NeuroLOG ANR-06-TLOG-024
Software technologies for integration of process, data and knowledge in medical imaging

Financé par ANR

Application pipeline specifications

Deliverable L5

Authors : D. Alech, M. Dojat, A. Gaignard, B. Gibaud, D. Godard, S. Kinkingnéhun, Y. Lévy, F. Michel, J. Montagnat, M. Pélégrini-Issac, X. Pennec, E. Pernod, J. Rojas Balderrama

Summary : This deliverable describes the NeuroLOG application pipelines and the security infrastructure set up to support the applications deployment on the NeuroLOG platform.

Document layout

1. Application pipelines

This section describes the NeuroLOG applications data analysis pipelines. The pipelines are both described informally and formally (using the Scufl data flow language). The formal Scufl documents can be found on the project wiki, in the applications section.

For each of the four applications, the clinical context and objectives of the data analysis pipeline is first introduced. The overall pipeline is then described. Individual processing tools involved are identified and their main properties are outlined. The pipeline results are finally described.

1.1. Multiple Sclerosis application

1.1.1. Clinical context

Magnetic Resonance Imaging (MRI) detects with high sensitivity white matter lesions (WML) in patients with Multiple Sclerosis (MS). Over the last 25 years, it has been increasingly used for diagnosis, prognosis and as a surrogate marker in MS trials. Conventional Magnetic Resonance (MR) sequences for MS include pre- and post-gadolinium (gd) T1-weighted (T1-w), T2-weighted (T2-w), proton density (PD) or FLuid Attenuating Inversion Recovery (FLAIR). These sequences have been developed to optimize the detection of the lesions in the white matter (WM) [4].

In cross-sectional and longitudinal studies, manual segmentation has been used to compute the total lesion load (TLL) in T2-w, PD-w, unenhanced and gd-enhanced T1-w MR sequences but this method is very time consuming and has large intra- and inter-operator variability [5]. Semi-automatic methods tend to reduce this variability, but there is a great promise that automatic methods will improve considerably lesion segmentation reproducibility, which is of critical importance when processing huge amounts of MR images, as in large multicenter clinical trials.

Automatic segmentation of WML is a complex task that requires considerable preprocessing of MRI data [6]. Our assumption is that WML segmentation methods significantly depend on the preprocessing tasks, most notably: registration, skull stripping, image denoising and intensity inhomogeneity correction.

Our purpose with the sharing of our lesion segmentation pipeline is two-fold: (1) to provide a method that makes the best use of multi-parametric data to obtain a robust and accurate segmentation of tissues and MS lesions, and (2) to investigate how the different preprocessing procedures impact on the segmentation results, especially in terms of robustness and accuracy.

1.1.2. Asclepios application pipeline description

This MS lesion segmentation algorithm has been developed by Dugas *et al.*. First, brain MRIs are normalized (spatially and in intensity) and skull-stripped. Then, a segmentation of the brain into the different healthy compartments

classes (white matter (WM), gray matter (GM), cerebro-spinal fluid (CSF)) is realized thanks to an expectation maximization algorithm. The segmentation captures Partial Volume Effects (PVE) at the boundaries of compartments to refine the precision of results. The segmented classes are used to segment lesion on the T2-FLAIR sequence. The expectation-maximization algorithm consists in iterating two steps: labeling of the image (Expectation step) and estimation of the Gaussian class parameters (Maximization step).

1.1.3. Processing tools

The application workflow involves the following services. A detailed description of these services aiming at ontologically describing them can be found on the NeuroLOG wiki web site.

- **Asclepios resampling service**: Resamples an image according to a transformation based on a Registration transformation (4x4 matrix, or a dense deformation field). Uses either the same sampling grid as input image, or a new one, through the specification of the new image geometry.
- **Asclepios conversion service**: Convert an image from a format to another without modifying the data content.
- **Asclepios Baladin registration service**: Registration (Rigid or non-rigid affine registration) to spatially align different modality images. A rigid registration procedure is used in this workflow. (The associated Processing Categories in the ontology are: Rigid registration; Affine normalization; Multi-modality affine coregistration; Mono-modality affine registration).
- **Asclepios EMBrainMask service**: Segmentation of brain mask, based on a priori information on grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF) coming from an atlas.
- **Asclepios EMBrainClassification service**: Segmentation of tissues, based on EM classification, and on a priori masks concerning GM, WM and CSF coming from an atlas. This algorithm takes into account extra classes due to partial volume effect (PVE) between grey matter and CSF.
- **Asclepios Binarisation service**: Create binary segmentation datasets from probabilistic segmentation datasets, based on the highest probability found at each voxel in the 4 images; creates one additional segmentation dataset containing all non-classified voxels.
- **Asclepios Unbias service**: Estimation of bias field and calculation of debiased image using WM and GM binary segmentation.
- **Zpar service**: Read the header of Inrimage file to get image size, voxel image size and type of data informations.

1.1.4. Results produced

The MS pipeline is used as a pilot workflow for porting on the EGEE infrastructure. Currently, the workflow has been deployed on the grid infrastructure, only until the Brain segmentation into the different healthy compartments classes. Figure 1 displays an example of the segmentation obtained by the execution of the current workflow on the EGEE grid.

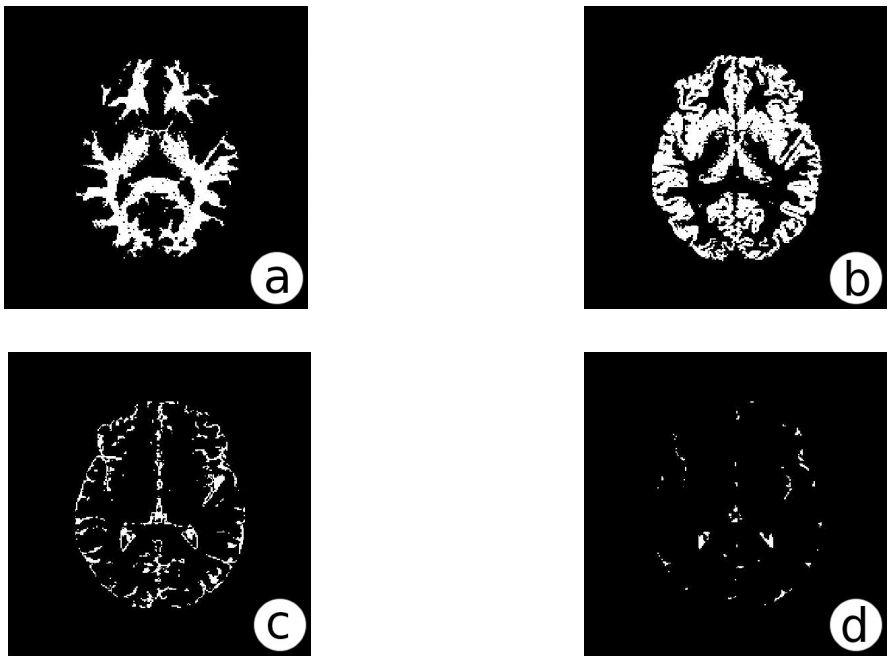

Figure 1. Binary segmentation from the workflow on the EGEE grid : a) White matter, b) Grey matter, c) CSF, d) PVE.

Time performance :

Figure 2 shows the difference of execution time between local and grid executions of the workflow for a variable number of patient images to process. It is important to underline that we are still working on this workflow. Some known optimizations still need to be implemented. The grid introduces an overhead that makes it non competitive when processing 3 patients or less. For more patients, the grid parallelism outperforms the local execution. The grid computing time is not constant due to variable grid load: the blue curve is a mean value over multiple experiments for each point. Error bars show the standard deviation.

