

Development of **NeuroMat** Open Databases [Computational Issues]

Amanda S. Nascimento
Ana Carolina Q. Simões
Claudia D. Vargas
Kelly R. Braghetto

First Workshop of FAPESP's Center of Neuromathematics
January, 2014

Agenda

Computational Issues

- ◆ NeuroMat Database: Its main purposes
- ◆ An Overview of Related Initiatives
- ◆ NeuroMat Database – Development
- ◆ NeuroMat Computational Resources

General Issues

- ◆ Study of Brachial Plexus Injuries
- ◆ Open Data in Neuroscience

Discussion Session

NeuroMat

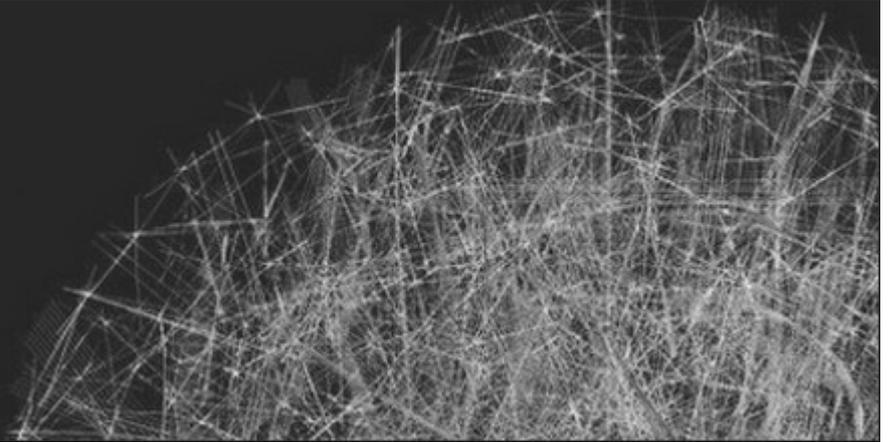
Technology transfer in years 1 and 2

"The first activity of the Center in technology transfer will be the development of a collection of open source tools for basic neuroscience research, databases handling and clinical practice, in particular with respect to diagnostics and rehabilitation [...]."

The initial stage will be gathering typical data, in order to design, implement and test fundamental algorithms for data handling. These will be packaged into reusable containers, mostly libraries and possibly plug-ins for existing software products. [...]

The technology produced by the project will be released as free and open source software in all stages."

NeuroMat Database



Its main purposes are to:

- ◆ Store in an efficient and secure manner all data produced in the project.
- ◆ Support research activities included in the NeuroMat project scope.

Expected Benefits I



- ♦ Facilitate the interaction between the project members.
- ♦ Create “standardized” formats to report experiments, analyses, etc.
- ♦ Support complex queries over project's data.

- ♦ Keep data provenance.
- ♦ Improve efficiency and security in data storage.
- ♦ Support the development of analysis tools.

Expected Benefits II



- Support reproducibility.
- Enable comparison of data across studies.
- Promote meta-analyses.

- Enable data reuse, to generate and test new hypotheses.
- Share with the scientific community all kind of data produced in the project.
 - Create access to data for those who cannot afford to pay for acquisition systems.

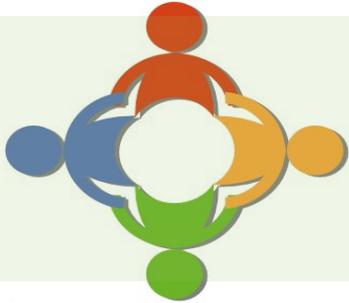
Related Initiatives

There is a growing experience with projects developing individual, often complementary, approaches to data storage and distribution that reflect the present fragmented state of neuroscience data representation.

- Standards to report experiments in neuroscience
- Computational neuroscience
 - Web portals for resource sharing
 - Databases
 - Analysis tools



Web Portals for Neuroscience Resources*



Large-scale projects are usually initiated and carried out by a consortium of research groups in the context of an ambitious research programme.

<https://www.humanbrainproject.eu>



Human Brain Project



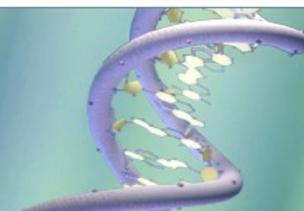
www.loni.ucla.edu/ICBM/

ICBM International Consortium for Brain Mapping

www.birncommunity.org

BIRN
Biomedical Informatics
Research Network

The Conduit for Biomedical Research



(*) Resources → data sets, software tools, materials, etc.

Limitations of Existing Proposals for Neuroscience Databases

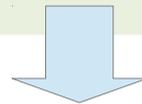
- Databases are seen as mere data repositories that do not necessarily create insights.
 - Inadequate documentation.
 - Unstructured Data.
 - Ineffective solutions for data curation.
-
- Accessing available databases is often over complicated.
 - They are not intuitive enough.
 - They frequently require computer knowledge and additional software installation.
 - They do not provide an infrastructure where heterogeneous databases can be viewed as a unique integrated repository (federation of databases).

What is a Database?

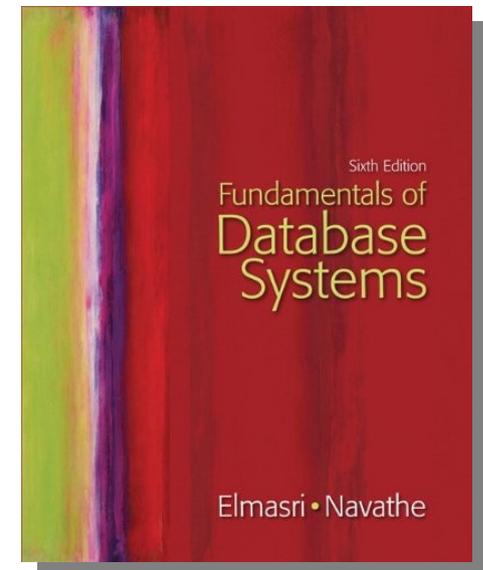
Def 1. “A **database** is a collection of **related data**.”

