Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 73 (2013) 85 – 92

Procedia
Social and Behavioral Sciences

The 2nd International Conference on Integrated Information

An Integrated e-science Analysis Base for Computation Neuroscience Experiments and Analysis

Kamran Munir*, Saad Liaquat Kiani, Khawar Hasham, Richard McClatchey, Andrew Branson, Jetendr Shamdasani and the N4U Consortium

University of the West of England, Coldharbour Lane, Bristol BS16 1QY, UK

Abstract

Recent developments in data management and imaging technologies have significantly affected diagnostic and extrapolative research in the understanding of neurodegenerative diseases. However, the impact of these new technologies is largely dependent on the speed and reliability with which the medical data can be visualised, analysed and interpreted. The EU's *neuGRID for Users* (N4U) is a follow-on project to neuGRID, which aims to provide an integrated environment to carry out computational neuroscience experiments. This paper reports on the design and development of the N4U Analysis Base and related Information Services, which addresses existing research and practical challenges by offering an integrated medical data analysis environment with the necessary building blocks for neuroscientists to optimally exploit neuroscience workflows, large image datasets and algorithms in order to conduct analyses. The N4U Analysis Base enables such analyses by indexing and interlinking the neuroimaging and clinical study datasets stored on the N4U Grid infrastructure, algorithms and scientific workflow definitions along with their associated provenance information.

Selection and peer-review under responsibility of The 2nd International Conference on Integrated Information

Keywords: Computational neuroscience; Data analysis; E-science; Data integration; Scientific workflow; Information service

1. Introduction

Unprecedented growth in the availability and accessibility of imaging data of persons with brain diseases has led to the development of computational infrastructures offering scientists access to the image databases themselves and e-science services such as sophisticated image analysis algorithm workflows (also referred to as pipelines), access to powerful computational resources, and visualization and statistical tools. Scientific e-infrastructures have been and are being developed in Europe and North America offering a suite of services for computational neuroscientists and seeking convergence towards a worldwide infrastructure that can constitute the

* Corresponding author. Tel.: +44-780-9682467, *E-mail address:* kamran.munir@cern.ch

1877-0428
Selection and peer-review under responsibility of The 2nd International Conference on Integrated Information
doi:10.1016/j.sbspro.2013.02.026

foundations of a global *virtual imaging laboratory*. However, existing infrastructures such as the existing work in the EU FP7 project entitled neuGRID [1], either address the needs of a restricted user group, e.g. Alzheimer's imaging neuroscientists working with a specific dataset and application, or are targeted at a highly specialized community.

A facilitated e-science environment needs to be developed where the broader neuroscience community will find a large array of scientific resources and services. Such an environment needs to span the scientific and global challenges of sophisticated automated image analysis on databases of unprecedented size, thereby enabling further progress in the understanding and cure of neurodegenerative (NDD), white matter (WMD) and psychiatric (PSY) diseases. Our continuing work in the EU FP7 project neuGRID for Users (N4U) [2] provides an e-science environment by further developing and deploying the neuGRID infrastructure to deliver a *Virtual Laboratory* (https://neugrid4you.eu) which will offer neuroscientists access to a wide range of datasets, algorithm applications, access to computational resources, services, and support. The laboratory, whose architecture is explained in Section 2, is not only being developed for imaging neuroscientists but is also being designed to be adaptable to other user communities. This paper presents the N4U Analysis Base within the context of the N4U virtual laboratory and highlights how this work paves the way for neuroscientists to access the integrated e-science environment of computational neuroimaging, which can enhance the prospects, speed and utility of the data analysis process for neurodegenerative diseases. The N4U infrastructure is described in detail in Section 2 and the Analysis Base requirements are presented in Section 3. The Analysis base architectural details along with a use-case are presented in Section 4 that also highlights how the Analysis Base is utilized by the N4U information and analysis Services. We conclude this paper in Section 5 and also discuss our future work related to the analysis base component.

