The Brainomics/Localizer database

Dimitri Papadopoulos Orfanos[a,*], Vincent Michel[e], Yannick Schwartz[c,a], Philippe Pinel[b,a,d], Antonio Moreno[b,a,d], Denis Le Bihan[a], Vincent Frouin[a]

[a] *CEA, DSV/I2BM, NeuroSpin, 91191 Gif-sur-Yvette, France*
[b] *INSERM, U992, Cognitive Neuroimaging Unit, 91191 Gif-sur-Yvette, France*
[c] *Parietal team, Inria Saclay Île-de-France, 91120 Palaiseau, France*
[d] *Univ. Paris-Sud, Cognitive Neuroimaging Unit, 91191 Gif-sur-Yvette, France*
[e] *Logilab, 104 boulevard Auguste Blanqui, 75013 Paris, France*

Abstract

The Brainomics/Localizer database exposes part of the data collected by the in house Localizer project, which planned to acquire four types of data from volunteer research subjects: anatomical MRI scans, functional MRI data, behavioral and demographic data, and DNA sampling. Over the years, this local project has been collecting such data from hundreds of subjects. We had selected 94 of these subjects for their complete datasets, including all four types of data, as the basis for a prior publication; the Brainomics/Localizer database publishes the data associated to these 94 subjects. Since regulatory rules prevent us from making genetics data available for download, the database serves only anatomical MRI scans, functional MRI data, behavioral and demographic data.

To publish this set of heterogeneous data, we use dedicated software based on the open-source CubicWeb semantic web framework. Through genericity in the data model and flexibility in the display of data (web pages, CSV, JSON, XML), CubicWeb helps us expose these complex datasets in original and efficient ways.

Keywords:
imaging genetics, database, semantic web

*Corresponding author
Email address: dimitri.papadopoulos@cea.fr (Dimitri Papadopoulos Orfanos)

1. Introduction

The Brainomics/Localizer database is a data repository containing datasets from 94 subjects with structural MRI scans, functional MRI data, behavioral and demographics data. DNA sampling has been performed on the subjects, but we cannot publish such data due to regulatory rules.

Datasets have been acquired by the in house Localizer project which initially planned to investigate inter-subject variability [1]. We have been collecting data from volunteer research subjects taking part in different studies carried out in our lab. The investigators of these studies agreed to provide behavioral and demographic data, anatomical MRI scans and DNA sampling. They also agreed to acquire a short fMRI sequence, approximatively 5 minutes long, after their functional imaging session, specifically for the Localizer project. We were thus able to collect data from a considerably larger number of volunteer research subjects than a single study could afford.

We have also been working on genetic neuroimaging in the context of the Brainomics project. We felt the need for a database that could index and expose heterogeneous data including MRI images, genetic data or behavioral data. We based our software developments on the CubicWeb semantic web framework and wrote specific CubicWeb modules to describe and visualize neuroimaging and genetic data. We decided to build a Brainomics/Localizer demonstrator based on the Localizer dataset. The resulting database is now publicly available[1] as well as the source code[2].

We also viewed the Brainomics/Localizer demonstrator as an opportunity to study the feasibility of opening up medical research data as support material for scientific articles. Regulatory rules differ from country to country and may hamper homogeneous publication of scientific data: we do not know of other public databases of research individual health information in France–and suspect there are very few.

2. Material and methods

2.1. De-identification of the database

The local ethics committee had initially approved the Localizer study. Starting the Brainomics/Localizer database effort, we addressed with the

[1] `http://brainomics.cea.fr/localizer`

[2] `https://github.com/neurospin/localizer`

ethics committee the publication of Localizer data as support material for one of the published paper [2], in order to facilitate replication of the results. The following paragraphs describe the agreement that led to the publication. Please note that not all described functionality has been implemented. More specifically we currently do not provide means to run calculations involving genetic data. As a result genetic information is currently not publicly available from the database, although it is internally available to the server.