“By data, we mean **known facts** that can be **recorded** and that have **implicit meaning**.”

Example: Names, telephone numbers, and addresses of the people you know.



These data can be recorded in an indexed address book or it can be stored in a hard drive, using a computer.



What is a Database?

Def 2. “A database has some source (miniworld) from which data is derived, some degree of interaction with events in the real world, and an audience that is actively interested in its contents.”

| ID | Name | Nick Name | Birthday |
|----|-----------------|-----------|----------|
| 1 | Angelina White | Angel | 11-23 |
| 2 | Leonardo Garcia | Leo | 06-09 |
| 3 | Robert Jones | Bob | 02-15 |
| | | | 07-28 |

| PersonID | ID | Address | City | Country | Type |
|----------|----|-----------------|------------|---------|-------------|
| 1 | 11 | 88 Histon Road | Cambridge | UK | Business |
| 1 | 12 | 2 London Road | Cheltenham | UK | Residential |
| 2 | 21 | 182 High Street | Cranleigh | UK | Residential |

| AddressID | ID | Area Code | Phone Number |
|-----------|-----|-----------|--------------|
| 11 | 111 | 1765 | 20 7123 4567 |

Life Cycle of a Database

The **life cycle of a database** is the cycle of **development and changes** that a database goes through during the **course of its life**.



The cycle typically consists of several stages.

Defining

Constructing

Manipulating

Sharing



Defining

Defining

Constructing

Manipulating

Sharing

Defining or **modelling** a database involves specifying the **data types**, **structures**, and **constraints** of the data to be stored in the database.

| Structure | Data Type | Size | Format | Constraint |
|-----------------|------------------------|-----------------|-------------------|--|
| register number | integer number | 5 digits | Ex: 12345 | uniqueness in the university |
| name | sequence of characters | Max.: 30 digits | Ex: "John Lennon" | - |
| birth date | date | - | Ex: 12/24/1991 | Min. Value = 01/01/1900 Max. Value = 01/01/2014 |
| ... | ... | ... | ... | ... |

→ Data requirements are collected from users (e.g., people that produce or use the data).

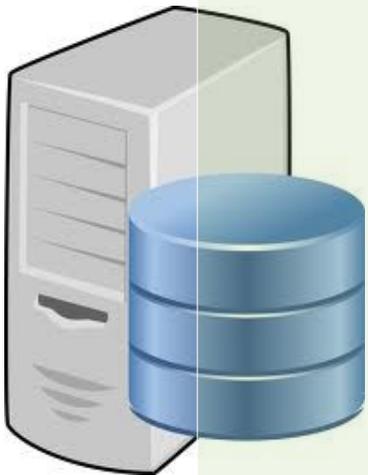
Constructing



Constructing the database is the process of **storing** the data on some **storage medium**.

Related concerns:

- Security (data integrity, access control).
- Efficiency (fault tolerance, replication).
- Periodic backups.



Manipulating



Manipulating a database includes functions such as:

- **Querying** the database to retrieve specific data.
- **Updating** the database to reflect changes in the miniworld.
- **Generating reports** from the data.

| | Reference | Forename | Surname | Address1 | Town | Coun |
|--------------------------|-----------|-----------|---------|---------------------|--------------|---------------|
| <input type="checkbox"/> | 1372 | Elizabeth | House | 521 Etiam Av. | New Rochelle | Angus |
| <input type="checkbox"/> | 1373 | Abel | Dunn | 903-5759 Magna Road | Rohnert Park | Co. Waterford |
| <input type="checkbox"/> | 1374 | Driscoll | Russo | 238-4308 Orci Av. | Hannibal | Northumberla |

Data are stored in or recovered from a database through the use of software tools specially developed to perform these tasks.

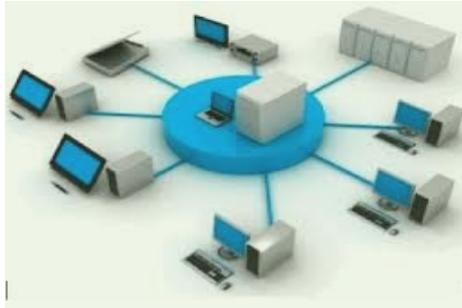
Sharing

Defining

Constructing

Manipulating

Sharing



Sharing a database allows multiple users and programs to **access the database simultaneously**.



Overview of the NeuroMat Database

Various Organizational Data

People and their affiliation.

Team members.

Working groups.

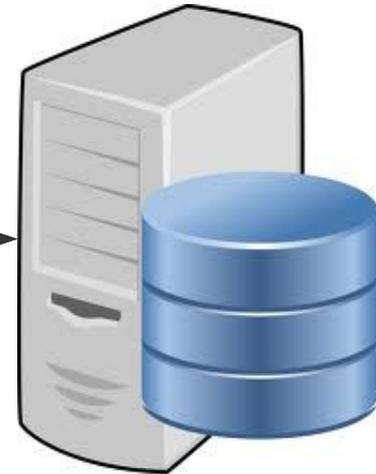
Projects.

Neuroscience Data

“Raw” Data.

Processing tools and derived data.

Other documents (articles, reports, etc.).



Examples of Data in NeuroMat

Data acquired from experiments:

- Electrophysiological (EEG, TMS, EMG, etc.).
- Neuroimaging (MRI, fMRI, etc.).



Important: not only “raw” data, but also provenance data → metadata

Derived data, generated by processing tasks (filtering, transformation, analysis, etc.).



Important: not only the derived data, but also information about the process used to derive the data.

Data Provenance

Frequently asked questions for Scientists: *



Where was a document found?

How was this data set produced?

Were all facts included in this decision?

Were all the latest figures included in this diagram?

Can this scientific experiment be reproduced?

“**Data provenance** covers the provenance of computerized data.

There are two main aspects of data provenance:

ownership of the data and data usage.”

Data Provenance in NeuroMat

We want to **keep track of the provenance information of all data and software tools produced in the project.**



These information are fundamental to enable researchers to make a correct use of the resources produced in NeuroMat.

