NeuroImage 144 (2017) 309–314

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

The Brainomics/Localizer database

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

ARTICLE INFO

Article history:
Accepted 25 September 2015
Available online 9 October 2015

Keywords:
Imaging genetics
Database
Semantic web

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 with these 94 subjects. Since regulatory rules prevent us from making genetic 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.

Introduction

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

Datasets have been acquired by the in-house Localizer project which initially planned to investigate inter-subject variability (Pinel et al., 2007). 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, approximately 5 min long, after their own 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 such heterogeneous 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 individual health data as support material for scientific articles. Indeed regulatory rules differ from country to country and may hamper homogeneous publication of scientific data. We do not know of other public research databases of individual health information in France – and suspect there are very few – and we were only able to spot a couple public databases created for educational purposes, NeuroPeda[3] being currently active. We have found lists of neuroscience databases[4] which point to sites serving individual health data mostly hosted in the United States, with a few exceptions such as MIRIAD[5] hosted in the United Kingdom. Differences in regulatory rules may partly explain this discrepancy.

* Corresponding author.
E-mail address: dimitri.papadopoulos@cea.fr (D. Papadopoulos Orfanos).

[1] http://brainomics.cea.fr/localizer.
[2] https://github.com/neurospin/localizer.
[3] http://acces.ens-lyon.fr/acces/ressources/neurosciences/Banquedonnees_logicielneuroimagerie.
[4] http://en.wikipedia.org/wiki/List_of_neuroscience_databases.
[5] http://miriad.drc.ion.ucl.ac.uk.

http://dx.doi.org/10.1016/j.neuroimage.2015.09.052
1053-8119/© 2015 Elsevier Inc. All rights reserved.

Material and methods

De-identification of the database

The local ethics committee had initially approved the Localizer study. Starting the Brainomics/Localizer effort to open up Localizer data, we voluntarily limited ourselves to a subset of the whole Localizer dataset. We chose to publish data related to a previous publication (Pinel et al., 2012) based on the following criteria:

i) The dataset should be seen as support material for published scientific results.
ii) If at all possible, the dataset should depend on a single initial agreement with the ethics committee (such agreements might cover many studies) to facilitate discussions with the PI and the ethics committee.

We quote excerpts of the text agreed upon with the ethics committee, translated from French. Please note that not all described functionality has been implemented. 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 is prohibited for 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 new conversion table is discarded.

As a result of this irreversible re-encoding, updates are not straightforward. In the event that we have to remove a subject, we would have to get back to the source data on our internal network, remove the 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 (Fischl, 2012) 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 Pinel et al. (2012). These actions should be crafted carefully to forbid retrieval of individual genotypic data.

Demographic and behavioral data Only data related to the publication (Pinel et al., 2012) are uploaded to the server. Such data do not present a risk of identifying the subject.

The ethics committee board decided demographic and behavioral data can be published without high privacy risks. Contrarily DNA samples collected elsewhere could easily be compared and matched with DNA data from the database. We have *not* implemented any interface to start calculations on the genetic data. This goes beyond the data publication effort and should use an additional layer of dedicated software on top of the database. Genetic data or calculations performed on genetic data are currently *not* available from the server.

Software infrastructure

The Brainomics genetic neuroimaging database

The need to manage data growing in volume and complexity has 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 neuroimaging data at a large scale are XNAT (Marcus et al., 2007) and COINS (Scott et al., 2011).

We chose CubicWeb[6], 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 data, genetic data, and associated demographic and questionnaires (Michel et al., 2013). We use it internally to query jointly genetic and neuroimaging data.

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

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 *Uniform Resource Identifier* or 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 (Simon et al., 2013).

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.

[6] https://www.cubicweb.org.