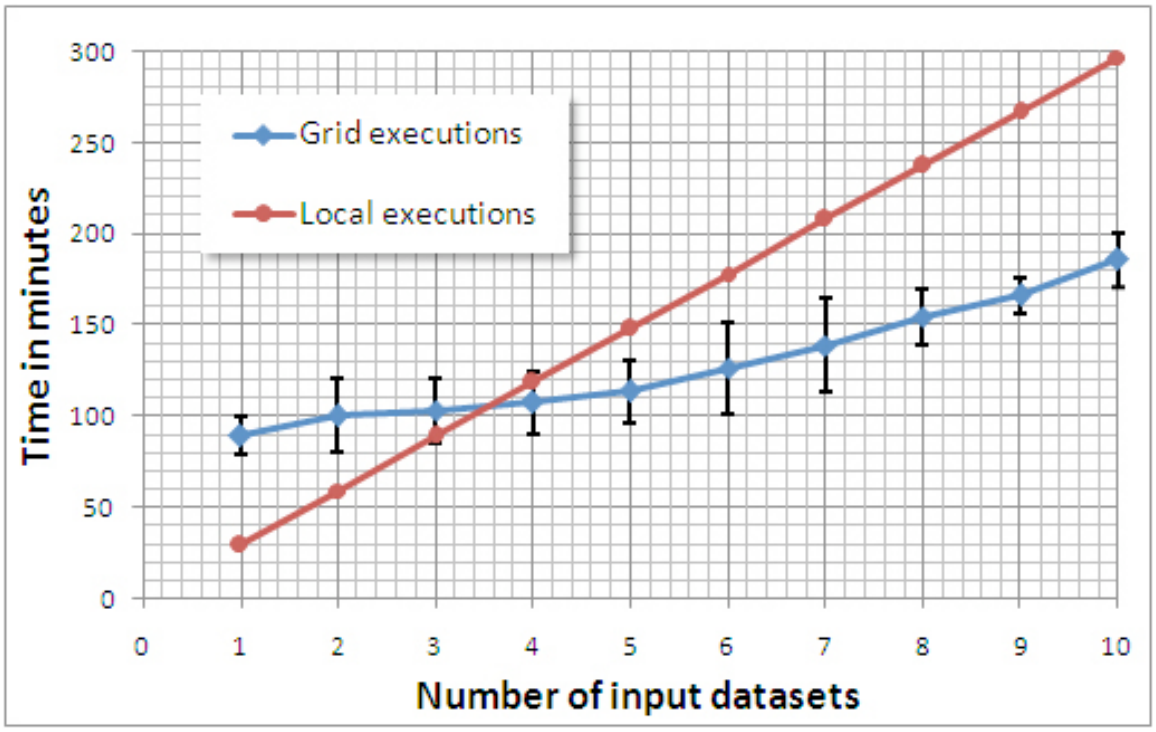

Figure 2. Local versus Grid execution time for a variable number of patient images.

1.1.5. Scufl representation

The workflow as it appears in the workflow manager interface is shown in Figure 3. The input data sources (light blue triangle) are patient images (T1, T2 and Proton Density MRIs), atlases or algorithm parameter values. Each bow corresponds to the invocation of one service described above (baladin registration, resampling (reech3d), skull stripping, segmentation, etc). The workflow outputs (light blue diamonds) are the segmented images for each class (WM, GM, CSF and Partial Volume Effect).

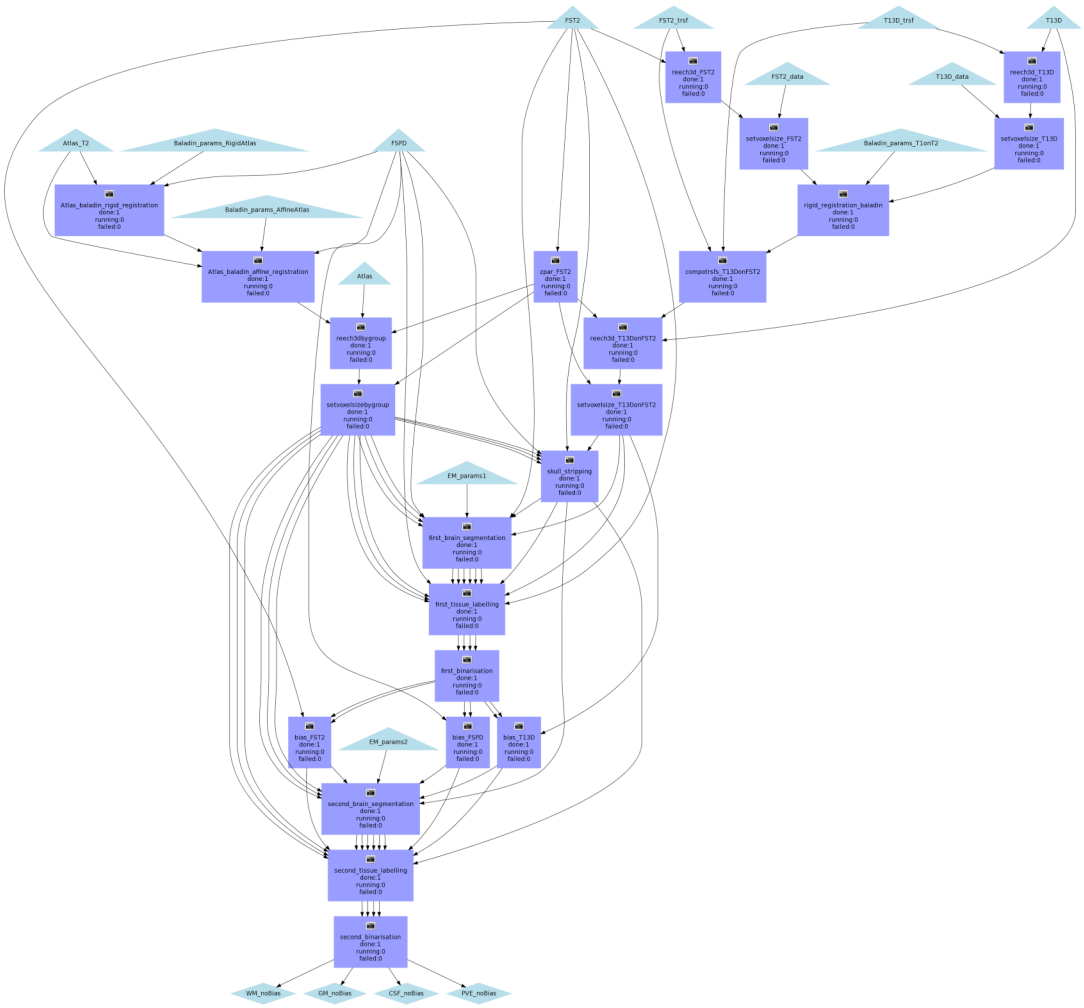

Figure 3. Scufl workflow in MOTEUR interface.

1.1.6. IRISA application pipeline description

The pipeline shown in Figure 4 aims at classifying tissues and lesions from multi-parametric MR data. The inputs of the pipeline are T1-w, T2-w, PD, gdT1 and FLAIR longitudinal image sequences. Several preprocessing are applied to each image, namely bias correction (which may be quite important using 3 Tesla scanners), denoising and intensity normalization. The images are registered into a common geometric reference system (T1 image at t0) and skull-stripped. All image modalities are then used in the segmentation procedure. The results are:

- A binary mask of the tissues: CSF, grey matter, white matter
- A binary mask of lesions appearing as T2 hyperintensities
- A binary mask of lesions appearing as black holes in T1 images
- A binary mask of lesions which are gd-enhanced.

Note: this algorithm of segmentation uses cross-sectional data yet, but a later version is planned that will be able to process longitudinal data as well.

1.1.7. Processing tools

The pipeline involves six processing tools:

- **Visages_registration_service**: This algorithm was developed in-house by Nicolas Wiest-Daesslé [5]; it register linearly one image onto another.
- **Visages_STREM_service**: STREM stands for Spatio Temporal Robust Expectation Maximization. This algorithm was developed in-house by Daniel Garcia-Lorenzo [6], based of previous work of L. Ait-Ali, S. Prima et al. [7]. It provides a segmentation of three tissues (White matter, Grey matter and Cerebrospinal fluid) and MS lesions (Black Holes, T2 Hyperintensities, Gadolinium enhancing lesions).
- **Visages_VipT1BiasCorrection_service**: This bias correction method was developed by J.F. Mangin and it is part of the Brainvisa software package. It computes a smooth multiplicative field which corrects for non stationarities. This field aims at minimizing the volume entropy.
- **Visages_Bet_service**: The well-known segmentation method (Brain Extraction Tool) developed by S.M. Smith. It is part of the FSL Library (FMRIB Software Library).
- **Visages_NLMeans_service**: An in-house developed algorithm (by Pierrick Coupé) for denoising, based on the Non-Local Means (NLM) algorithm [8], which has shown better performance than other state-of-the-art methods.
- **Visages_intensity_normalisation_service**: Not described, yet (to be done).

A detailed description of these services aiming at ontologically describing them can be found on the NeuroLOG wiki web site.

1.1.8. Results produced

The previous methods are being used in the IRISA/Visages environment (especially by D. Garcia's in the context of his PhD work).

The pipeline has been modeled, as well as its individual components. These components were modeled as individual services using the description form defined in WP2. A first version of the corresponding workflow was then modeled (in Scufl language, used in the MOTEUR implementation) from these descriptions.