Example of Data Provenance in NeuroMat

Provenance information of an human EEG signal:

- Acquisition system (equipment model, manufacturer, software, ...).
- Equipment setting (sampling rate, amplifier filter, ...).
- Electrode placement system (international 10-20 system, ...).
- Size of the electrode cap (S, M, L).
- Information about the protocol of the experiment.
- Information about who conducted the experiment (affiliation, research team, ...).
- Information about the subject of the experiment (gender, age, medical records, ...).

Development of the NeuroMat Database

- We divided the database of **NeuroMat** in two big modules:
 - **Module of “raw” data.**
 - Module of derived data.



Important: we are currently working on the design/construction of the first module.

- **Development approach:**
 - gather data requirements of one laboratory at time;
 - consider the (good) work already done in other related initiatives.

First case study (foundations of **NeuroMat** database):

Laboratory of Neuroscience and Rehabilitation
Institute of Neurology Deolindo Couto
Federal University of Rio de Janeiro

<http://controlemotor.com.br/indc-npnr/>



Standards to Report Experiments in Neuroscience

The Minimum Information for Biological and Biomedical Investigations (MIBBI) project*:

“promotes extant efforts developing minimum information (MI) guidelines for the reporting of biological and biomedical science to the wider community.”

- ♦ Examples of MI guidelines under MIBBI project:
 - ♦ **MINI** – for neuroscience investigations
 - ♦ **MINEMO** – for event-related potential (ERP)/EEG data
 - ♦ **MIfMRI**** – for fMRI studies

*<http://www.biosharing.org/standards/mibbi>

**<http://www.fmrimethods.org/>

The CARMEN* Project and the MINI** Guideline

Minimum Information about a Neuroscience Investigation (MINI): Electrophysiology

Frank Gibson^{*1}, Paul G Overton², Tom V Smulders³, Simon R Schultz⁴, Stephen J Eglén⁵, Colin D Ingram⁶, Stefano Panzeri⁷, Phil Bream⁴, Evelyne Sernagor⁶, Mark Cunningham⁶, Christopher Adams⁶, Christoph Echtermeyer⁸, Jennifer Simonotto¹, Marcus Kaiser¹, Daniel C Swan⁹, Martyn Fletcher¹⁰, Phillip Lord¹

The following section, detailing the reporting requirements for the use of electrophysiology, is subdivided as follows:

1. General features
2. Study subject
3. Task
4. Stimulus
5. Behavioral event
6. Recording
7. Time series data

...

Reporting requirement for electrophysiology

1. General features

- (a) Date and time
- (b) Responsible person or role
- (c) Experimental context
- (d) Electrophysiology type

2. Study subject

- (a) Genus
- (b) Species
- (c) Strain
- (d) Cell line
- (e) Genetic characteristics
- (f) Genetic variation
- (g) Disease state
- (h) Clinical information
- (i) Sex
- (j) Age
- (k) Development stage

Checklist that identifies the minimum information required to report the use of electrophysiology in a neuroscience study.

* Code Analysis, Repository & Modelling for E-Neuroscience: <http://www.carmen.org.uk/>

** <http://www.carmen.org.uk/standards/mini.pdf>

The NEMO* Project and the MINEMO** Guideline

Minimal Information for Neural Electromagnetic Ontologies (MINEMO): A standards-compliant method for analysis and integration of event-related potentials (ERP) data.

Frishkoff G, Sydes J, Mueller K, Frank R, Curran T, Connolly J, Kilborn K, Molfese D, Perfetti C, Malony A.

Subset of MINEMO terms that are required to save data to the NEMO portal (in addition to unique ID for each table)

1. Research lab (General Features)
 - a. Institution
 - b. Principal investigator (PI)
2. Experiment (General features)
 - a. Experiment paradigm(s)
3. Publication
 - a. Publication type
 - b. DOI or File location (Path)
4. Study subjects (Group characteristics)
 - a. Diagnostic classification
 - b. Genus
 - c. Species
 - d. Age (average)
 - e. Gender (#male, female subjects)
 - f. Handedness (#RH, LH subjects)
 - g. Native language (modal)
5. Experiment condition
 - a. Experiment condition
 - b. Experiment task (Instructions)
6. Stimulus presentation
 - a. Target stimulus type
 - b. Target stimulus modality
7. Behavioral data collection
 - a. Response type
 - b. Response modality
8. EEG Data collection
 - a. Electrode array (Layout)
 - b. Sampling rate
9. EEG/ERP Data preprocessing
 - a. ERP event
 - b. ERP epoch length (in ms)
 - c. ERP baseline (pre-target) duration
 - d. Offline reference
10. EEG/ERP Data file
 - a. Data file contents (EEG data type)
 - b. Data file format
 - c. Data file location (URI)

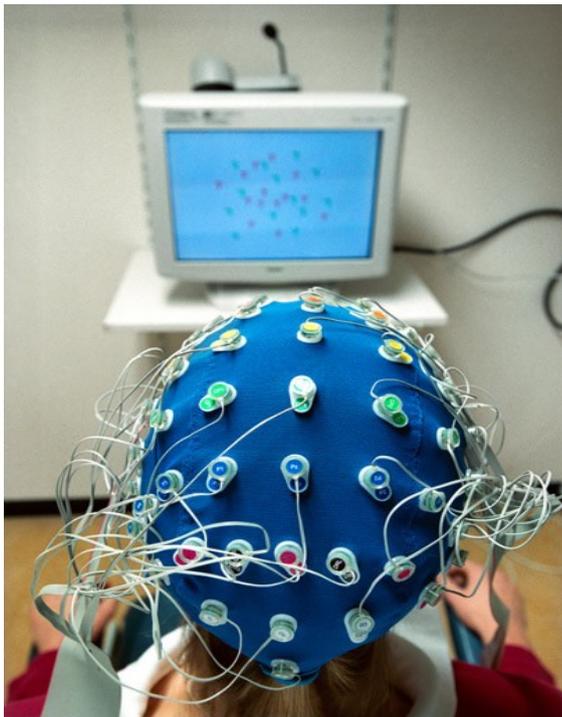
- ◆ MINEMO extends MINI (Minimal Information for Neuroscience Investigations) to the ERP domain.
- ◆ Checklist terms are explicated in NEMO, a formal ontology that is designed to support ERP data sharing and integration.

