Brainomics: Harnessing the CubicWeb semantic framework to manage large neuromaging genetics shared resources

David Goyard, Antoine Grigis, Dimitri Papadopoulos Orfanos, Vincent Michel, Vincent Frouin, Adrien Di Mascio

Brainomics: Harnessing the CubicWeb semantic framework to manage large neuromaging genetics shared resources

D. Goyard[1], A. Grigis[1], D. Papadopoulos Orfanos[1], V. Michel[2], V. Frouin[1*], A. Di Mascio[2]

[1] CEA DSV NeuroSpin, UNATI, 91 Gif sur Yvette, France.
[2] Logilab, Paris, France.
[*] vincent.frouin@cea.fr.

Abstract *- **In neurosciences or psychiatry, large multicentric population studies are being acquired and the corresponding data are made available to the acquisition partners or the scientific community. The massive, heterogeneous and complex data from genetics, imaging, demographics or scores rely on ontologies for their definition, sharing and access. These data must be efficiently queriable by the end user and the database operator. We present the tools based on the CubicWeb open-source framework that serve the data of the european projects IMAGEN and EU-AIMS.***

Index Terms *- **Imaging Genetics, Magnetic Resonance Imaging, Medical Informatics.***

I. INTRODUCTION

The recent wide availability of high-throughput genomic and neuroimaging devices makes it possible to study neurodegenerative diseases or psychiatric syndromes through Population Imaging studies (PIs) that comprise several thousands of subjects. Those studies like IMAGEN (impulsivity in adolescents) [1] or EU-AIMS (autism) [2] aim to gather heterogeneous measurements like genotyping, MR neuroimaging, EEG, behaviour scores and neuropsychological questionnaires. PIs are multicentric and the massive, complex and heterogeneous datasets collected are intended to be shared. Classically those data flow is a three steps workflow that consists of (i) data collection, (ii) data alignment, quality control, and possibly high level image processing and (iii) data indexing and exposition. Several frameworks support one or several out of the three different steps listed above. For example the XNAT project underpins numerous neuroimaging databases, including some shared international resources [3]. Openclinica [4] is recognized to ease the collection of electronic data like eCRF or questionnaires. Finally laboratory information management system were developed to collect genomic measurements like BaSE [5].

To the best of our knowledge no general framework addresses the situations with very heterogeneous measurements. Specific adaptations of one existing software or bespoke developments have to be performed. We participated in the Brainomics project (french ANR IA) which includes one workpackage devoted to this issue. The package was twofolds: (i) to develop specific Python modules using the CubicWeb framework (Logilab, Paris) and (ii) to demonstrate feasibility and scalability of a fully featured server of a Shared Data Service (SDS). Our specifications retained the need for a web user interface to provide data entry forms, to browse the data or to upload/download datasets. We wished to uncouple quality control or processing from the SDS and we brought specific attention to the versatility and performance of the download. Finally we considered the query language was a central issue: we need to script the requests to the SDS for the processings, the same language must underpin the requests rendered via the web interface and finally this language should bring interoperable features.

II. MATERIALS AND METHODS

The CubicWeb framework. CubicWeb (CW) is a framework developed by Logilab that follows the semantic web approach: data are structured using ontologies for easier sharing, access, and processing. It also enables data federation and enrichment of local information from external resources. In the Brainomics project, we leveraged CW, and defined domain specific modules denoted ”cubes” easily connected to one another. CW is built upon well established core technologies: SQL, Python, web technologies (HTML5 and Javascript) and is developed under the LGPL license. CW defines its data model with Python classes and generates the underlying SQL tables. The queries are expressed via the RQL language which is similar to W3C’s SPARQL. CW 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 HTML pages or trigger external processings. The separation of queries and views holds major advantages: i) the same data selection may have several representations, ii) data can be exported in several other formats (e.g. XCEDE or MAGE-ML) without modifying the underlying data storage. CW has a security system, coupled to the data model definition, that grants a fine-grained access to the data. CW may run as a standalone application or be smoothly integrated in an Apache platform with LDAP and SFTP (*cf.* Figure 1.).

Development of domain specific cubes. We developped one cube 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 like Subject, Center, Assessment; the *neuroimaging* (resp. *genetics*) cube defines entities and relations like Scan, Scanner (resp. SnpVariant, Platform, GenotypeMeasurement). Each cubes implements the corresponding views (navigation, download) triggers and access rights. Connected together, those cubes and others are used to build the complete SDS for the IMAGEN and EU-AIMS projects.