1.1.9. Scufl representation

The Scufl data flow shown in Figure 4 exhibits a similar pre-processing chain for each modality (T1-w, T2-w, PD, gdT1 and FLAIR). The modality sources contain temporal series of acquisition from a single patient. After pre-processing, all images are registered on the reference image (T1-w image acquired at time

0). All modalities are used simultaneously for the final segmentation step. One segmentation result is produced for each time instant.

The current Scufl workflow can only handle a single patient at a time. Extensions of the Scufl language that are studied in the context of another ANR project (GWENDIA, ANR-06-MDCA-009) will enable manipulation of images from different patients without confusion in the future.

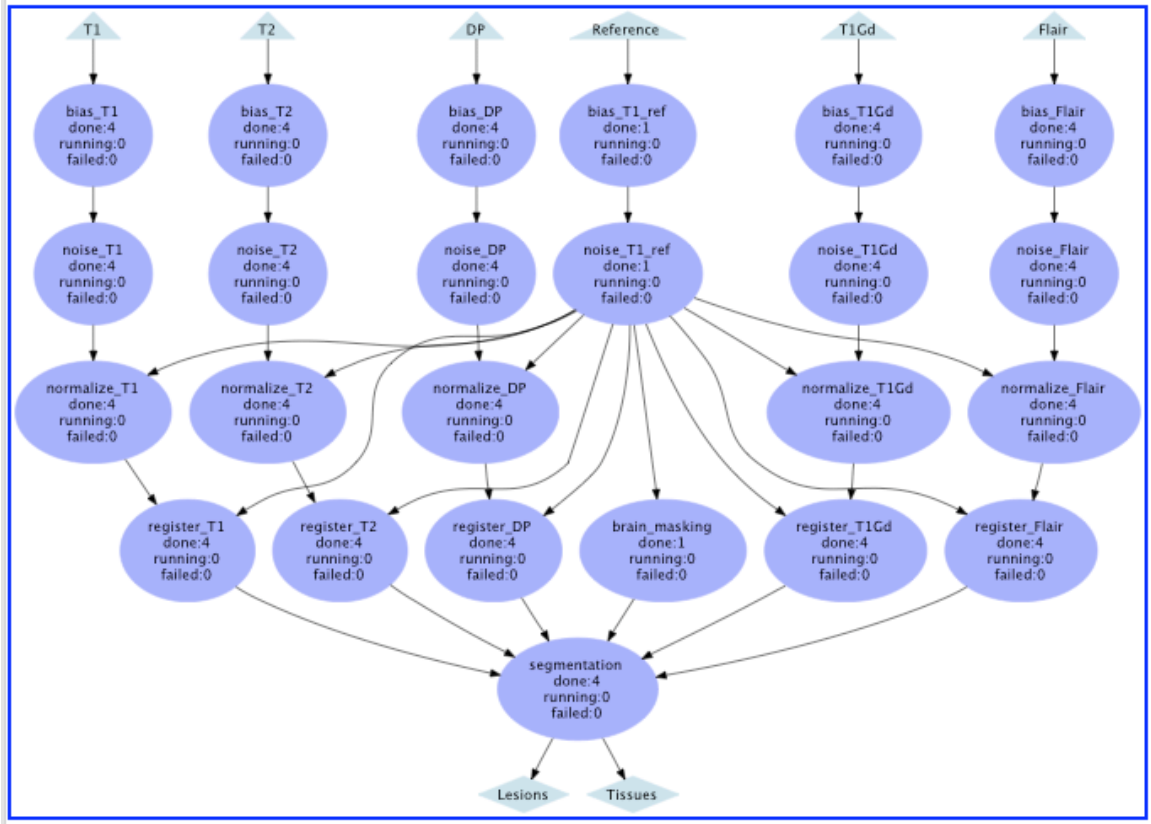

Figure 4. Scufl workflow in MOTEUR interface.

1.2. Stroke application (GIN)

This application goal is to segment lesions following an acute stroke event and access their evolution in term of size and position. The subject is followed up at three times, e0, e1, and e2 (about 3 hours, 3-4 days and more than 30 days after the stroke event). Three types of MRI are used: T2, FLAIR and Diffusion weighted imaging. The relevance of information contained in each image depends of the time concerned (for instance the lesion is well defined in Diffusion weighted image at e0 and in FLAIR at e2). For segmentation, two algorithms are available at GIN: 1) segmentation of each image separately and 2) conjoint segmentation of the three images (under development). For each processing steps tools are available at each NeuroLOG partner and will be tested in order to define the optimal processing pipeline.

The current clinical protocol used is named **Virage** and was approved by our local ethical committee. It is routinely used at the Michallon Hospital in Grenoble (FR) in the neurology unit managed by Pr Marc Hommel and under the supervision of Dr Assia Jaillard. Images are acquired on a Philips Intera 1.5T. Images are available in the GIN lab on CD in the **DICOM format**. Philips format (.par and .rec) files could be stored if needed). Presently, no local database is available. The construction of such a database is under study in interaction with partners of this project. For processing image data are in NIFTI or ANALYZE (SPM) format.

1.2.1. Processing tools

- Spatial images realignment. T2 at e0 is considered as the reference image. Flair and Diffusion are firstly realigned onto the corresponding T2 (rigid realignment) for each time considered. The two T2 are then realigned onto the reference T2. The transformation matrices obtained at these two steps are then composed. Images are not resampled at this stage. We use the SPM2 realignment algorithm. Algorithms available at NeuroLOG partners can be tested (Inputs: target and source images. Outputs: transformation matrices composed with the initial matrix).

- Temporal sequences realignment onto a template. T2 at e0 is firstly realigned on a template T2 (non rigid registration). Realignment parameters are then applied (composition with the output of the previous step) to all images. We use the SPM2 realignment algorithm. Algorithms available in NeuroLOG partners can be tested (Inputs: template, source image, and the list of images to transform Outputs: transformation matrices composed with the initial matrix for all images.

- 8 bits Conversion. 12 bits or 16 bits images are transformed to 8 bits images (GIN binary available). (Input: image to transform. Output: transformed image, histogram visualization to check the validity of the threshold used above which grey level intensities are set to 255).

- Skull-stripping. Brain is extracted onto T2 à e0 (Input: image to process, Output: brain image and binary brain mask). We used Bet (binary) from FSL library. The binary mask is then applied (multiplication) for brain extraction to all images and all times

- Lesion segmentation. We use LOCUS for all images separately or to all three images conjointly for each time (under development see SELMIC project http://r2-d2.ujf-grenoble.fr/selmic/doku.php). (Input: image(s) to segment Output: 3 tissues images + lesion, + debiaised image).

1.2.2. Results produced

For each step, output images should be visualised by the end user. The final result is the lesion labelling (see Figure 5, right), position and volume. Automatic lesion segmentations can be refined by user manual intervention. This requires the simultaneous visualisation (superposition by transparency) of initial and labelled images.

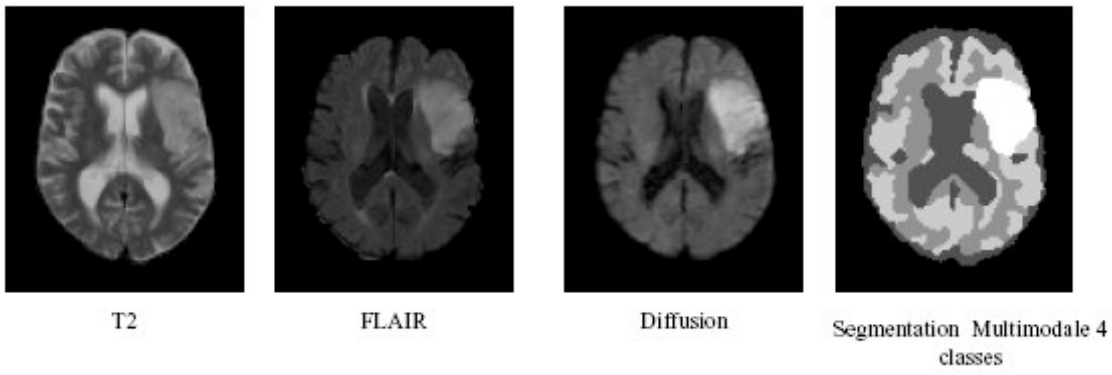

Figure 5. Input images (T2, diffusion and FLAIR) and output segmentation example.

1.2.3. Scufl representation

The Scufl data flow displayed in Figure 6 shows the spatial realignment of FLAIR and diffusion images onto the T2 image. The T2 temporal sequence of images is aligned onto the reference (T2 at e0) in the temporal registration step. Spatial and temporal registration matrices are multiplied to provide the final transformation aligning each modality at each time instant onto the reference image. The images are then resampled in this common frame, debiased and segmented.