* Neural ElectroMagnetic Ontologies: <http://nemo.nic.uoregon.edu/wiki/NEMO>

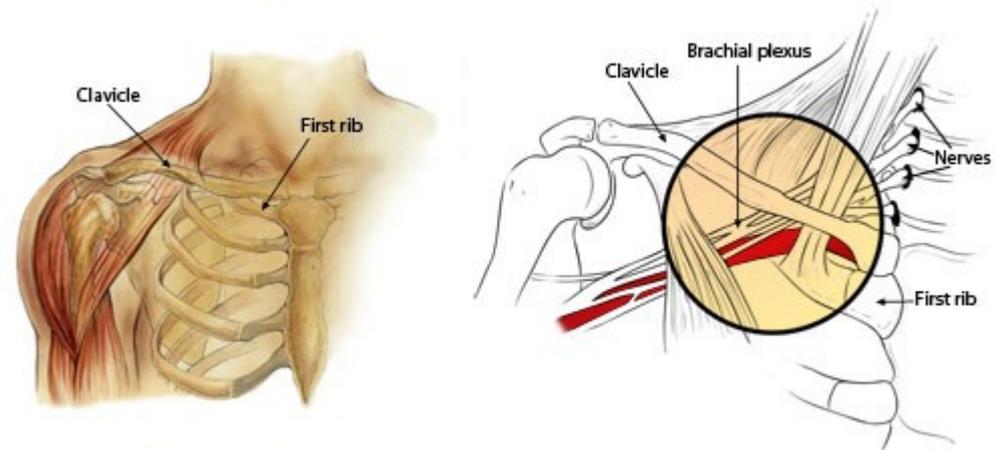
** <http://www.ncbi.nlm.nih.gov/pubmed/22180824>

Data Requirements at INDC

- Data collected in the experiments conducted by Claudia's team at INDC:
 - Experiments involving EEG, TMS, EMG, and Stabilometry
 - Patients with brachial plexus injuries



EEG



Brachial plexus

NeuroMat Database Model – Overview

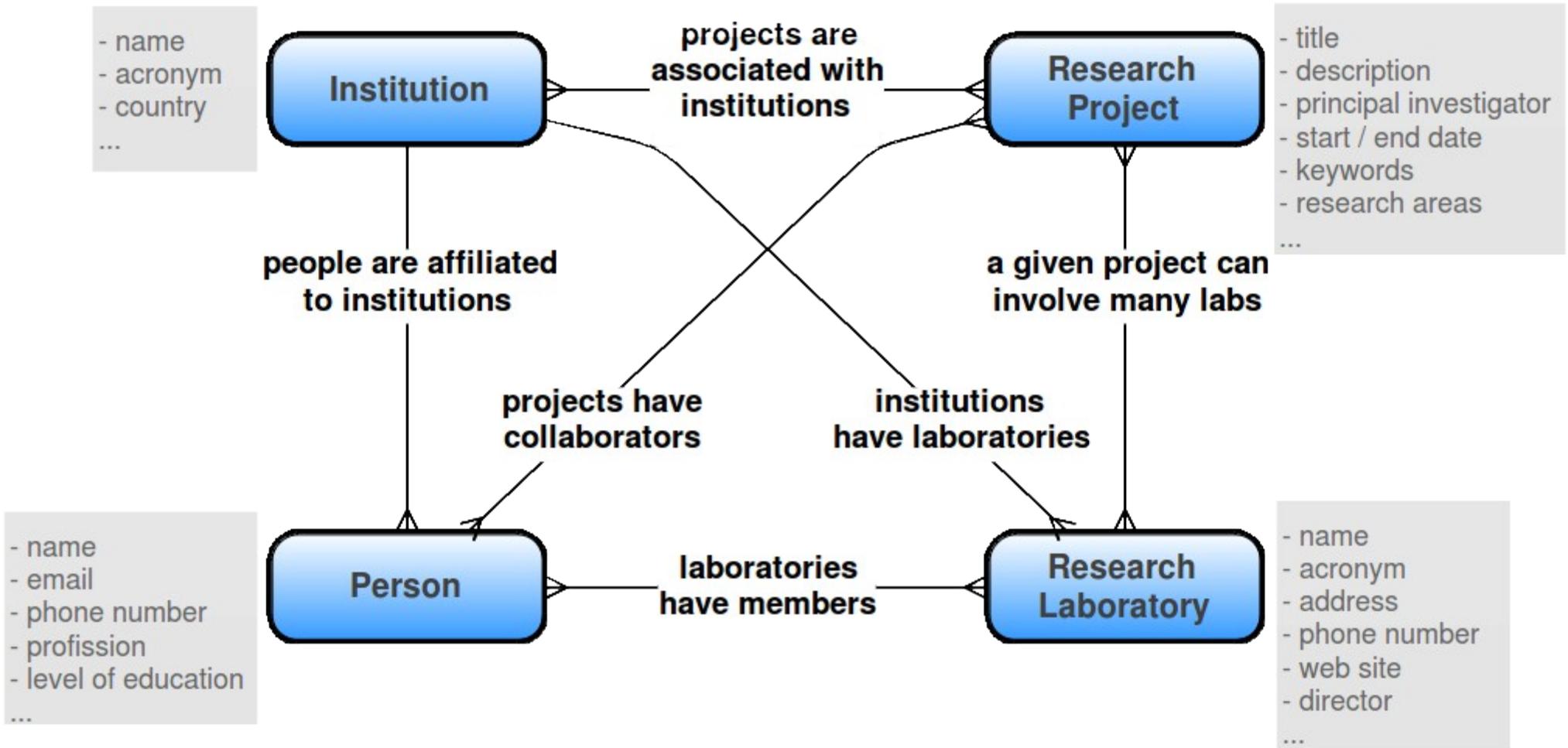
Current state – 5 modules:

- Organizational Structure.
- Experiment Protocol.
- Electrophysiological Data Acquisition.
- Behavioral Data Acquisition.
- Documents.