2. The N4U Infrastructure

In N4U, the development of a virtual laboratory infrastructure was identified as a major requirement in providing a facilitated e-science environment where the broader neuroscience community can find a large array of scientific resources and services. The N4U virtual laboratory, whose architecture is illustrated as Figure 1, offers neuroscientists access to a wide range of datasets, scientific pipelines, algorithm applications, computational resources and services. This virtual laboratory is mainly developed for imaging neuroscientists but designed in a way that it should remain adaptable to other user communities. Recently, a significant amount of work has been carried out in computational neuroscience research and, in particular for studies of Alzheimer's disease [3]. The main body of work in this area includes; for example, neuGRID [1], Neurolog [4], LONI [5], CBRAIN [6], and BIRN [7]. In these efforts data gathering, management and visualisation has been successfully facilitated, however the constituent data is captured and stored in large distributed databases. Such type of data management, even with the availability of dedicated and powerful software tools and hardware infrastructure, makes it unrealistic for clinical researchers to constantly review, process and then analyse dynamic and potentially huge data repositories for research. In future, it is highly likely that such data volumes and their associated complexities will continue to grow, especially due to the increasing digitization of (bio-) medical data. Therefore, users need their data to be more accessible, understandable, usable and shareable. Moreover, the danger at present is that the general lack of integration of the underpinning data infrastructures with user-defined interfacing services - coupled with the issues of unprecedented growth in data volumes and with information heterogeneity - may stifle research and delay the discovery of potential treatments of major and increasingly common diseases.

The N4U virtual laboratory has been designed following a user-led approach to provide access to the infrastructure resident data and to enable the analyses demanded by the biomedical research community. This virtual laboratory enables clinical researchers to find clinical data, pipelines, algorithm applications, statistical tools, analysis definitions and detailed interlinked provenance in a user-oriented environment. This has been

achieved by basing the N4U virtual laboratory on an integrated Analysis Database, which has been developed by following the detailed user requirements from both neuGRID and N4U projects. This integrated analysis database is entitled the "N4U Analysis Base" and its conceptual design is depicted in Figure 1. The N4U analysis base addresses the above practical challenges by offering an integrated medical data analysis environment to optimally exploit neuroscience workflows, large image datasets and algorithms to conduct scientific analyses. The high-level flow of data and analysis operations between various components of the virtual laboratory and the analysis base are also highlighted in Figure 1. The N4U analysis base enables such analysis by indexing and interlinking the neuroimaging and clinical study datasets stored on the N4U Grid infrastructure, algorithms and scientific workflow definitions along with their associated provenance information. Moreover, once the neuroscientists conduct their analysis by using this interlinked information, the analysis definitions and resulting data along with the user profiles are also made available in the analysis base for tracking and reusability purposes.

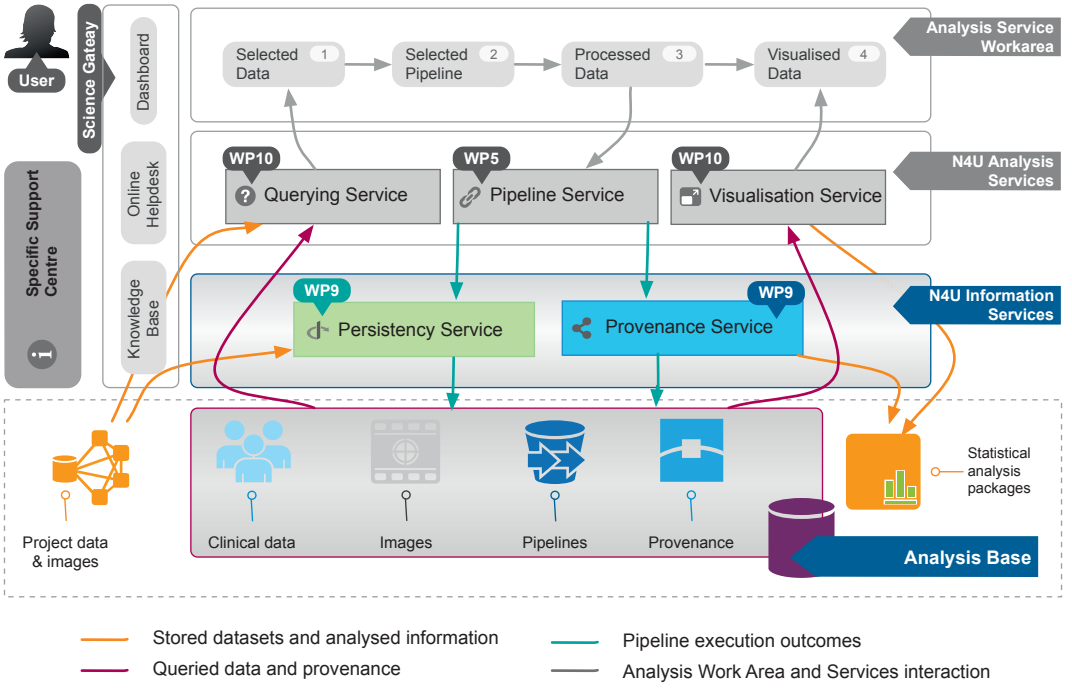

Figure 1: The detailed N4U Virtual Laboratory Infrastructure Powered by the N4U Analysis Base