The current Scufl workflow can only handle a single patient at a time. Extensions of the Scufl language that are studied in the context of another ANR project (GWENDIA, ANR-06-MDCA-009) will enable manipulation of images from different patients without confusion in the future.

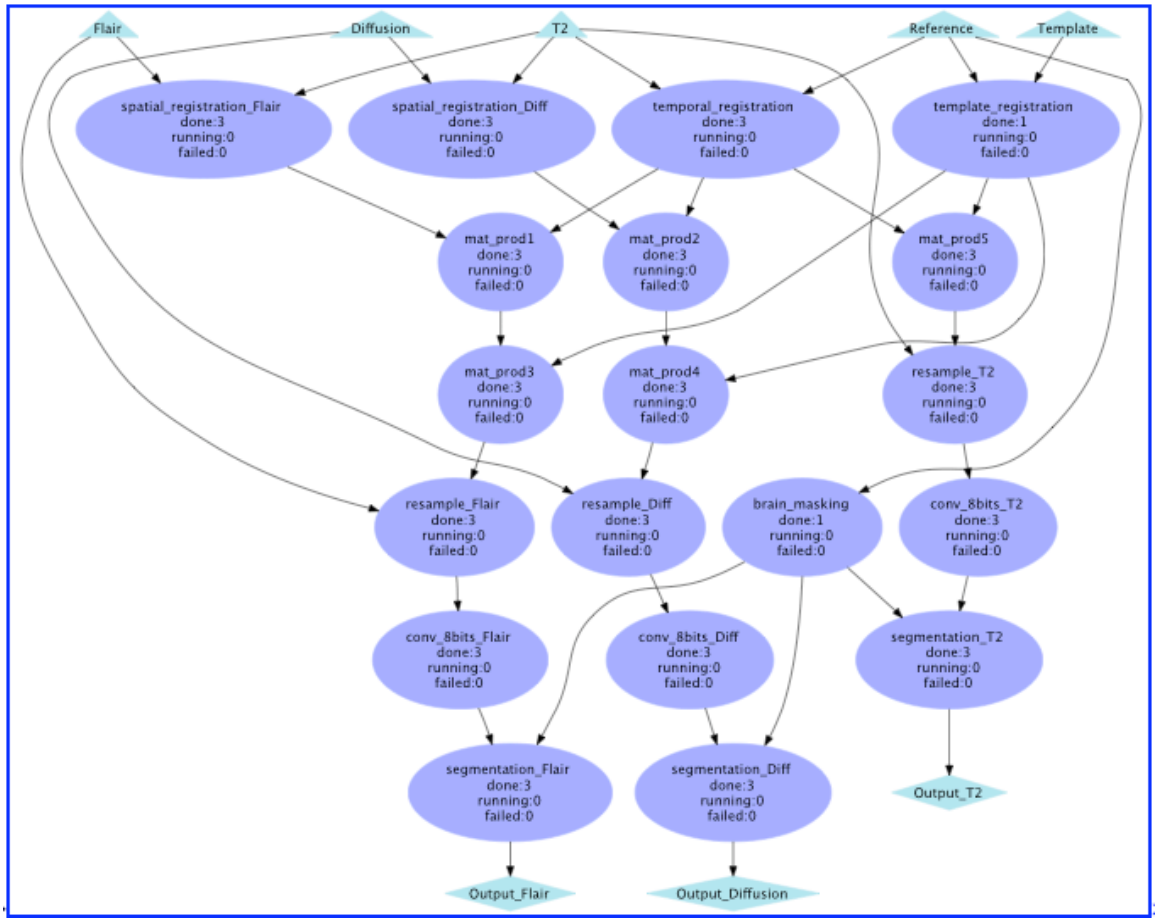

Figure 6. Scufl workflow in MOTEUR interface.

1.3. Stroke / tumours application (IFR49)

The aim of this so called *AnaCOM* method is to create anatomo-clinical maps from patients with brain lesions. These maps are created using the anatomical MRI of patients who are administered neuropsychological tests.

Beside the scientific interest of a better knowledge of the brain, these maps have a clear medical interest. Actually, these anatomo-clinical maps are designed to locate the specific brain loci that are linked to a specific function. Therefore we will be able to build up or complete brain anatomo-clinal atlases, which will be used by neurosurgeons to plan their interventions for tumor resections, and neurologists to understand which functions are impaired by a stroke and the probability for a patient to recover following a specific re-education.

1.3.1. Application pipeline description

The method consists of two independent but complementary steps:

- **preprocessing**: individual MRIs are registered to a common standardized space. Brain lesions are segmented to obtain regions of interest (ROIs).
- **statistical analysis**: the score obtained by a given patient in the neuropsychological test is introduced in the voxels of the segmented ROI of this patient. Lesions are overlapped and voxels are gathered in clusters according to the patients brain lesions which cover them. The patient's score for each cluster is statistically compared to values obtained for control subjects, which allows for the creation of statistical maps.

A) Preprocessing: registration and segmentation step

- **Image Acquisition**: The MR images required for AnaCOM studies are anatomical scans on which lesions boundaries are clearly visible. 3D T1 anatomical MRI allows one to see both vascular and tumor excision lesions.

  Image file type: Analyze (.img/.hdr).
- **Artefact Removal**: A coarse binary mask is created to exclude signal abnormalities during registration. "Abnormalities" consist of the brain lesions as well as artifacts. The resulting mask is composed of two values: 0 in the voxels with an abnormal signal value and 1 in the others. The mask can be larger than the actual lesions. This operation is still manual.

  Processing tool: MRICro (cf. 1.4.2)
- **MRI Registration**: MRIs are all registered to the same template, according to the T1 Montreal Neurological Institute (MNI) stereotactic space. This registration is performed using first an estimation of the optimum 12-parameter affine transformation to match images (Ashburner et al., 1997). Then, a second step accounts for global non-linear shape differences, which are modeled by linear combination of smooth spatial basis functions.

  Processing tool: SPM2 (cf. 1.4.2)

  Image file type: Analyze (.img/.hdr)
- **Lesion Segmentation**: This segmentation still requires the manual intervention of an expert. The result is a binary image in which voxels values are equal to 1 inside the lesion and 0 elsewhere. This provides us with Regions of interest (ROI). In the stroke application, the infarcted part of the tissues is considered.

  Image file type: Analyze (.img/.hdr)

B) Statistical analysis: creation of the Anatomo-Clinical Maps

- **Maximum Overlap Map**: All the maps containing the ROIs are superimposed on each other. This yields a map in which voxels value is

the number of overlapping ROIs. This map is called *Maximum Overlap Map*.

Image file type: Analyze (.img/.hdr)

- **Assignment of performance to the lesion**: For each patient, the value 1 inside the lesion ROI is replaced by the score this patient obtained in the neurological test.

  Image file type: Analyze (.img/.hdr)

- **Mean Value Map**: All the scored images are summed, then divide by the thresholded Maximum Overlap Map. It results in a Mean Value Map.
- **Half Length 95% Confidence Interval**: Using the Maximum Overlap Map, the masked mean value map and every patient's scored image, the Half Length 95% Confidence Interval ($CI_{95\%}$) map is generated. $CI_{95\%}$ gives an estimated range of values which is likely to include an unknown population parameter. The estimated range is calculated for each voxel of the image from the neuropsychological test scores of the patients having a lesion in that voxel. $CI_{95\%}$ is calculated as follows:

$$CI_{95\%}(x) = \bar{x} \pm 1.96\frac{s}{\sqrt{n}},$$

where *x* is the value obtained in a voxel, $\bar{x}$ is the average of the values obtained by the patients whose lesions overlap in that voxel, n is the number of patients whose lesions overlap in the voxel, and *s* is the standard deviation of the test score values in the voxel.

Image file type: Analyze (.img/.hdr)

- **p-statistical map:**
  - **Labeling**: Each sub-region defined by the contiguous overlap of ROIs in the mean value map is labeled with an integer value. This yields a label map.

    Image file type: Analyze (.img/.hdr)

  - **Vectorization**: Using the label map and the score map for each patient, a text file is generated that contains the list of the patient's scores for each label. The list is a vector (the label) of vectors (list of scores corresponding to the label).
  - **Statistical Test**: For each labeled sub-region, the scores of the patients (via the vector text file) are compared to the scores obtained by control subjects at the same test. Depending on whether the score obtained by the controls follows a normal distribution, different statistical tests are applied: Student T-Test for a normal distribution or Kolmogorov-Smirnov or Wilcoxon test for other distributions. This yields a statistical map (the so-called Anatomo-Clinical Map) in which voxels value is the p-value of the statistical test

    Image file type: Analyze (.img/.hdr)

- **Bonferonni and Holm correction thresholds**: From the p-values map, we create a text file, which contains the Bonferroni/Holm thresholds. These thresholds may be used in a post-processing step to perform a correction of multiple testing on the p-value map. Note that the AnaCOM pipeline itself does not perform the correction.