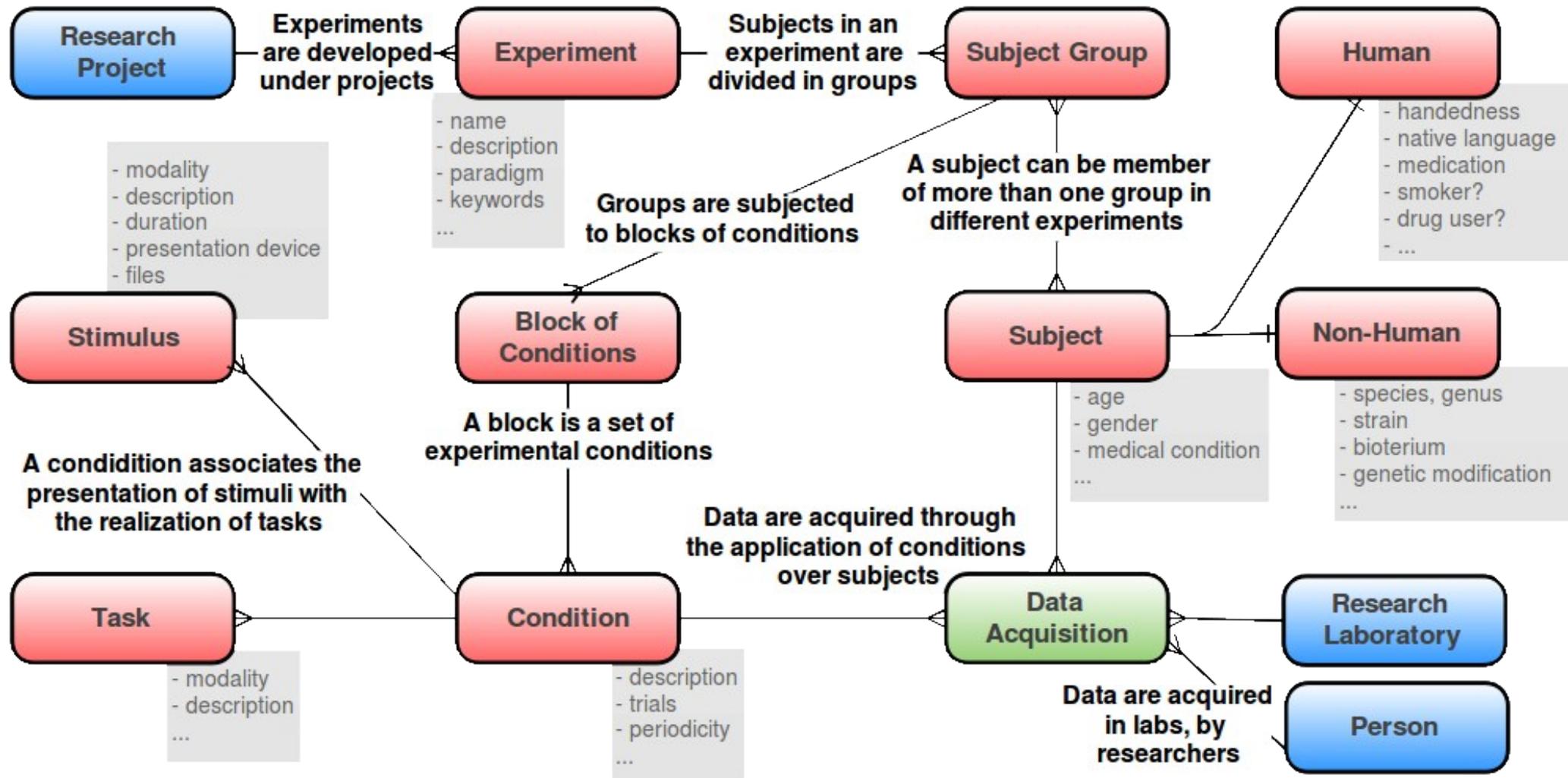
Other previewed modules:

- Histopathology Data Acquisition.
- Molecular Data Acquisition.
- Neuroimaging Data Acquisition.
- Derived Data.