Building on the analysis base, the N4U virtual laboratory provides the environment for users to conduct their analyses on sets of images and associated clinical data (as shown in Figure 1). In-addition to the analysis base, the N4U virtual laboratory comprises the following major components (1) Information Services; (2) Analysis Services; (3) The Analysis Service Work Area; and (4) Science Gateway and Specific Support Centre. The Information Services (see Figure 1) comprise the Persistency and Provenance services within the N4U virtual laboratory. Here, the persistency service complements the analysis base by acting as a wrapper for information storage. The persistency service provides appropriate interfaces for storing the meta-data of datasets e.g. ADNI (Mueller et. l., 2005) into the analysis base such that these datasets, which are actually stored in their entirety on the N4U grid infrastructure (or other similar repositories), become indexed in the analysis base. The Analysis Service within the N4U virtual laboratory provides access to tracked information (images, pipelines and analysis outcomes) for querying/browsing, visualisation, pipeline authoring and execution. The Analysis Service Work Area is a facility for users to define new pipelines or configure existing pipelines to be run against selected

datasets and dispatch to conduct analysis. Finally, the Science Gateway in the N4U virtual laboratory provides facilities that include a Dashboard (user interface), an Online Help Desk and several Service interfaces for users to interact with the underlying set of N4U services. The Online Help Desk is a one-stop assistance facility with which the user can interact to learn about the N4U virtual laboratory and how to interact with it via the dashboard.

3. The N4U Analysis Base Requirement

The N4U analysis base should fully support different components and services of the above-mentioned N4U Virtual Levorotary by having the capability of for example (a) indexing all external clinical datasets (b) registering neuroscience pipeline definitions and/or associated algorithms (c) storing provenance and user-derived data resulting from pipeline executions on the Grid (d) providing access to all datasets stored on the gird Infrastructure and (e) storing users' analysis definitions and linking them with the existing pipelines and datasets definitions. In order to meet these requirements, specific data structures and software interfaces to access and manipulate these data structures had to be designed and implemented. While the modelling of the analysis base will be described in the next section, here we provide a discussion on some of the major requirements, their associated dependability requirements and some of important design decisions that we had to take for the implementation of analysis base.

3.1. Indexing of External Clinical Datasets

In the N4U setup, the datasets are stored in the distributed Grid infrastructure. In order to make this data available to the users (e.g. neuro- and citizen scientists) of different N4U services, a mechanism is required that is usable by the software components, which will eventually also be used by human users. Here, providing a direct access to these datasets is not ideal, because a user - who may be preparing to carry out an analysis through the Analysis Services - may need to select part of the dataset based on certain characteristics. Filtering giga-bytes of data at runtime is a non-trivial task, unless it is methodically indexed beforehand.

3.2. Register Neuroscience Pipeline Definitions and/or Associated Algorithms

Neuro- and citizen-scientists will use the N4U virtual laboratory to execute complete pipelines or algorithms over different datasets stored in the Grid infrastructure. Therefore, the N4U virtual laboratory is required to provide a list of pipelines (and algorithms) that are executable on the N4U grid infrastructure. Moreover, these pipelines may have constraints such as being restricted to certain datasets (or formats) as inputs, algorithms that can be employed within those pipelines, etc. These constraints and other characteristics are specified in pipeline definitions. Therefore the analysis base is required to index the pipelines already registered (or when they become available) in the N4U infrastructure. Consequently, for completeness, other entities such as algorithms and toolkits related to such pipelines are also indexed in the analysis base.

3.3. Storing Provenance and User-derived Data Resulting from Pipeline Executions

Any provenance information generated as a result of some user analysis needs to be stored for cross-analysis, verification and comparison purposes. This provenance information should contain references to the data set items that were used in the analysis, the algorithms employed, pipeline execution and outcome details, user information, etc. Use-cases such as these required not only the data sets, pipeline and algorithm definitions to be indexed with respect to their contents and properties, but also required the storage of provenance information linked to these entities. Therefore the analysis base provides a mechanism for storage of the provenance

information along with relationships/references of datasets and any data derived from the original data sets as a result of user analysis.