Development of upload/download cubes. Specific cubes for data transfers were developed for our local SDSs. A cube dedicated to data upload allows fast web-form generation so that users can bundle data and meta-data during uploads. Big files are saved on a filesystem and their path and meta-data are indexed. A specific download cube converts the result of a query to a virtual entry in FUSE file system that in turn can be served through a SFTP server; convenient access rights are set by the cube. This cube avoids data compression or duplication and leverage SFTP clients for data transfer efficient mechanisms.

III. RESULTS

Logilab and Brainomics have set up 2 demonstrators of CW capabilities in terms of neuroimaging database: http://brainomics.cea.fr/localizer or clinical follow-up database: http://www.brainomics.net/demo/.

Data collection. These specific cubes greatly expedited the construction of a complete collect database for EU-AIMS. The EU-AIMS cube is built upon a dozen of cubes like the ones presented previously. EU-AIMS SDS can handle any kind of data produced by the project: genetics, neuroimaging, clinical, EEG, and eye-tracking data coming from 10 centres across Europe. Each acquisition centre directly writes entities inside the database. This requires to set precise access permissions at the level of the entity to avoid data deletion or overwriting. Daily asynchronous procedures navigate the database to check all new entities and acknowledge or throw out the data.

Data exposition. IMAGEN database serves heterogeneous data for 2000 subjects: original Nifti scans (strtuctural, functional, diffusion weighted scans), genotyping, gene expression, methylation, demographics and questionnaires. Data download is provided through a SFTP server: as soon as a subset of data is selected the user may save in the web interface and download the corresponding files using SFTP technology; the corresponding metadata are downloaded as spreadsheets.

IV. DISCUSSION-CONCLUSION

Using the CW framework we designed a set of specific cubes dedicated to shared data service. These cubes include both the model of the data structure and the default

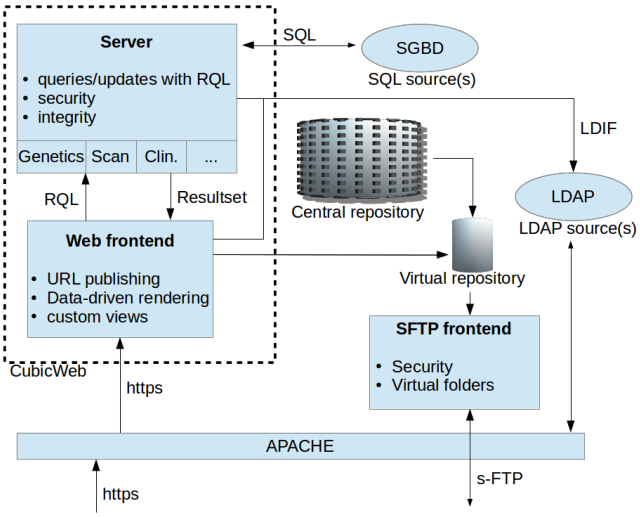

Figure 1: Architecture of the SDSs.

ways to render or export the results of a query. Those different cubes may be arranged to serve data pertaining to different domains. The RQL language brings the possibility to script requests which is useful for the domestic QC/processing of the data. The RQL language enables to create rich views that render the complexity of the data: secondary views, facets. Finally, we will benefit from the RQL *From* clause to implement interoperability: the CW web front end may query concurrently several internet semantic web servers exposing genetics (for instance Entrez-Gene) or clinical meta-information [6].

ACKNOWLEDGMENTS

This work was supported by ANR-10-BINF-04.

REFERENCES

[1] G Schumann and *coll. The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology.* Molecular psychiatry, *15 (12):1128–39, 2010.*

[2] K Ashwood and coll. European clinical network: autism spectrum disorder assessments and patient characterisation. European child & adolescent psychiatry, *2014.*

[3] D Marcus and coll. Human Connectome Project informatics: quality control, database services, and data visualization. NeuroImage, *80:202–19, 2013.*

[4] https://docs.openclinica.com/ (8th january 2015).

[5] L Saal and coll. BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. Genome Biology, *3 (8):1–6, 2002.*

[6] Dan Hall, Michael F Huerta, Matthew J McAuliffe, and Gregory K Farber. Sharing heterogeneous data: the national database for autism research. Neuroinformatics, *10(4):331–9, 2012.*