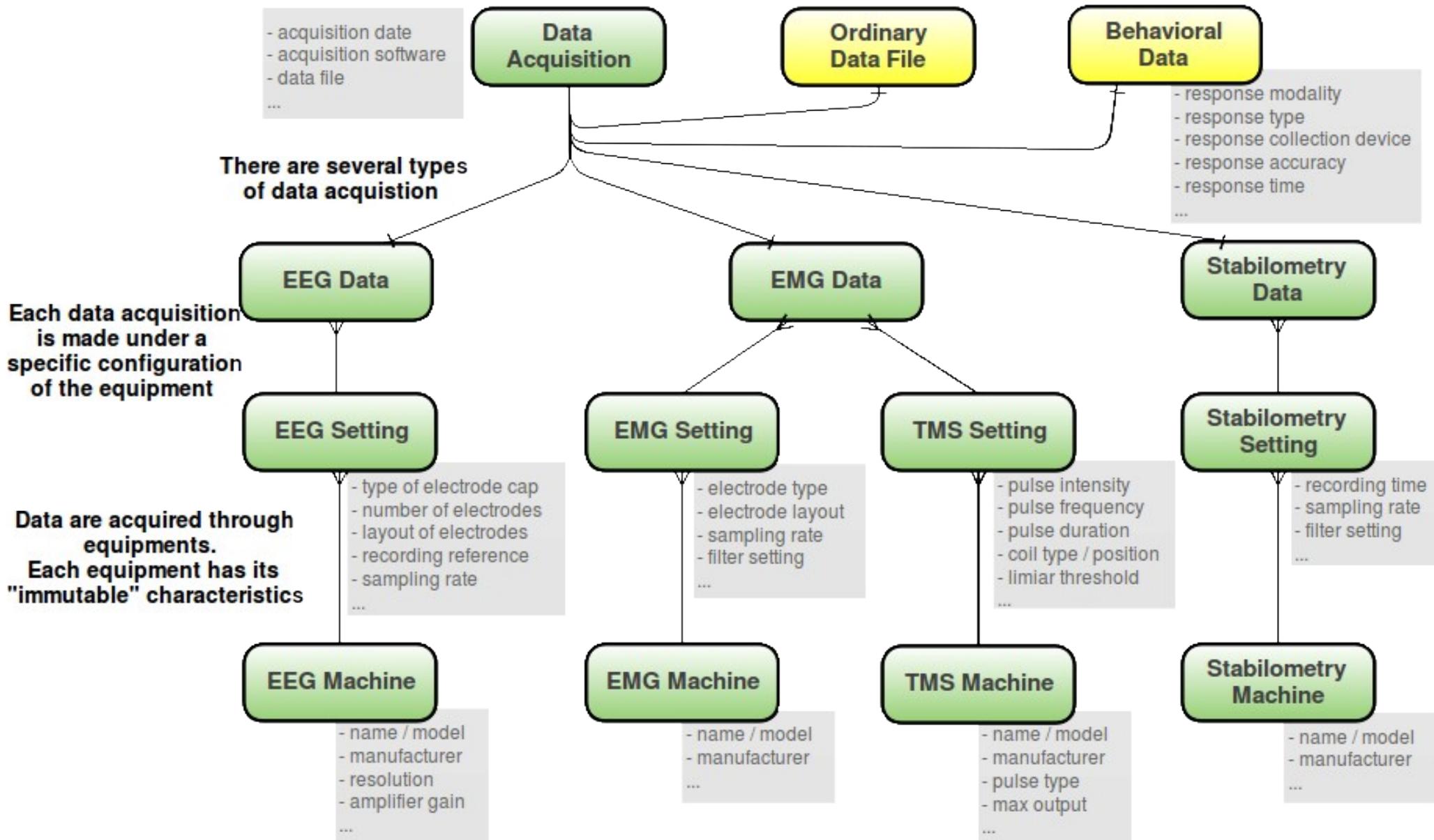
Organizational Structure Module



Experimental Protocol Module



Electrophysiological and Behavioral Data Acquisition Modules



Study-Specific Data

The structure of the database modules can “accommodate” an important portion of all data that can be collected in an electrophysiological experiment.

→ Data whose structure is common for all experiments. In other words, all data that can be described in terms of the standardized structure defined by the database model.



Problem: Experiments may result in other kinds of data, with variable structure. Generally, these data are study-specific, collected by means of questionnaires.

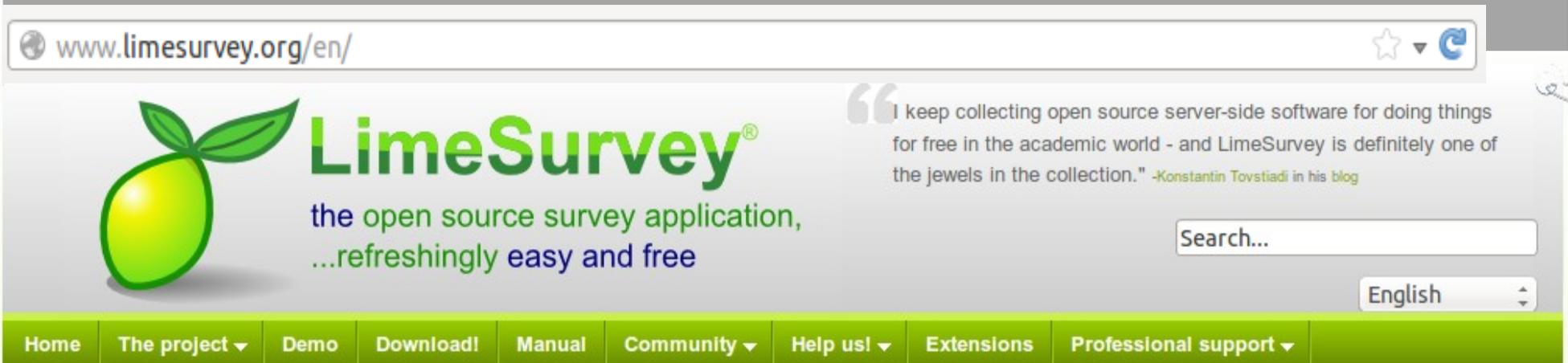
Example: Each study conducted at INDC gathers data from the volunteers by means of questionnaires designed by the researchers responsible for the study.

Computational solution proposed to
digitalize study-specific data:
→ Digital Questionnaires

Digital Questionnaires

- There exist several software systems that enable users to create digital questionnaires and make them available online.
- Some of these systems are very “powerful”:
 - Rich set of question structures and presentation formats
 - Data collected through the questionnaires can be stored in local databases, hosted in “private” servers → improved security

In NeuroMat, we are using a free, open source questionnaire system:

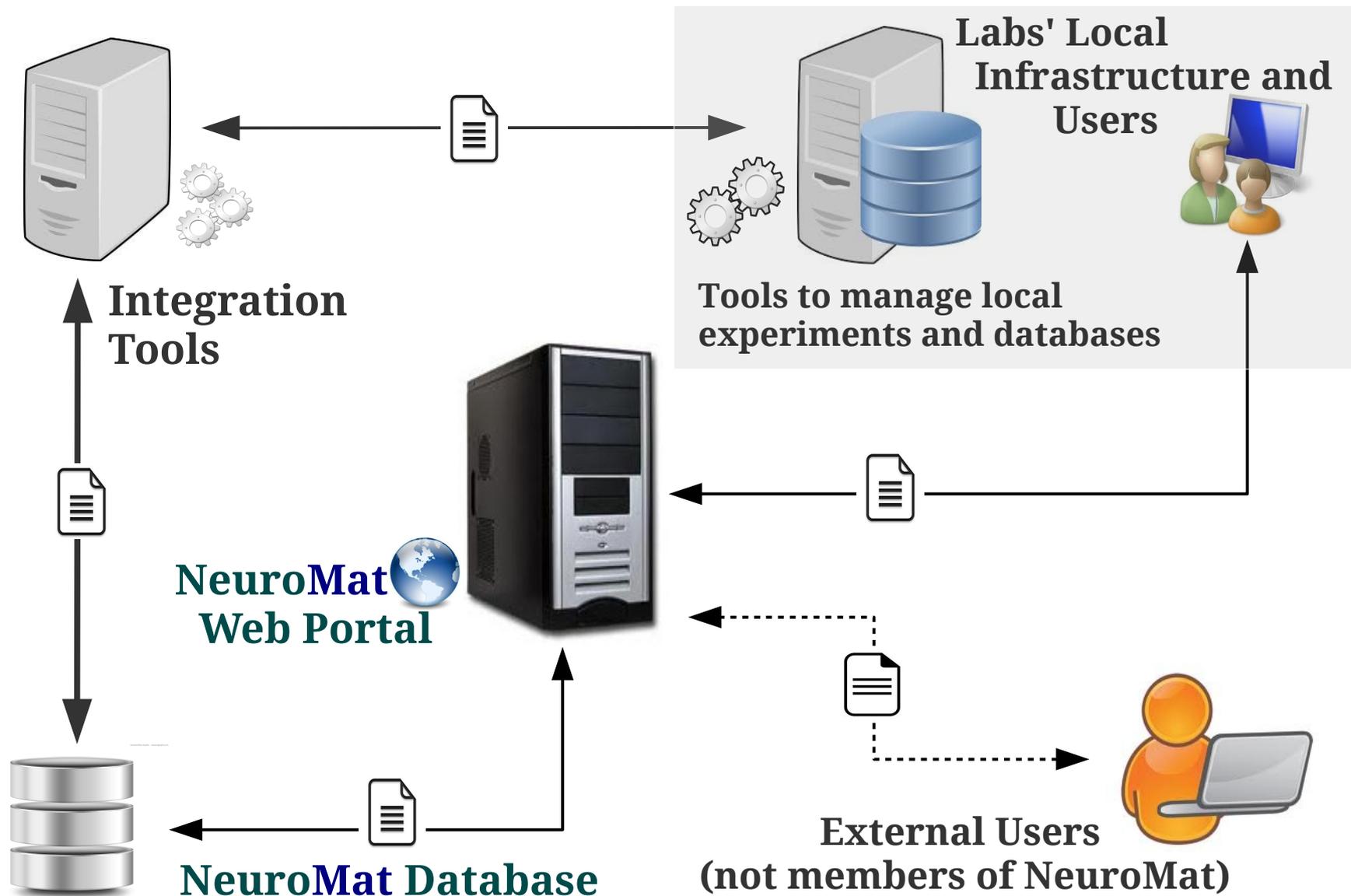


The screenshot shows the LimeSurvey website homepage. At the top, the address bar displays www.limesurvey.org/en/. The main content area features the LimeSurvey logo, which is a green lime with two leaves, followed by the text "LimeSurvey®" in a large green font. Below the logo, it says "the open source survey application, ...refreshingly easy and free". To the right of the logo, there is a quote: "I keep collecting open source server-side software for doing things for free in the academic world - and LimeSurvey is definitely one of the jewels in the collection." attributed to "-Konstantin Tovstiadi in his blog". Below the quote is a search bar with the text "Search..." and a language dropdown menu set to "English". At the bottom, there is a green navigation bar with the following links: Home, The project ▾, Demo, Download!, Manual, Community ▾, Help us! ▾, Extensions, and Professional support ▾.