3.4. Exporting Datasets to the Analysis Base:

In order to index the datasets stored in the N4U infrastructure, the datasets are required to be exported into the analysis base. Exporting the whole data sets into the analysis base is not feasible because the datasets are (and will be) prohibitively large in size and hence access. To meet the objectives of the analysis base, only indexes to the data sets need to exist inside the Analysis Base. Creating these indexes is not straight forward, especially in the absence of any metadata associated with the datasets and differences in the format of the various datasets being considered in the N4U project. By incorporating a metadata definition and mapping approach within the Persistency Service, discussed later in this paper, the indexes to the datasets are stored in the analysis base.

4. The N4U Analysis Base

The N4U Analysis Base is an information catalogue for the users and a central repository for the Analysis and Querying services to perform search queries for different datasets, algorithms and pipelines in the N4U. Users who are part of the N4U community are able to index Pipelines, Algorithms and Data Sets, which are stored by the Persistency Service, within the analysis base. Only the users who are marked in the system as active are able to perform these actions. The schema has been designed to facilitate enabling or disabling a user in the analysis base.

Another objective for the analysis base was to store provenance information related to the pipeline and its analysis, data or image files generated as a result of a pipeline execution. This provenance information aids the N4U user community in reviewing the execution of the pipelines and confirming the outcome produced as a result of their analysis. It also provides a mechanism to the users in exploring the problems with their pipelines or the given input, in case of failures. The analysis base has been designed in a way that it separates the pipeline, algorithm specifications with the pipeline, algorithm execution related information, respectively. Thus, this enabled rich provenance querying, including queries about the general structure of the pipelines, the activities in the pipelines, the links between different pipeline activities and activity results. Moreover, any analysis performed on the N4U datasets is also linked with the provenance of those analyses for tracking and re-usability purposes. In the following sections, we describe the interaction of these N4U virtual laboratory persistency and provenance services with the analysis base.

4.1. N4U Analysis Base Schema

In order to map the requirements (discussed earlier in Section 0), the N4U analysis base schema has been designed, a simplified version of which is shown as **Error! Reference source not found.**. This schema is composed of five main entities, which include User, Pipeline, Algorithm, Analysis, and DataSet and are briefly described as follows:

- User: In order to provide ownership of activities and data in the N4U, user information is maintained in the N4U analysis base.
- Pipeline: A pipeline (or workflow) authored by a user in a workflow-authoring tool such as LONI. As discussed earlier in Section 3, pipeline definitions are indexed in the analysis base.
- Algorithm: Each pipeline is composed of different algorithms, which can be thought of as tasks, to perform different types of processing on the neuro-images in a neuroscience analysis.
- Analysis: Execution of a pipeline by providing the required input values (or datasets) is termed conducting an analysis in the N4U terminology.

E DataSet: The files used as inputs for analysis or produced as outputs during an analysis are termed a dataset. A dataset can have a single file or a set of files.

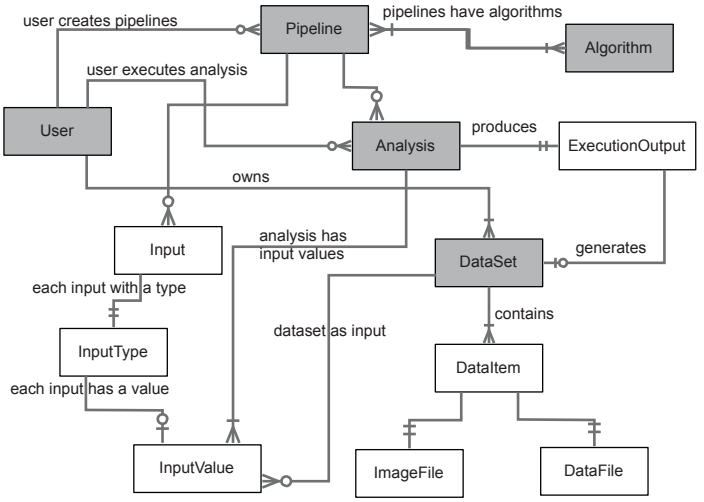

Figure 2 A simplified schema of the N4U analysis base database

4.2. Persistency Service

The Persistency Service provides appropriate interfaces for storing the meta-data of data sets (e.g. the ADNI data sets) into the analysis base such that those data sets, which are actually stored in their entirety on the N4U Grid infrastructure, become indexed in the analysis base. This index can then be used by other services for allowing users to define their analysis and carry out relevant queries using these data sets. The operation of the Persistency Service and its interaction with the analysis base is shown in *Figure 3:* .

