Establishing and Evaluating Access to Scientific Data
EUDAT Human Brain Project
Big Databases and Cloud Services
Project Report

Tom Wiesing

May 27, 2016

[1] EdN:1

[1] EdNote: Make an abstract

Contents

1 Introduction

1.1 Overview

Multi-dimensional arrays arise naturally in multiple occasions and thus the paradigm of Array Databases has become more and more important. They can be used for a variety of scientific applications, ranging from satellite imagery, over medical imaging techniques to mathematically interesting objects.

One such array database is the Rasdaman system [Bau+98]. It promises to allow users to "storing and querying massive multi-dimensional arrays, such as sensor, image, simulation, and statistics data appearing in domains like earth, space, and life science" [Gmb16].

In this project we want to establish and evaluate access to scientific data. In particular we want to insert this scientific data into the Rasdaman database and evaluate how well the database can handle this kind of data. In this case we want to work together with EUDAT and the Human Brain Project.

1.2 Collaborators

The Human Brain Project [Pro16b] is a European Commission Future and Emerging Technologies Flagship Project with the aim of "providing research infrastructure in the fields of Neuroscience, computing and brain-related medicine". In particular we collaborated with Huanxiang Lu and Dr. Sean Hill from the École polytechnique fédérale de Lausanne to gain access to the Human Brain Database. This archive of multiple sources is a database of (not only) human brain scans.

We also collaborated with Peter Wittenburg and Daan Broede from EUDAT [EUD16]. EUDAT is a "collaborative Pan-European infrastructure for research data services, training and consultancy" and will be used to host the data and resulting interfaces created in this project.

1.3 Project Components

Concretly this project consists of (1) gaining access to the scientific data provided by the Human Brain Database, (2) determining how (and if) this data can be represented inside the Rasdaman system, (3) developing a method to properly ingest the data into Rasdaman, (4) asking the collaborators about useful queries that can be performed on the data, (5) running the queries and gaining new insight into the data and finally (6) evaluating how well Rasdaman was able to deal with the provided data and developed queries.

The section of this report is as follows: In Section 2 we introduce in detail the provided dataset. We then continue in Section 3 by describing how we ingested the data into Rasdaman. Next we describe the queries that we developed over the course of this project in Section 4 before concluding with an evaluation and outlook for future work in Section 5.

2 The Dataset

2.1 The Human Brain Database Dataset

The dataset we have worked with for this project comes from the Human Brain Database project. In total it is about 1 TB in size. It consists of a collection of different brain scans from different sources, each of different resolutions and shapes. Due to the size of this dataset we have not yet accessed all of it.

So far we have only worked with a small subset in order to test our methods. This subset is significantly smaller and is sourced from the SPM Anatomy toolbox [Jül16]. It has a size of under 10 MB and is a brain atlas consisting of 27 different brain scans. Each scan has a resolution of about 150 x 150 x 200 pixels.

2.2 Neuroimaging methods

Before we describe the dataset in more detail we briefly introduce the reader to the most commonly used neuroimaging methods. This provides the neccessary background to understand what the data means. Neuroimaging is the process of scanning the brain and generating images of it. There are two different methods that are commonly used to achieve this.

X-ray computed tomography (commonly known as CT scan) is a method which produces multiple two-dimensional X-Ray images of an object to be scanned. With the help of computers these images are then assembled into a full three-dimensional scan. This allows to look inside an object without having to physically disassemble or open it.

Another common technique is called Magnetic resonance imaging and also known as an MRI scan. MRI scans do not use X-Rays to investigate an object but instead abuse the quantum spin of atoms. By applying a magnetic field, the spins of atoms are forced to align. Once this magnetic field is removed, the atoms return to equilibrium and produce very faint RF emissions. These can be detected, measured and then assembled into a three-dimensional image.

Both methods produce a 3-dimensional cube with scalar values at each point as the output, although MRI scans usually give higher resolution than X-Rays. For this reason our dataset consists mostly of MRI scans.

2.3 The HDF format and BBIC encoding

When we first received the dataset it was provided in HDF5 format. HDF stands for Hierarchical Data Format and is a data format that was originally developed at National Center for Supercomputing Applications and is now mainted by the HDF Group [Gro16]. It is designed to store and organize large amounts of data and has two different kinds of objects, datasets and groups. Datasets are multi-dimensional arrays of homogeneous type. Groups are containers for datasets and sub groups. This results in a filesystem-like structure for HDF5 files.

Even though the data is stored inside an HDF file, it was encoded using the BBIC format. This was an internal format used mostly for an image viewer used by our collaborators. Our collaborators stated that "the bbic format contains a stack of tiled images of different level/resolution. Our image viewer works just like the Google Maps. Every time we zoom in or out, the viewer sends a query to the image service to retrieve certain tiles of images of certain slice at certain resolution" [Lua]. This makes it very difficult for us to extract the original three-dimensional cube that represents the brain scan. What makes it even more difficult is the fact that there is very little documentation available because the "bbic format was developed by a former colleague who has left for long time before I joined" [Lub].