1.3.2. Processing tools

The AnaCOM method is developed in the Python language using BrainVISA software for the main interface (Institut Fédératif de Recherche IFR-49, Orsay, France, http://www.brainvisa.info/). Modules are used from:

- AIMS and VIP Libraries (SHFJ-CEA, ORSAY, France, http://www.brainvisa.info/) (Cointepas et al., 2001) provided in the BrainVISA package;
- SPM2 (Welcome Department of Imaging Neuroscience, Institute of Neurology, London, UK, http://www.fil.ion.icl.ac.uk/spm) (Friston, 1995) in conjunction with MATLAB version 6.5 (The Mathworks, Inc., MA, USA);
- R library for the statistics (http://www.r-project.org/) (Ihaka and Gentleman, 1996).

The manual segmentations of the lesions and artefacts of the images of the patients are performed with MRIcro (http://www.mricro.com).

1.3.3. Results produced

Five maps are produced (3D images in the common standardized space):

- maximum overlap;
- labels;
- mean score value;
- 95% Confidence Interval; and
- p-value

as well as two text Files:

- labels coordinates in the common standardized space;
- bonferroni/Holm thresholds.

1.3.4. Scufl representation

The Scufl data flow is show in Figure 7. The first 3 processing steps (artifacts removal, registration and lesion segmentation) correspond to the preprocessing. The rest of the workflow deals with the statistical analysis.

It is to be noted that this workflow cannot be executed with MOTEUR: the accumulation processor correspond to data flow reduction (data synchronization barriers) that cannot be expressed currently. Extensions of the Scufl language

that are studied in the context of another ANR project (GWENDIA, ANR-06-MDCA-009) will enable data synchronization in the future.

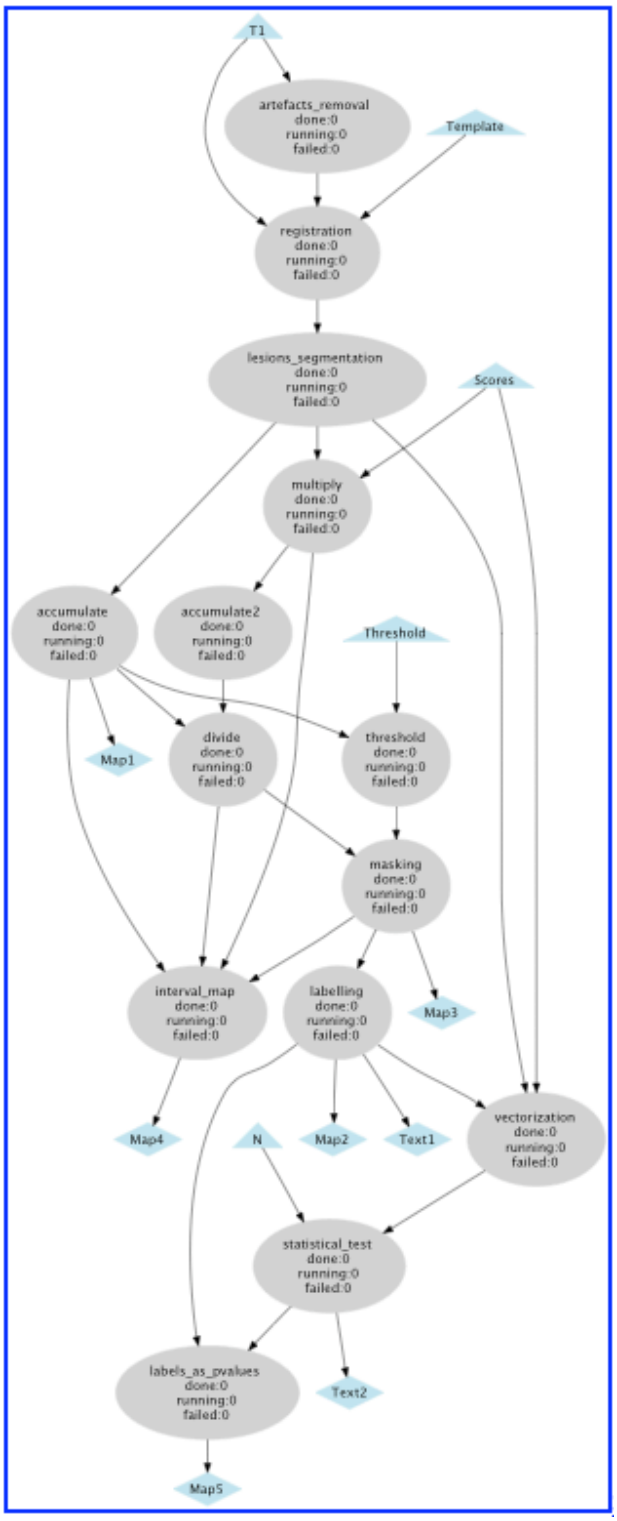

Figure 7. Scufl workflow in MOTEUR interface.

2. Security use cases study

This section describes various security use cases to fulfill the application requirements for data protection. The aim of this section is to assess the capability of the NeuroLOG infrastructure to enact the applications while guaranteeing data protection.

2.1. Security architecture overview

The overall NeuroLOG software architecture described in [3] is drafted in Figure 8. The NeuroLOG platform involves a single *Registry* root server for coordination of the platform, multiple intercommunicating *Site* servers which are implementing most of the middleware functionality and per-site *Clients* from which the user authenticate and connect to the system.

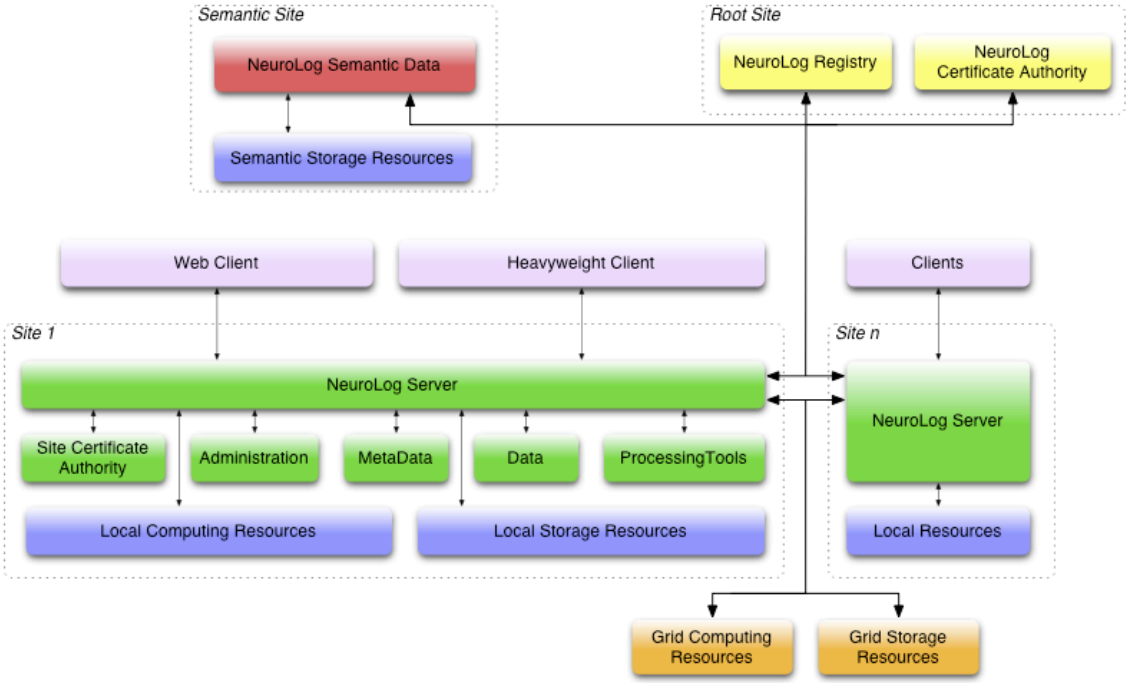

Figure 8. NeuroLOG software architecture (from [3]).

To ensure scalability and sites independence, the administration of the platform is distributed and handled by all site managers. In particular, sites are responsible for granting access to the platform for their users.

2.1.1. NeuroLOG certification chains

Users are authenticated through standard X509 certificates. A hierarchical certification authority is set up in the platform: a root authority, administrated by the Registry administrator is delivering certificates to all participating sites (site CA certificates). It delegates certification ability to each site administrator for her

local users. End users receive individual certificates that are thus resulting from a 2-levels signature chain (root CA and site CA) as illustrated in Figure 9.