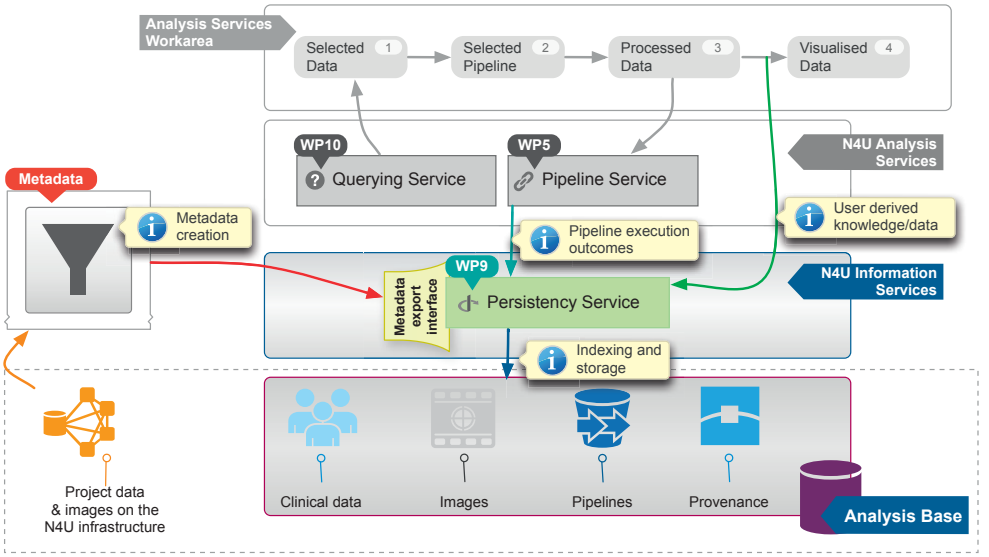

Figure 3: The N4U Persistency Service and its interaction with the Analysis Base

Exporting the whole data sets into the analysis base is not feasible because the data sets will be prohibitively large in size. To meet the objectives of the analysis base, only indexes to the data sets need to exist inside the analysis base. In the case of the ADNI dataset, the contents are organised in multi-level folder structures that contain outputs of different clinical studies at different stages. Each stage contains dozens of image outputs that themselves are organised in multi-level directory structures and associated clinical subject and study data in XML format. There is no formal meta-data associated with the data set as a whole. To automate the process of indexing such a data set, and to accommodate any changes to the original data source, it is essential to devise a metadata structure that is uniform across various revisions of the data set. In the absence of such a metadata format in the ADNI dataset, we have devised one. This metadata format is specified in an XML Schema Definition (XSD) that models items within a data set as a (related) collection of images and data files (files containing study and subject information). The metadata of the ADNI data set is generated through a software programme, entitled the DatasetCrawler, which browses (crawls) through the data set directories (on the Grid infrastructure) and records the structure, names, properties and URLs to the files included in the data set contents. The DatasetCrawler then generates metadata XML files that conform to the metadata schema. These metadata files are then exported to the Persistency Service, which then indexes the metadata in the analysis base. Thus the Persistency Service complements the N4U analysis base, acting as a wrapper for storing information into the analysis base and retrieving such information.

4.3. Provenance Service

Simply creating and executing pipelines is not enough on its own, it is important that results, as and when required, should be reproduced and reconstructed using past information. The Provenance Service keeps track of the origins of the data and its evolution between different stages and services. The provenance service allows users to query analysis information, to regenerate analysis workflows, to detect errors and unusual behaviours in past analyses and to validate analyses. The service support and enable the continuous fine-tuning and refinement of the pipelines in the N4U project by capturing (1) pipeline specifications; (2) data or inputs supplied to each pipeline component; (3) annotations added to the pipeline; and (4) execution outputs or errors generated during analysis.