> Before publishing the data, we anonymize it in an irreversible way by re-encoding all subject identifiers and discarding the conversion table. Data is stored on an online server and made available to the broader scientific community as a web service. Users can access the data from a web browser.

In our lab, any mention of name, social security number or other similar data are prohibited in all the data acquired for research purposes. Instead we use a subject identifier; the correspondence between this local identifier and sensitive data, such as names, is kept securely. Conversion methods are hosted on a specific system restricted to medical staff. The publication process requires that local identifiers are converted into new random identifiers and the conversion table is discarded.

As a result of this irreversible re-encoding, updates are not straightforward. In the event we have to remove a subject, we would have to get back to the source data on our internal network, remove data based on local subject identifiers and then re-encode them to new random identifiers.

> *Imaging data* In addition to re-encoding subject identifiers, anatomical MRI images are defaced.

We used the *mri_deface* tool of Freesurfer [3] to deface anatomical images.

> *Genetic data* The very nature of genotyping data strongly identifies a subject, by mere comparison to other genotyping data collected elsewhere. As a result genetic data cannot be downloaded from the server. Users can nevertheless start calculations on the server itself from a user interface, using the genetic data of all subjects as a parameter. The results of such calculations are images like those presented in Fig. 1 of [2]. These actions

should be crafted carefully to forbid retrieval of individual genotypic data.

Demographic and behavioral data Only data related to the publication [1] are uploaded to the server. Such data do not present a risk of identifying the subject.

We feel it is hard to identify an individual by comparing the demographics and behavioral data available from the database to similar data collected elsewhere. Contrarily DNA samples collected elsewhere could easily be compared and matched with DNA data from the database.

2.2. Software infrastructure

2.2.1. The Brainomics genetic neuroimaging database

The need to manage data growing in volume and complexity have led the neuroimaging field to rely increasingly on database infrastructures. These databases typically provide support for multiple data types *e.g.*, brain images, behavioral and demographic data, neuropsychological scores, and genetic data. Popular solutions for storing such data at a large scale are COINS [4] and XNAT [5].

We chose CubicWeb[3], which is an alternative open-source framework. We customized it for the requirements of imaging genetics. The resulting Brainomics genetic neuroimaging database permits deep integration of imaging and genetic data [6]. We use it internally to query jointly genetic and neuroimaging data–but as already explained the publicly available database currently excludes genetic data.

XNAT and COINS not only expose neuroimaging data, they also collect and even process data. In contrast we focused on exposing and offering different views on the data, including web pages for human consumption, mechanisms for download, and semantic-web queries run from processing software.

2.2.2. The CubicWeb framework

The CubicWeb framework follows the semantic web approach: data are exposed using ontologies for easier sharing, access, and processing, and each item is identified by a unique ID (called *Uniform Resource Identifier* or

[3] `https://www.cubicweb.org`

URI). CubicWeb is built upon well established core technologies such as SQL, Python and web standards (HTML5 and JavaScript). It has been successfully used in large semantic web and knowledge management projects [7].

One major part of a CubicWeb application is the data model, defined as *entities* and *relations* by Python classes, from which CubicWeb generates the underlying SQL tables. It is thus possible to query the data via the RQL language which predates but is similar to W3C's SPARQL. This language provides an abstraction over the underlying database, queries being expressed in terms of business logic rather than low-level SQL schema. For example, *Query all the scans of male subjects* can be expressed in RQL as *Any X WHERE S is Subject, S gender "male", X is Scan, X concerns S*.

