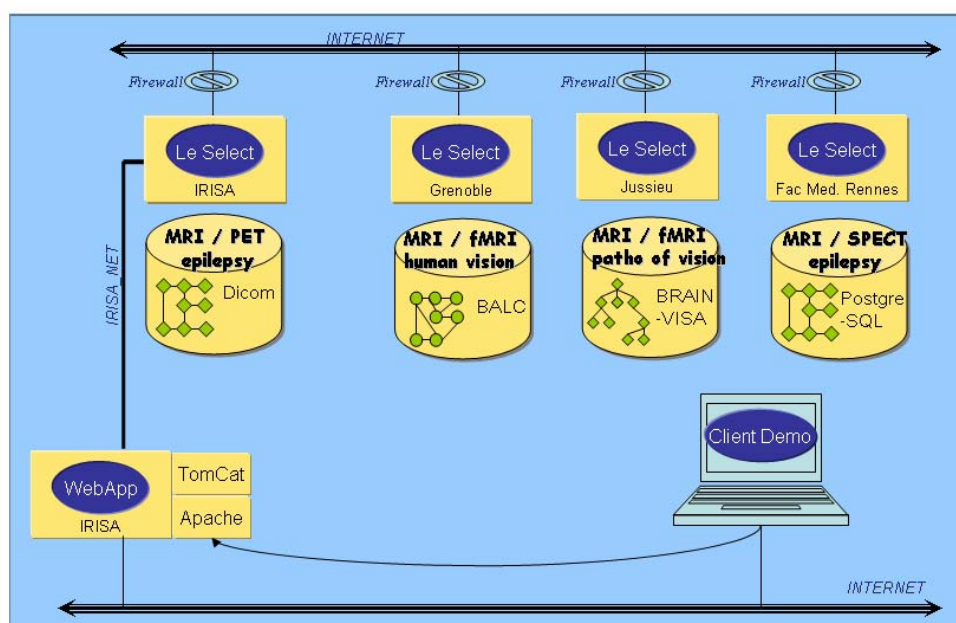

Fig. 2. Architecture of the Neurobase demonstrator: the upper part shows the four resource repositories; imaging data can be queried and image processing dataflows can be launched from a web browser, thanks to the web application.

4. Conclusion

This project allowed us to get a better insight of the issues raised by the design of federated systems to share data and processing tools in neuroimaging. Concrete results consist in a draft domain ontology and a demonstrator. Our perspectives concern extension and refinement of our draft domain ontology, so as to support interworking with other federative projects such as BIRN (Biomedical Informatics Research Network). Another perspective concerns the implementation using a GRID infrastructure to benefit from powerful computing capabilities, useful to process large masses of data, and general purpose services such as security mechanisms absolutely required for large scale deployment.

Acknowledgements

References

[1] Viewpoint: Neuroscience, Neuroimaging database, by the Governing Council of the Organization for Human Brain Mapping (OHBM), Science 2001, 292:1673-1676.

[2] Van Horn JD, Grafton ST, Rockmore D, Gazzaniga MS, Sharing neuroimaging studies of human cognition, Nature Neuroscience 2004, 7:473:481.

[3] Toga AW, Neuroimage databases: The good, the bad and the ugly, Nature Reviews/Neuroscience 2002, 3(4):302-309.

[4] Brinkley JF, Rosse C, Imaging informatics and the Human Brain Project: the role of structure, Yearbook of Medical Informatics 2002, 131-148.

[5] Kötter R, Neuroscience databases: tools for exploring brain structure-function relationships, Philosophical Transactions of the Royal Society London 2001, 356:1111-1120.

[6] Wiederhold G, Genesereth M, The conceptual basis for mediation services, IEEE Expert 1997, 12(5):38-47.

[7] Rousset MC, Bidault A, Froidevaux C, Gagliardi H, Goasdoué F, Reynaud C et al., Construction de médiateurs pour intégrer des sources d'information multiples et hétérogènes: le projet PICSEL, Revue I3 Information – Interaction – Intelligence 2002, 2(1):9-58.

[8] Guarino N, Welty C, Ontological analysis of taxonomic relationships, Proceedings of ER-2000: International Conference on Conceptual Modeling, 2000, LNCS Springer Verlag, 210-224.

[9] Noy NF, Sintek M, Decker S, Crubézy M, Fergerson RW, Musen MA, Creating semantic web contents with Protégé 2000, IEEE Intelligent Systems 2001, 16(2):60-71.

[10] Masolo C, Borgo S, Gangemi A, Guarino N and Oltramari A, The WonderWeb Library of Foundational Ontologies and the DOLCE ontology, 2003, WonderWeb (EU IST Project 2001-33052) Deliverable D18.