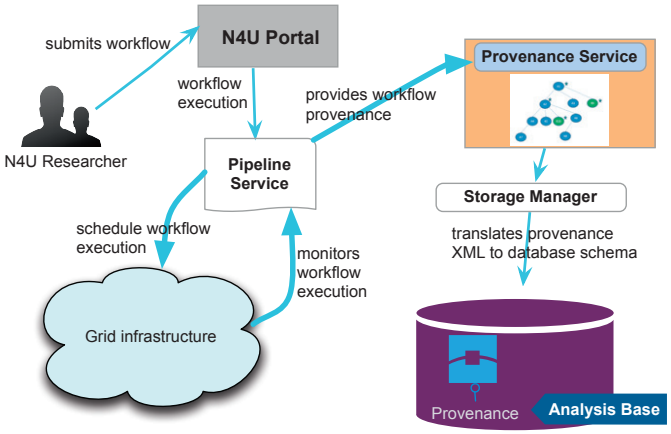

Figure 4: Detail of provenance capturing in the N4U Virtual Laboratory

Figure 4 shows an overview of provenance capturing in the N4U virtual laboratory. When a user submits a pipeline through the N4U Portal for execution, it is called an analysis. The submitted workflow defines the flow and inter-dependencies of activities i.e. algorithms and their inputs. Upon receiving the workflow, it is then

passed onto the Pipeline service. The Pipeline service is responsible for parsing the submitted workflow for consistency checks, and for identifying suitable resources on the Grid for its execution. The Pipeline service creates a scheduling plan for the workflow and schedules the workflow to the available suitable resources on the Grid. This service is also responsible for periodically monitoring the execution status of the submitted workflow. In the case where a workflow fails during execution due to any runtime problems, this service will react to this, and reschedule the workflow to some other Grid resource. Along with the status monitoring, the pipeline service retrieves the execution output along-with other details and update the provenance service about this. The Provenance service exposes interfaces to record provenance information of a pipeline and record it along with the output and error logs in analysis base via Storage Manager.

5. Conclusion

This paper has presented the design and development of the N4U Analysis Base within the N4U Virtual Laboratory, which has been designed following a user-led approach to provide access to the Grid infrastructure resident data and to enable the data analyses demanded by the biomedical research community. The N4U Analysis Base enables such analyses by indexing and interlinking the neuroimaging and clinical study datasets stored on the N4U Grid infrastructure, algorithms and scientific workflow definitions along with their associated provenance information. Moreover, once the neuroscientists have conducted their analyses by using this interlinked information, the analysis definitions and resulting data, along with the user profiles, are also made available to the registered scientific community for tracking and reusability purposes. This work has paved the way for neuroscientists to access the integrated e-science environment of computational neuroimaging, which has dramatically enhanced the prospects, speed and utility of the data analysis process for neurodegenerative diseases. Our future efforts geared towards indexing a larger number of clinical datasets and providing interactive visualisation interfaces for the users of the N4U virtual laboratory.

Acknowledgements

This work is funded by the EU 7th Framework Programme under the N4U project (reference 283562).

References

[1] Redolfi, A., McClatchey, R. et al., (2009). Grid infrastructures for computational neuroscience: the neuGRID example. Future Neurology, November 2009, 4(6), (pp. 703-722), DOI 10.2217/fnl.09.53

[2] Frisoni, G. B., et al., (2012), N4U: expansion of neuGRID services and outreach to new user communities (poster). 9th e-Infrastructure Concertation Meeting of the European Grid Infrastructure, 22 Sept. 2011, URL: https://neugrid4you.eu/conferences

[3] Mueller, S.G. and Weiner, M.W. and Thal, L.J. and Petersen, R.C. and Jack, C. and Jagust, W. and Trojanowski, J.Q. and Toga, A.W. and Beckett, L., The Alzheimer's disease neuroimaging initiative, Neuroimaging Clinics of North America, vol 15, no 4, pg. 869, 2005, NIH Public Access.

[4] Wali, Bacem and Gibaud, Bernard, 2012, Semantic annotation of image processing tools, Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics, ISBN 978-1-4503-0915-8

[5] Dinov Ivo et al. 2009, Efficient, distributed and interactive neuroimaging data analysis using the LONI pipeline, Journal of Frontiers in Neuroinformatics, Vol 3(22).

[6] CBRAIN Project, http://cbrain.mcgill.ca/. Last accessed 25th August 2012.

[7] Grethe JS. et al., 2005, BIRN: Biomedical informatics research network: building a national collaboratory to hasten the derivation of new understanding and treatment of disease. Stud Health Technol Inform. 2005;112:100-9. PMID: 15923720.