Moreover, CubicWeb implements a mechanism to expose information in several ways called *views*. Being defined in Python, the views are applied on query results, and can produce any kind of output, such as web pages, but also binary data or even trigger external processing. The separation of queries and views holds major advantages:

i) The same data selection may have several representations, *e.g.* the subject *S65*, defined by a single query (*Any X WHERE X is Subject, X identifier "S65"*) can be viewed as HTML, or downloaded in the JSON, RDF or CSV formats (see Listing 1). Each couple *(query, view)* is identified by a unique Universal Resource Locator (URL).

ii) Data can be exported in several other formats (*e.g.* XCEDE or MAGE-ML interchange formats) without modifying the underlying data storage. The data model can be performance-oriented, adding a new ontology for sharing the data being simply a new view to define.

```
http://brainomics.cea.fr/localizer/dataset?rql=Any X WHERE X is Subject, X identifier "S65"
http://brainomics.cea.fr/localizer/dataset?rql=Any X WHERE X is Subject&vid=csvexport
http://brainomics.cea.fr/localizer/dataset?rql=Any X WHERE X is Subject&vid=xcede
```

Listing 1: Example of URLs containing RQL queries. They permit to uniquely identify data associated with the queries in the Localizer database. From top to bottom: select subject "S65" and by default display a web page, select all subjects and return the results as a CSV tabular file, and select all subjects and return the results in the XCEDE format.

Finally, CubicWeb has a security system, coupled to the data model definition, that grants fine-grained data access rights. CubicWeb may run as a standalone application or use Apache as a front end and an alternative for logging, monitoring or authentication purposes. CubicWeb may also use LDAP as an alternate source for user credentials and information.

2.2.3. Development of domain specific modules

We developed one *cube* (CubicWeb module) per data type. Each cube is connected to the others – if needed – in the final database schema. The *medicalexp* cube contains the definition of general *entities* (or classes of objects) like Subject, Center, Assessment; the *neuroimaging* (resp. *genetics*) cube defines entities and relations like Scan, Scanner (resp. SnpVariant, Platform, GenotypeMeasurement). The concepts have been be modeled as much as possible upon the XCEDE [8] schema. Each cube implements the corresponding views (navigation, download), triggers and access rights. Connected together, those cubes and others are used to build the complete Brainomics/Localizer database (Fig. 1). Like the CubicWeb framework, these cubes are distributed under the LGPL license and the source code is available from the CubicWeb web site[4].

These domain specific modules add the capacity to download data in the XCEDE [8] format, in addition to the JSON, RDF and CSV formats natively supported by CubicWeb.

3. Description of the repository

3.1. Purpose of the database

Our database was designed to publish data from the Localizer project [1] and more specifically the subset of 94 subjects examined in [2], and make it available to the broader scientific community. Our intent was to set up a demonstrator for the software we have developed in the context of our Brainomics project.

We provide a static set of data. In the short term we have no plans for adding data from other subjects of the Localizer study.

[4] `https://www.cubicweb.org/project/cubicweb-brainomics`

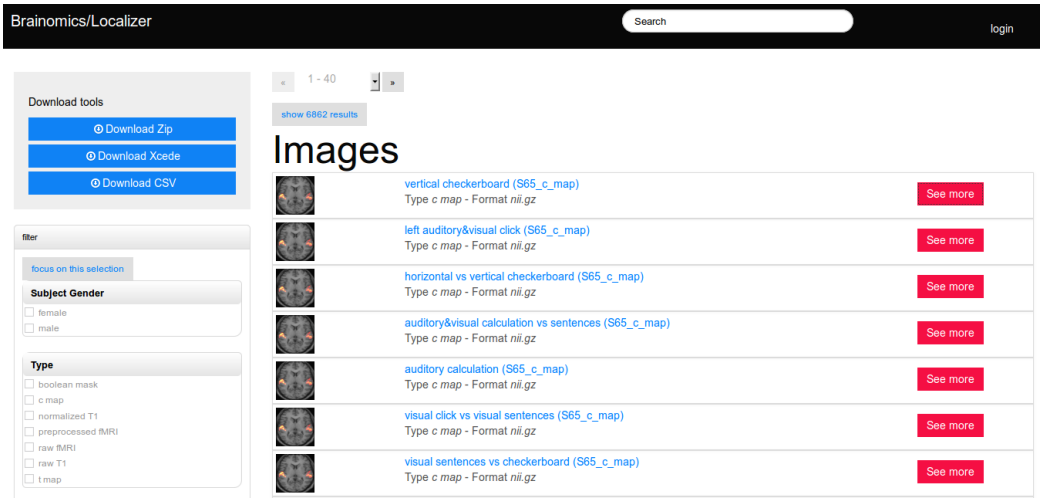

Figure 1: Screenshot of the Brainomics web interface, showing the downloading and filtering *facets* (menus) on the left, and the summary representation of the images stored in the database.