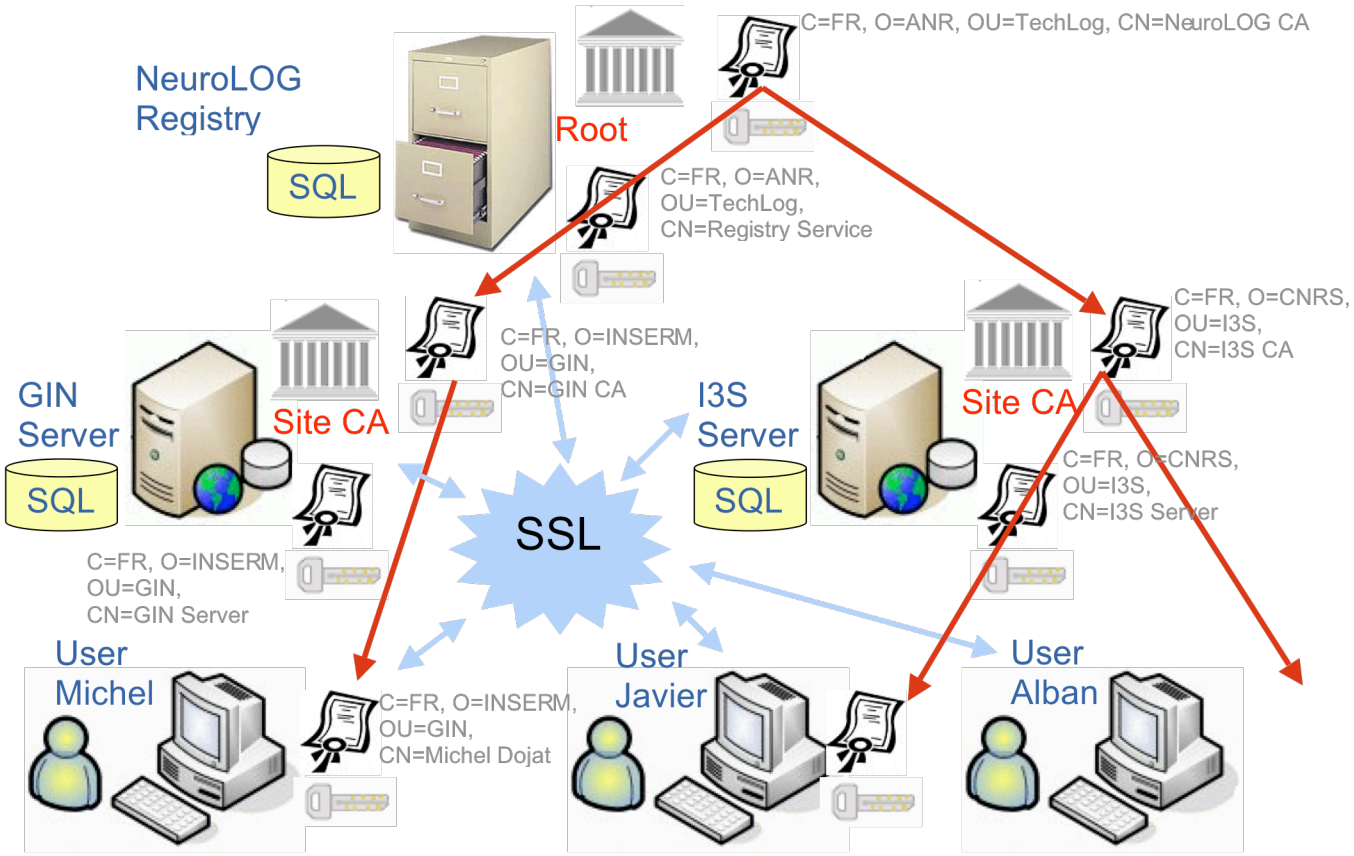

Figure 9. NeuroLOG authentication scheme.

Figure 9 shows a deployment example with 2 sites (GIN and I3S). The GIN site has 1 registered user (Michel) and I3S has two users (Javier and Alban). The registry and the site have 2 certificates each: a *CA certificate* used to sign other certificates and a *server certificate* used to authenticate each server. The red arrows show the certification signature chains. The blue arrows show the possible authenticated communication between participants. It is important to note that since the authorization scheme is decentralized, all servers do not know about the existence of all users. The certificate chain validation mechanism is used to authenticate foreign clients: a client is recognized as a valid entity if its certificate is ultimately signed by the NeuroLOG CA.

2.1.2. Grid CA and other credentials

The EGEE grid middleware uses X509 certificates to authenticate users that are very similar to the NeuroLOG certificate. However, the Grid and the NeuroLOG certificates are not recognized by the same authentication authorities

and cannot be exchanged. As a consequence, the NeuroLOG middleware needs to handle two certificates per user and to use each certificate depending on the service (Grid or NeuroLOG service) invoked on behalf of the user.

Other credentials will be managed *e.g.* for accessing SQL databases. The joint use of different technologies enforces multi-credentials management and synchronization.

2.2. Inter-servers communications

To protect data and control access rights, all communication between NeuroLOG distributed components have to be encrypted and authenticated. From [3], it was decided that Java components would communicate through Java RMIs. The RMI communications will use the standard SSL layer to authenticate and encrypt the messages exchanged. As discussed below however, SSL was designed as a client-server communication mechanism and the porting of RMI over SSL in a distributed (multi-clients) system does not guarantee a satisfying security level for the NeuroLOG medical use cases.

2.2.1. RMI over SSL

Java RMI use the JRMP protocol to exchange objects and trigger methods invocation between Java programs through regular socket communication channels. For securing RMI, JRMP messages can be exchanges over secured sockets.

The Secured Socket Layer (SSL) is based on asymmetric cryptography authentication techniques: it uses X509 certificates and associated private keys. A client C owns an X509 public certificates C_{cert} and an associated private key C_{key}. A server S recognizes a number of client certificates that are stored locally in a trust store. When a client connects, the server and the client proceed with an hand-shaking: the server checks the identity of the client. If C_{cert} is valid (*i.e.* it is not outdated and it was signed by a known certification authority) the socket is opened and the communication begins. Otherwise the server closes the connection. An SSL channel can be configured to perform dual hand-shaking: both the client and the server authenticate to each other. The NeuroLOG middleware ensures the distribution of respective client and server public certificates to the needed trust stores upon their signature. It uses dual hand shaking (thus servers own certificates too).

When using SSL for RMI calls, the SSL handshaking happens when the socket between a client and a server is created, prior to the first JRMP message exchange. Once the connection established, the server is guaranteed that it is receiving messages from a trusted client. However, after handshaking the server cannot distinguish between JRMP messages received from different trusted clients: the JRMP messages carry no identifying information. Two successive messages could come from two different users originating from a same host or from two different hosts. Since all users do not have the same access rights, it is needed for a server to control the identity of the caller for each RMI invocation in order to check the client access rights. A solution to this problem is discussed in section 2.2.2. Furthermore, clients are not known from foreign servers and

cannot be directly authorized. The identification mechanism is therefore extended in section 2.2.3.

2.2.2. Server – server communications

In a site server to site server (or Registry) communication, one of the servers is acting as a client for the other server. As described above, after establishing the SSL connection, there is no control on the client identity. The client is known to be trusted but it could be *any* of the trusted clients (i.e. a user from this site or another site server). To ensure re-identification of the client, each remote method invocation is guarded by the control of the identity of the client at application level as illustrated in Figure 10. Sensitive invocations are encapsulated in a command pattern responsible for the generation of an authentication token. This token is ciphered by the client with its private key and passed as parameter of each sensitive remote invocation. The server then retrieves the client certificate from its SSL context. Re-identification of the client is finally guaranteed by the success of token deciphering.