2.4 The ATLAS viewer

The BBIC format is used for the so-called ATLAS viewer. This is based on the imaging service mentioned above and is available at [Pro16a]. The viewer allows the user to interactively explore a subset of the Human Brain Database.

The atlas viewer has three main features: (1) selection of different brain scans, (2) looking at different two-dimensional cuts through the original data (Figure 1) and (3) overlaying of brain region masks (Figure 2).

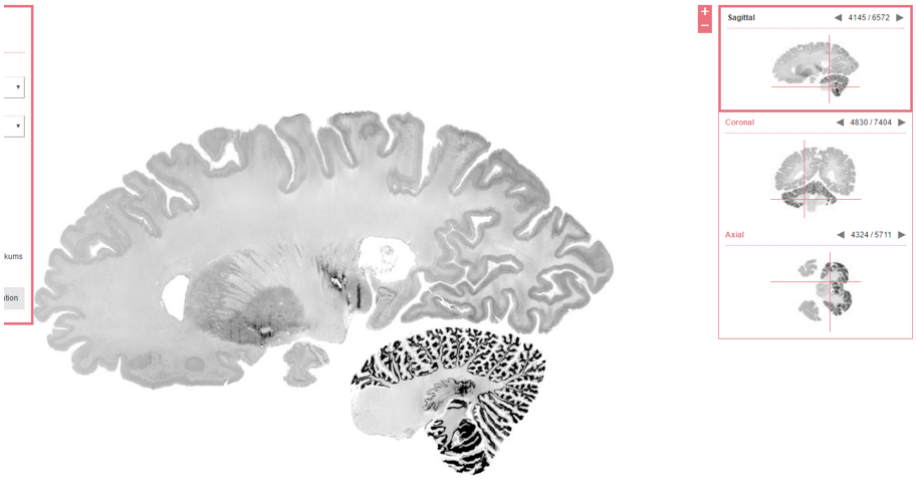

Figure 1: A screenshot of the ATLAS viewer showing a basic cut through a single dataset.

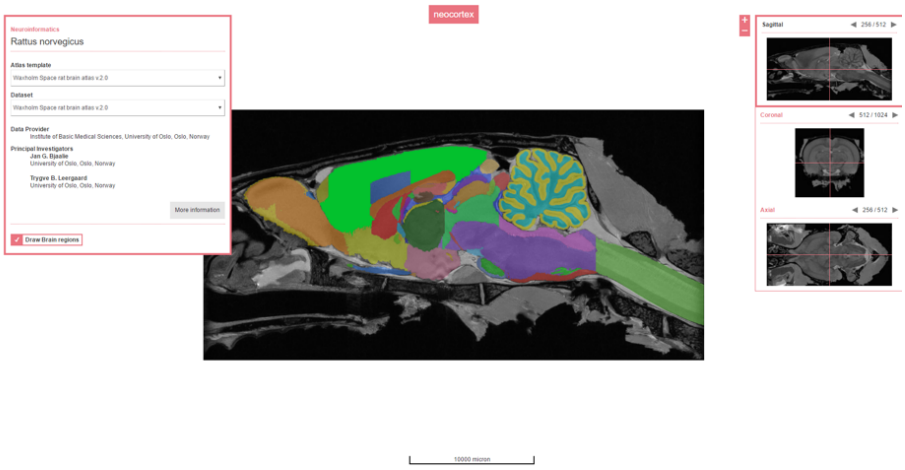

Figure 2: A screenshot of the ATLAS viewer showing an overlayed brain region mask.

3 Data Ingestion

[2] EdN:2

4 Querying the Dataset

[3] EdN:3

5 Outlook & Conclusion

[4] EdN:4

References

[Bau+98] Peter Baumann et al. "The Multidimensional Database System RasDaMan". In: *Proceedings ACM SIGMOD'98*. Seattle, Washington, USA, June 1998.

[EUD16] EUDAT. *EUDAT - Research Data Services, Expertise & Technology Solutions*. 2016. URL: `https://www.eudat.eu/` (visited on 05/27/2016).

[2] EdNote: Write this
[3] EdNote: Write this
[4] EdNote: Write this

[Gmb16] rasdaman GmbH. *the rasdaman raster array database – rasdaman*. 2016. URL: http://www.rasdaman.org/ (visited on 05/27/2016).

[Gro16] The HDF Group. *HDF5*. 2016. URL: https://www.hdfgroup.org/HDF5/ (visited on 05/27/2016).

[Jül16] Forschungszentrum Jülich. *SPM Anatomy toolbox*. 2016. URL: http://www.fz-juelich.de/inm/inm-1/EN/Forschung/_docs/SPMAnatomyToolbox/SPMAnatomyToolbox_node.html (visited on 05/27/2016).

[Lua] Huanxiang Lu. *The BBIC data format*. private communication.

[Lub] Huanxiang Lu. *The BBIC developers*. private communication.

[Pro16a] Human Brain Project. *Atlas Viewer*. 2016. URL: https://nip.humanbrainproject.eu/atlas/ (visited on 05/27/2016).

[Pro16b] Human Brain Project. *The Human Brain Project*. 2016. URL: https://www.humanbrainproject.eu/ (visited on 05/27/2016).