3.2. Available data

Of the hundreds of subjects acquired in house by the Localizer project, we selected a subset of 94 subjects for their complete datasets [2] including anatomical MRI scans, functional MRI data, behavioral and demographic data, and DNA sampling. Retained subjects were mostly young Caucasian highly educated, mostly normal readers, 49 women and 45 men. The exact age of 4 subjects remains unknown, the rest were from 18 through 49 years old – mean age was 24.7 years old. All subjects were right-handed and native French speakers.

Demographic data displayed by the database include gender, age at the time of inclusion, handedness, native language and a family identifier which helps identifying siblings.

Behavioral data aimed to create a rough cognitive profile of each subject. Each profile contains scores for 126 questions covering education, developmental disorders, reading difficulties, basic numerical knowledge, arithmetical skills, visuo-spatial abilities, and visuo-motor abilities.

Two 3 T MRI scanners have been used by our lab for routine acquisitions over time, a Bruker 3 T scanner and a Siemens Trio Tim 3 T scanner. The MRI data were acquired on either of these scanners.

Data processing was performed with SPM8 [5]. The anatomical scan was spatially normalized to the ICBM152 T1-weighted brain template defined by the Montreal Neurological Institute using the default parameters (including the nonlinear transformations and trilinear interpolation). Raw and processed MRI data are available in the NIfTI format. Both raw and normalized T1-weighted anatomical MRI scans are provided for each subject. Functional MRI data includes raw EPI scans, preprocessed fMRI scans and statistical parametric maps. The fMRI experimental design as well as data processing are described in more detail in the initial Localizer article [1]. Let us only cite the challenging constraints taken into account when designing the sequence:

- the sequence had to be short, so as to disrupt as little as possible the main protocol. We choose 5 minutes for performing 100 trials.
- we aimed to obtain for each subject a description of different levels of functional architecture, from sensorimotor areas (perception and action) to more associative areas involved in reading, language processing and calculation.
- we aimed to capture in 5 minutes most of the individual networks related to each task.
- individual networks described in 5 min had to be reproducible over sessions and time.

Images are made available as NIfTI files. They can be downloaded as ZIP files from the Brainomics/Localizer server. Other data can be viewed in tabular form in the Web interface and exported to a variety of formats such as JSON, RDF, CSV and XCEDE, as described in paragraph 2.2.2.

A URI is associated to every entity stored in the database, such as a Subject, a Scan or a ScanGenotypeMeasurement. Furthermore, a URI is also associated to every *(query, view)* couple used to select and display or download data. Such URIs can be kept as a permanent link to complex sets of data.

[5] `http://www.fil.ion.ucl.ac.uk/spm/software/spm8/`

```python
import numpy as np

from sklearn.cross_validation import LeaveOneLabelOut, cross_val_score
from sklearn.linear_model import LogisticRegression
from nilearn.datasets import fetch_localizer_contrasts
from nilearn.input_data import NiftiMasker

# fetch the specified tasks from the localizer database

tasks = [
    'horizontal checkerboard',
    'vertical checkerboard',
    'sentence reading',
    'calculation (auditory cue)',
    'calculation (visual cue)',
    'left button press (auditory cue)',
    'left button press (visual cue)',
    'right button press (auditory cue)',
    'right button press (visual cue)',
]

localizer_data = fetch_localizer_contrasts(tasks, get_tmaps=True)
images = np.array(localizer_data['tmaps'])

# we denote the statistical maps as X, and the target to predict as y
masker = NiftiMasker(standardize=True, memory='cache')
X = masker.fit_transform(images)
y = np.array(['calc' if 'calculation' in img else 'non-calc' for img in images])

# we perform a leave-one subject out cross validation
cv = LeaveOneLabelOut(ext_vars.subject_id.values)

# we use a LogisticRegression classifier with default parameters
clf = LogisticRegression()

scores = cross_val_score(clf, X, y, cv=cv, n_jobs=-1, verbose=1)
print 'scores mean=%.02f, std=%.02f' % (np.mean(scores), np.std(scores))
```