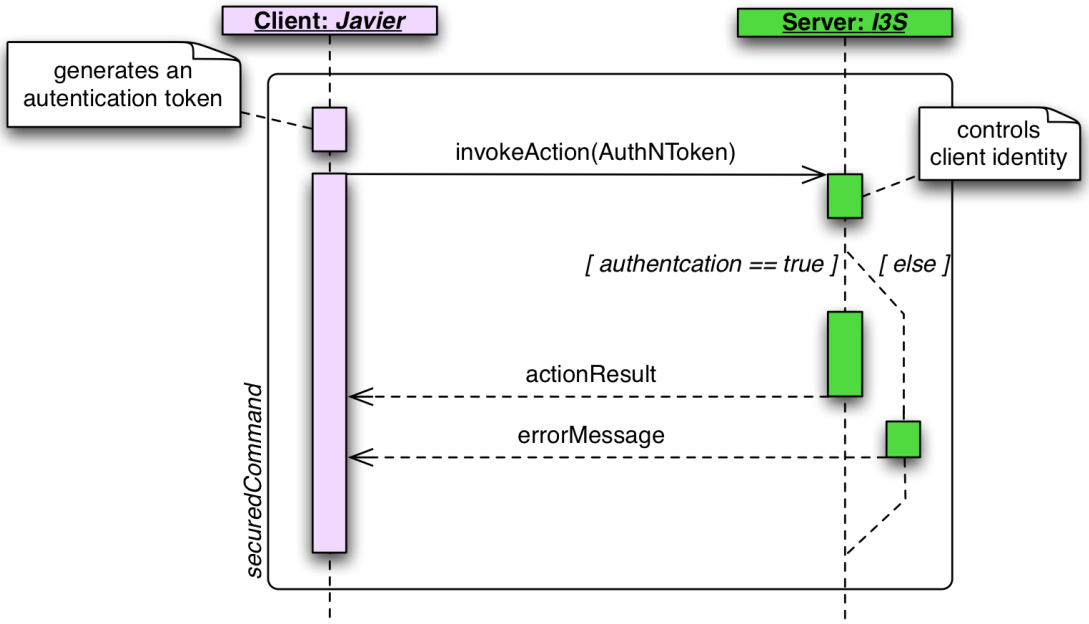

Figure 10. Client authentication sequence diagram.

2.2.3. User client – server communications

A user client may connect to the server it is registered to or to a foreign server. The user client to user server communication case is similar to the server-to-server case discussed in section 2.2.2 and it is treated with the same security pattern. As the client trusts the root certification authority, SSL communications are allowed between a client registered to its “local” server, and a “foreign”

server. The active SSL channel is used in the same way to re-identify at application level, each caller for a sensitive operation.

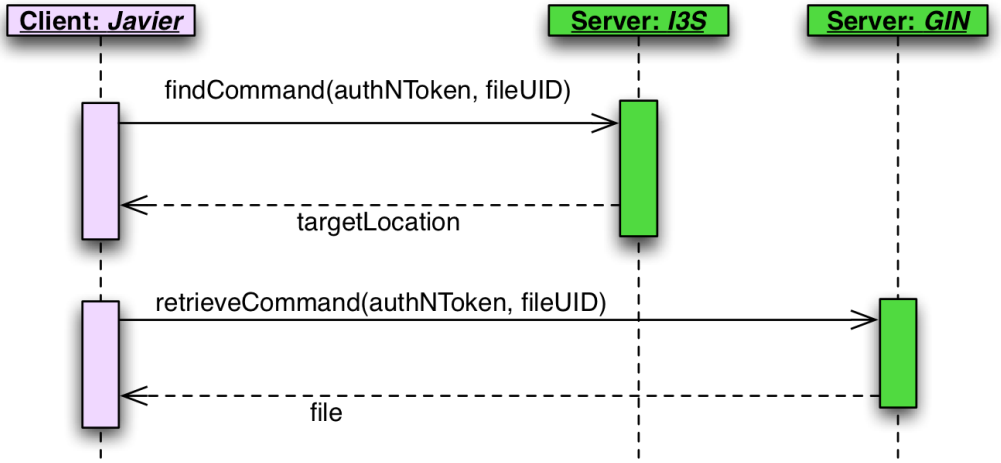

Figure 11. Remote retrieving sequence diagram.

Figure 11 shows the request of a file hosted by a remote site. The client Javier connects to its server I3S and requests its find service. I3S server finds the location of the requested file and returns it to Javier. Finally, Javier connects to GIN and invokes its retrieve service, GIN checks that Javier has a valid authentication token and performs the operation.

2.2.4. Client – Grid communications

The direct communications between a client and the grid will not depend on Java RMI and therefore will not rely on the same mechanism as described above. The client – grid communication will rely on the Grid X509 certificate of the requested user. The EGEE grid security relies on the GLOBUS Security Infrastructure (GSI) [1], which involves registering the clients' distinguished names into any grid server a client can potentially connect to. As a consequence, a client authorization relies on the propagation of its identity to a potentially very large number of resources. This is typically performed once every few hours on the EGEE grid. In the NeuroLOG middleware, we have adopted a technique that is based on the more standard SSL layer. The advantages are the lightweight, low dependency (SSL is installed on any system and host nowadays), the dynamic system (clients identities do not need to be broadcasted) and the seamless integration with Java RMI (no GSI layer exists in Java RMI to the best of our knowledge). The counterpart is the complex session management and session delegation scheme described above.

The EGEE grid infrastructure also makes use of VO Management Servers (VOMS) for users authorization. All users are registered in a centralized VOMS server. The VOMS server declares groups and roles that can be bound to the user identities by the VO administrator. When a user creates an X509 proxy, the server is contacted and X509v3 extensions may be added to the proxy

depending on the belonging of the user to groups and roles. The extensions are carried with the user proxy and can be checked by any service receiving a request from this user to make authorization decisions. This approach has the advantage to seamlessly integrate authorization credentials with the user identification proxy. The counterpart is that it relies on a centralized server (bottleneck that has to be queried at each proxy creation requested) and on a centralized authorization authority (the VO administrator).

In NeuroLOG instead, it was decided to distribute the authorization control to the different site administrators as described in the NeuroLOG Security Policy document [2]. No super-administrator has the power of authorizing some users to access files on different sites and site administrator solely keep control of the data they are managing. This security scheme is better adapted to the requirements of medical users. It is also completely distributed and does not require a VOMS-like centralized server. However, for files administrated on the EGEE grid infrastructure, the middleware will depend on the Grid client and therefore on the EGEE biomed VO manager.

2.3. Firewalls

Firewalls represent the main mechanism for network security. They are present at all levels, on server and computer side, on intranet side and on the Internet side. The firewall role is to enforce network security policy, its task is based on filtering data exchange. The filters use criteria such as:

- Network packet origin/destination (IP address, TCP or UDP ports, network interface…),
- Network adress translation (NAT),
- Packet option (validity, fragmentation…),
- Content conformity (HTTP packet has to contain HTTP content),
- User identification.

In our project we are facing firewalls at several locations as illustrated in Figure 12.

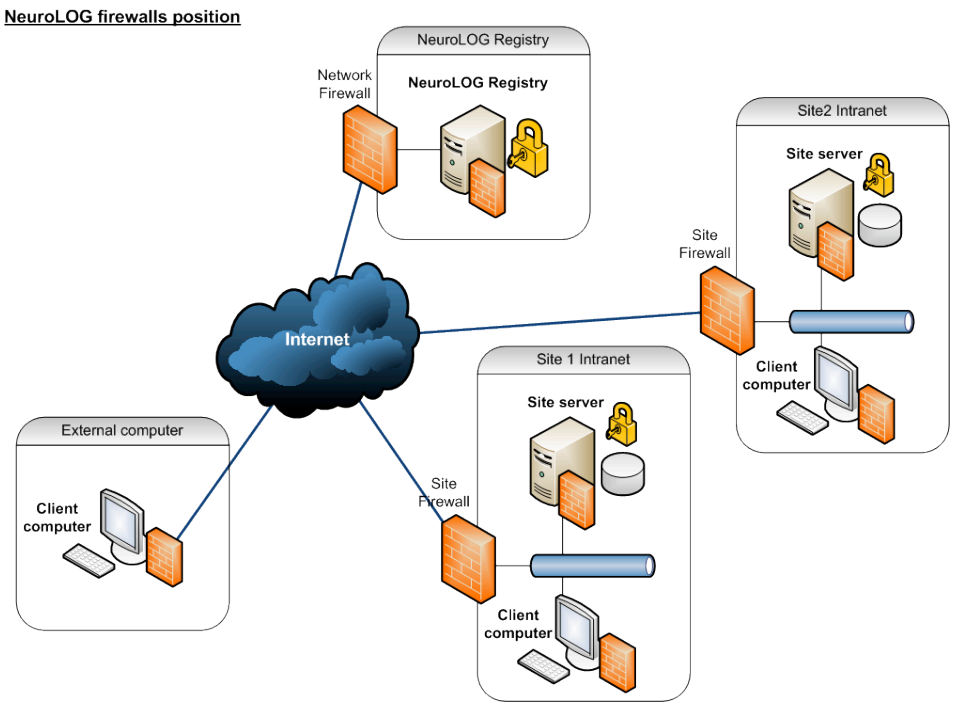

Figure 12. NeuroLOG firewalls.

2.3.1. NeuroLOG theorical architecture

In order to protect the NeuroLOG Registry it will be mandatory to define fine firewalls rules, only providing access to NeuroLOG site servers, identified by their IP and MAC addresses, and exposing only necessary TCP/IP ports used by the project.