Listing 2: Example of NiLearn's Localizer fetcher: the Localizer database data are directly downloaded from Python code and used to learn a model that predicts calculation tasks.

3.3. Quality control and review of the data

All anatomical MRI scans have been examined by radiologists for possible health issues. A *summary sheet* (Fig. 2) was generated for every subject to evaluate the quality of the fMRI acquisition. A script based on SPM scripts generates movement curves or *glass brain*. For each contrast a specific region of interest (ROI) has been defined. For each subject, the quality of a contrast is evaluated as the ratio of activated voxels in the ROI to the mean number of voxels activated in the ROI across all 94 subjects. Good contrasts are defined by a ratio of over 10%. We checked the summary sheets one by one

and all 94 selected subjects were considered good enough with scores of 5 or 6 good contrasts out of 6.

Demographic data were obtained from different sources, questionnaires performed by the nurses and other questionnaires by the experimenters. We cross-checked redundant information and looked for suspect values wherever possible while merging information from these different sources.

We manually reviewed behavioral data for each subject in the same way, checking that the relevant tables in the database matched correctly the paper questionnaires from which they were extracted.

DNA sample processing was outsourced to a specialized lab. Genetics data were roughly controlled by us too, mainly in order to include them in the proper format.

3.4. How to access the data

Data are available to everyone, without prior authentication, at `http://brainomics.cea.fr/localizer`. A registration process had initially been set up but later removed because it deterred users from trying to access the data. The same mechanism is used for imaging and phenotypic data. Imaging data can be selected in the web user interface, and then packed into a ZIP file that can be downloaded. Phenotypic data are directly available in tabular form in the web user interface and can be downloaded as a CSV file among other formats.

Data are made available under the Creative Commons 3.0 licence (CC BY 3.0). We ask our work is acknowledged in publications that use data from the Brainomics/Localizer database. Thanks to the permissive license, the statistical parametric maps are also downloadable from NeuroVault [9], as we aim to make the Localizer data available to a broader audience. The fMRI data can be downloaded from the NiLearn Python library [10] as an example dataset, which may be used as a testbed to gauge analytic techniques.

3.5. Data updates

Users do not authenticate on our server. We therefore have no way of updating them other than posting information on the web site.

Since this database comes in support of an existing publication [2], we do not plan on adding new subjects or modifying existing subjects. In the event of subject withdrawal, as explained in paragraph 2.1, regulatory rules leave us no choice but using new random identifiers for all subjects. The only

option in that case would be to inform of a removal, without being able to identify the removed subject.

3.6. Future developments

The current file download architecture is similar to the one found in XNAT: a ZIP file is created containing the files associated to a specific query and can then be downloaded. We plan on developing a faster and more robust process, providing persistent links to the files associated to the result of a query.

Although we do not plan on hosting new data in this specific database, we do operate other databases running the same CubicWeb/Brainomics software, for projects such as IMAGEN [11] or EU-AIMS [12] which keep collecting and exposing new data.

The long term plans for this resource is to keep it alive as a public example instance of our software. Since the data itself is mostly frozen, the database can be seen as a "technical object" that can be handed over to the team of engineers and IT specialists who will be managing the other databases based on the same software.

4. Conclusion

We opened up heterogeneous data from 94 subjects of the Localizer project, selected for the completeness of available data including anatomical MRI scans, functional MRI data, behavioral and demographic data. DNA sampling cannot be made publicly available due to regulatory rules. The 94 datasets can be downloaded under a permissive license.

The data are made available on a dedicated server hosting the Brainomics/Localizer database. The database is built upon the CubicWeb semantic web framework and a few extension modules that define the data model and *views*. It supports standard metadata formats in neuroimaging such as XCEDE and defines permanent URLs for each queryable entity in the database.

5. Acknowledgements

This work was supported by ANR-10-BINF-04. We thank Stanislas Dehaene for his participation to the creation and for the sharing of the Localizer database. We thank the Inria-CEA Parietal team and in particular Virgile Fritsch for the Localizer data fetcher in NiLearn.

[1] P. Pinel, B. Thirion, S. Meriaux, A. Jobert, J. Serres, D. Le Bihan, J.-B. Poline, S. Dehaene, Fast reproducible identification and large-scale databasing of individual functional cognitive networks, BMC Neurosci. 8 (1) (2007) 91. doi:10.1186/1471-2202-8-91.

[2] P. Pinel, F. Fauchereau, A. Moreno, A. Barbot, M. Lathrop, D. Zelenika, D. Le Bihan, J.-B. Poline, T. Bourgeron, S. Dehaene, Genetic variants of FOXP2 and KIAA0319/TTRAP/THEM2 locus are associated with altered brain activation in distinct language-related regions, J. Neurosci. 32 (3) (2012) 817–825. doi:10.1523/JNEUROSCI.5996-10.2012.

[3] B. Fischl, Freesurfer, NeuroImage 62 (2) (2012) 774–781. doi:10.1016/j.neuroimage.2012.01.021.

[4] A. Scott, W. Courtney, D. Wood, R. de la Garza, S. Lane, M. King, R. Wang, J. Roberts, J. A. Turner, V. D. Calhoun, COINS: an innovative informatics and neuroimaging tool suite built for large heterogeneous datasets, Front. Neuroinform. 5 (2011) 33. doi:10.3389/fninf.2011.00033.

[5] D. S. Marcus, T. R. Olsen, M. Ramaratnam, R. L. Buckner, The extensible neuroimaging archive toolkit, Neuroinformatics 5 (1) (2007) 11–33. doi:10.1385/NI:5:1:11.

[6] V. Michel, Y. Schwartz, P. Pinel, O. Cayrol, A. Moreno, J.-B. Poline, V. Frouin, D. Papadopoulos Orfanos, Brainomics: A management system for exploring and merging heterogeneous brain mapping data, poster presented at OHBM 2013, 19th Annual Meeting of the Organization for Human Brain Mapping, 16-20 June 2013, Seattle, United States (2013).

[7] A. Simon, R. Wenz, V. Michel, A. Di Mascio, Publishing bibliographic records on the web of data: Opportunities for the BnF (French national library), in: The Semantic Web: Semantics and Big Data, Vol. 7882 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2013, pp. 563–577. doi:10.1007/978-3-642-38288-8_38.

[8] D. B. Keator, S. Gadde, J. S. Grethe, D. V. Taylor, S. G. Potkin, A general XML schema and SPM toolbox for storage of neuro-imaging results and anatomical labels, Neuroinformatics 4 (2) (2006) 199–211.

[9] K. Gorgolewski, T. Yarkoni, S. Ghosh, R. Poldrack, J.-B. Poline, Y. Schwartz, T. Nichols, C. Maumet, D. Margulies, NeuroVault: a web repository for sharing statistical parametric maps, poster presented at OHBM 2014, 20th Annual Meeting of the Organization for Human Brain Mapping, 8-12 June 2014, Hamburg, Germany (2014).

[10] A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, G. Varoquaux, Machine learning for neuroimaging with scikit-learn, Frontiers in Neuroinformatics 8 (14). doi:10.3389/fninf.2014.00014.