The NeuroLOG site servers will have more flexible firewall rules, because they can be accessed by inside and outside site client. In this case we will have to expose TCP/IP ports used by NeuroLOG servers without restricting the IP address origin of the connection.

The NeuroLOG Client will have to authorize ports used by the applications without restriction for destination.

2.3.2. Specific deployment architecture

As the NeuroLOG project takes place in medical research area, we also have to deal with research and clinical organism network architectures where the security policy is paranoid.

In order to manage security, network organism is divided in sub-networks. Each sub-network can communicate directly with another one or through an Internet gateway. A gateway acts as a security entity, managing network exchange between networks, commonly as firewall. Like routers, it also routes network packets and manages authorization to access other networks.

In order to expose the NeuroLOG Registry server and site servers to the world we will be confronted to gateways with their firewall and NAT (Network Adress

Translation) rules. On one side, firewall rules authorize only certain IP ports to be exposed and on the other side, NAT manages network stream redirections.

In a paranoid Network security management mode, that the NeuroLOG project could be confronted to research and medical organism, it will be necessary to discuss with the network administrator in order to find the best solution to the NeuroLOG site infrastructure deployment. Several solutions exist:

- NeuroLOG server in DMZ (Demilitarized Zone): the server is fully exposed to the world; it can be accessed by intranet site but also by Internet. The NeuroLOG security software manages security.
- NeuroLOG server is under a firewall and a gateway. Firewall and NAT rules provide access to local and distant clients. The server is well protected and only useful IP and ports are exposed to the world.
- Same solution as above but with firewall rules only authorizing defined external IP range address to connect themselves to the NeuroLOG server. This solution will refuse nomad client that are not member of a specific network.
- A variation of previous solution, by specifying the IP anc MAC addresses of authorized clients.

2.4. Data access control and transfer

2.4.1. Security policy

As described in the NeuroLOG Security Policy proposal or NSP [2], files access control and transfer require strict user identification. Each user is registered into one site by the site administrator. Access to data is controlled at the group level: as many user groups as needed may be created and data files are individually controlled by group. Complementary to the NSP, all data exported from a site will be anonymized and encrypted prior to transmission for protection as detailed in NeuroLOG deliverable 3 [1].

Database access is managed at group level. Based on SQL 92, this process can apply security rules on tables and columns but not on rows. If we want to manage access at row level, we will have to implement user row access mechanism in the NeuroLOG security software, at the SQL request level. A second problem is that we will have to replicate user's profiles and authorization on all sites.

We are waiting on the NeuroLOG users returns about the data and file access restrictions.

2.4.2. Files encryption and anonymization

Anonymization and Patients reidentification

In order to guaranty patient respect and following the CNIL recommendations, patient data are anonymized when they are imported in the NeuroLOG solution. This anonymisation process is conform to the DICOM Basic Application Level Confidentiality Profile (PS 3.15-2007 Page 33, E 1.1 De-identifier)

As this project takes place at the same time in a research and clinical context it is also necessary to provide a way to re-associate the anonymous data to a

specific patient. One of the simplest solutions is to follow the DICOM protocol. DICOM does not provide patient specific unique identifier but study unique identifier. So we can keep, in NeuroLOG site-specific database, a relation between a study unique identifier and a specific patient in our anonymization mechanism.

DICOM provides a solution to permit files identification: all personal data are kept encrypted. The encryption key is property of the origin site. DICOM provides several TAG to manage encryption:

- Encrypted Content (0400,0520) Attribute,
- Encrypted Attributes Sequence (0400,0500),
- Encrypted Content Transfer Syntax UID (0400,0510)

The DICOM encryption algorithms are either AES or Triple-DES in all possible key lengths.

Another solution is to provide a NeuroLOG Patient unique Identifier. This solution asks for a second important question: how to guaranty that a patient is not present several times, with several identifiers, in our data?

It can be useful to provide a patient identifier composed by his own private data, guarantying his uniqueness. We can obtain this last identifier by calculating a checksum on patient records. The result can be compared to a fingerprint proper to a patient. For example, we have a patient, a man, named “John DOE”, born in “Paris”, on 04/07/1963. We can reduce this information to a string :MJOHNDOE19630704PARIS and calculate a checksum based on SHA1 or MD5 algorithm. We obtain a string, “aeBl0lBCd6WL9HnM9OvhA==” with MD5 or “iWTobLFcG26pcsfrRkobJTrecow=” with SHA1 algorithms. The resulting string does not provide any sensitive information on the patient.

If this patient goes to another NeuroLOG site in order to have a medical study, we can search for his fingerprint in the NeuroLOG solution. This solution is not perfect because we cannot avoid real homonymous problem that occurs when two people have the same patient informations.

Encryption

In order to guaranty data confidentiality, the files can be encrypted using AES algorithm describes in section Data encryption of Technical report: Specification of the NeuroLOG architecture [1]. The files will be encrypted in two specific contexts:

- Locally on NeuroLOG site to protect sensitive data,
- Before exchange between two sites, site and external client or between site and Grid.

2.4.3. Grid data

If users need to access Grid resources, they will have to be authorized through a Virtual Organization (VO). This process uses the gLite middleware interface in

order to manage users authentication, based on X.509 certificates, and authorization. Once authenticated, users can access to GRID Data with GridFTP.

GridFTP is a high-performance, secure, reliable data transfer protocol optimized for high-bandwidth networks. It is based upon the Internet FTP protocol. GridFTP uses basic Grid security on both control (command) and data channels.

3. Conclusions

This deliverable outlined the NeuroLOG partners' application pipelines and important security considerations in executing these pipelines. The 4 applications data flows, detailed in Section 1, are well understood. This preliminary study has shown that flexibility in the workflows design is expected and that the pipeline developers will probably implement variations. It has also been recognized that some extensions of the Scufl workflow language are required to address the use cases presented. This will be addressed in collaboration with the GWENDIA project (ANR-06-MDCA-009). Future work will also be needed to clearly identified what intermediate result should be recorded during pipeline execution and how to interface with the NeuroLOG semantic data manager.

The security of all communication is vital in the system and raises complex issues. An elaborated identification and data protection scheme for an RMI-based distributed platform has been implemented. In addition, data access, whether hosted by NeuroLOG servers or grid servers, will be strictly controlled.

4. Bibliography

[1] L3: Specification of the NeuroLOG architecture components. NeuroLOG technical report (ANR-06-NLOG-024-L3), November 2007.

[2] J. Montagnat, D. Godard: NeuroLOG Security Policy proposal. NeuroLOG technical report (ANR-06-NLOG-024), December 2007.

[3] A. Gaignard, J. Montagnat: NeuroLOG Software Architecture. NeuroLOG technical report (ANR-06-NLOG-024), January 2008.

[4] Miller, D.H. et al.: The role of magnetic resonance techniques in understanding and managing multiple sclerosis. Brain 121 (Pt 1) (Jan 1998) 3–24

[5] Grimaud, J. et al.: Quantification of MRI lesion load in multiple sclerosis: a comparison of three computer-assisted techniques. Magn Reson Imaging 14(5) (1996) 495–505

[6] Zijdenbos, A.P. et al.: Automatic "pipeline" analysis of 3-D MRI data for clinical trials: application to multiple sclerosis. IEEE TMI 21(10) (Oct 2002) 1280–1291

[7] N. Wiest-Daesslé, S. Prima, S.P. Morrissey, C. Barillot. Validation of a new optimisation algorithm for registration tasks in medical imaging. In 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI'2007, Pages 41-44, Washington, États-Unis, Avril 2007.

[8] D. Garcia Lorenzo. Combining Robust Expectation Maximization and Mean Shift algorithms for Multiple Sclerosis Brain Segmentation. MICCAI workshop on “Medical Image Analysis on Multiple Sclerosis (validation and methodological issues)”, *New York University ·NYC ·USA.*

[9] L.S. Aït-Ali, S. Prima, P. Hellier, B. Carsin-Nicol, G. Edan, and C. Barillot, "STREM: A Robust Multidimensional Parametric Method to Segment MS Lesions in MRI," in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2005, vol. 3749, Lecture Notes in Computer Science, J. S. Duncan and G. Gerig, Eds. Palm Springs, CA, USA: Springer-Verlag, 2005, pp. 409.

[10] P. Coupé, P. Yger, S. Prima, P. Hellier, C. Kervrann, C. Barillot. An Optimized Blockwise Non Local Means Denoising Filter for 3D Magnetic Resonance Images. IEEE Transactions on Medical Imaging, 27(4):425-441, Avril 2008.