[11] G. Schumann, E. Loth, T. Banaschewski, A. Barbot, G. Barker, C. Büchel, P. J. Conrod, J. W. Dalley, H. Flor, J. Gallinat, H. Garavan, A. Heinz, B. Itterman, M. Lathrop, C. Mallik, K. Mann, J.-L. Martinot, T. Paus, J.-B. Poline, T. W. Robbins, M. Rietschel, L. Reed, M. Smolka, R. Spanagel, C. Speiser, D. N. Stephens, A. Ströhle, M. Struve, the IMAGEN consortium, The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology, Mol. Psychiatry 15 (12) (2010) 1128–1139. doi:10.1038/mp.2010.4.

[12] K. L. Ashwood, J. Buitelaar, D. Murphy, W. Spooren, T. Charman, European clinical network: autism spectrum disorder assessments and patient characterisation, Eur. Child Adolesc. Psychiatrydoi:10.1007/s00787-014-0648-2.

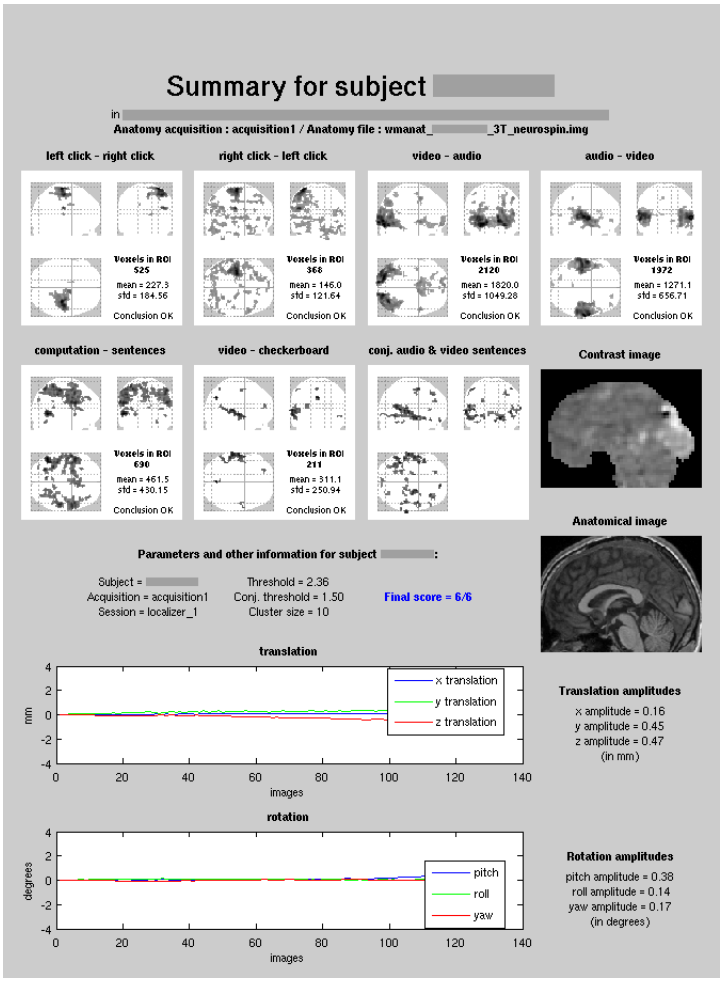

Figure 2: This summary sheet generated for each of the subjects helps assess the quality of the fMRI acquisition. The summary sheet shows, in the lower half, the translation and rotation movement curves for the subject, with maximal amplitudes printed beside the graphics. It also shows sagittal views of the normalized anatomy and of an unthresholded contrast in order to check that normalization has been performed correctly. In the upper half of the sheet, six selected contrasts and a conjunction of two contrasts are shown on *glass brain* figures. The goal of these images is to verify that the main activations are in the expected regions of the brain. This is done visually but also a quick test is performed: for each contrast, regions of interest and adapted thresholds are defined and, if the activation for the subject is good enough, the contrast is considered correct. Therefore, if the six contrasts are good, the *final score* that appears in the middle part of the sheet will be equal to 6. If only two of them are correct, the score will be 2/6. Other processing parameters are also shown in the middle and top of this summary sheet.