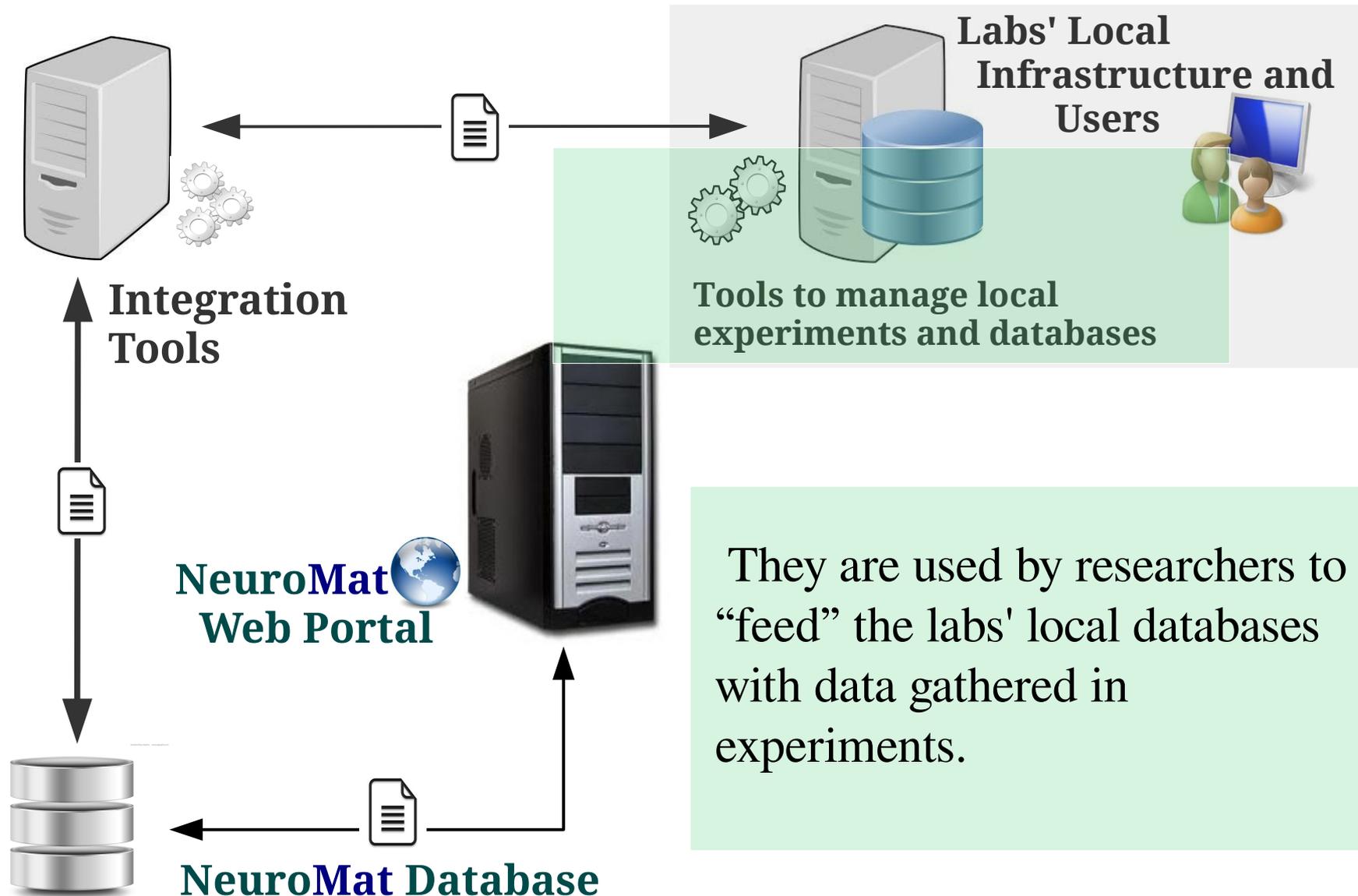
Next Steps

- ◆ To extract data requirements from other research groups.
- ◆ To add in the model neuroimaging experiments.
- ◆ To add in the model “derivated data” (and their provenance information).
- ◆ To develop the software tools that will interact with the **NeuroMat** database.

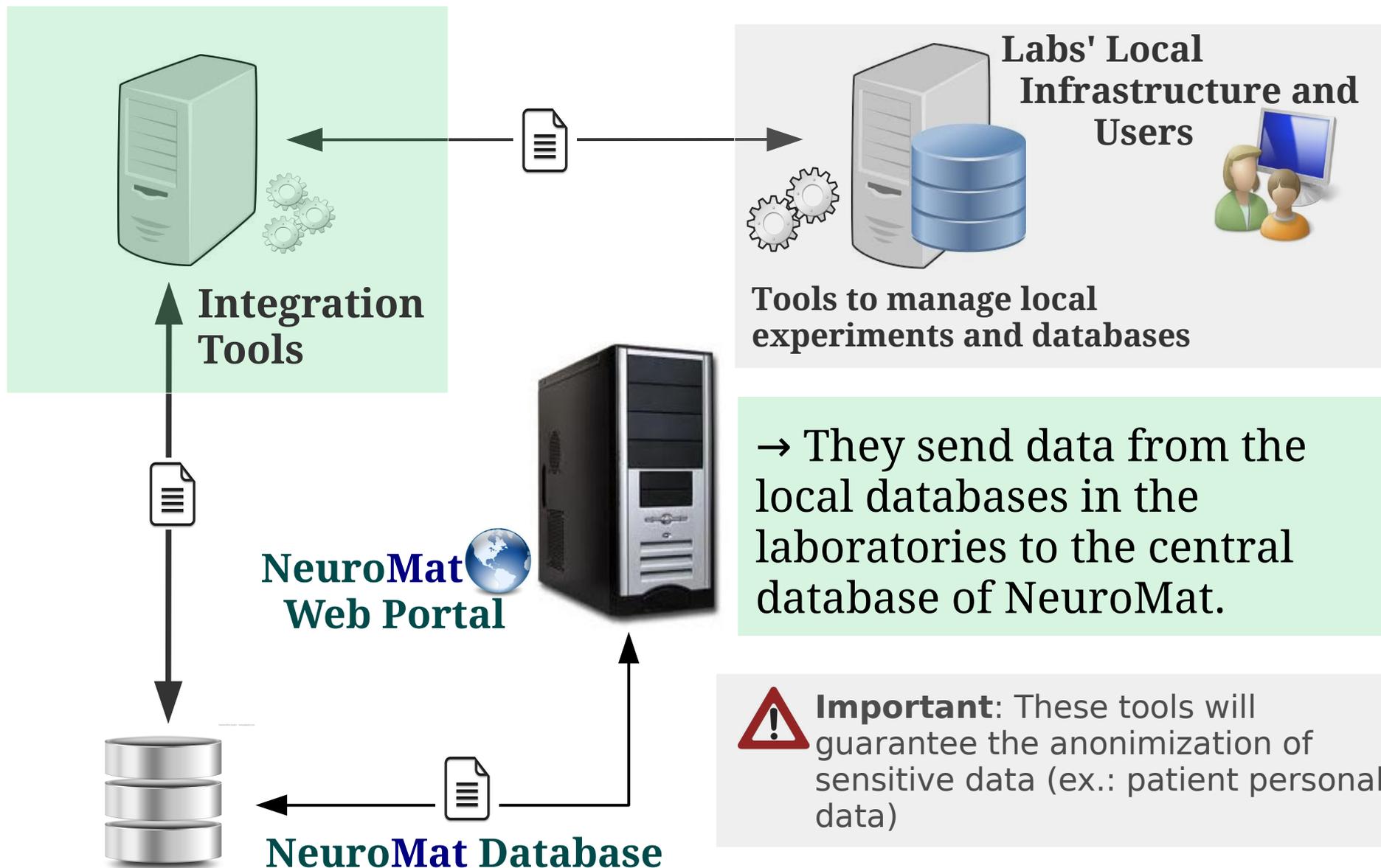
An Overview of NeuroMat Computational Resources



Tools to Manage Local Experiments and Databases

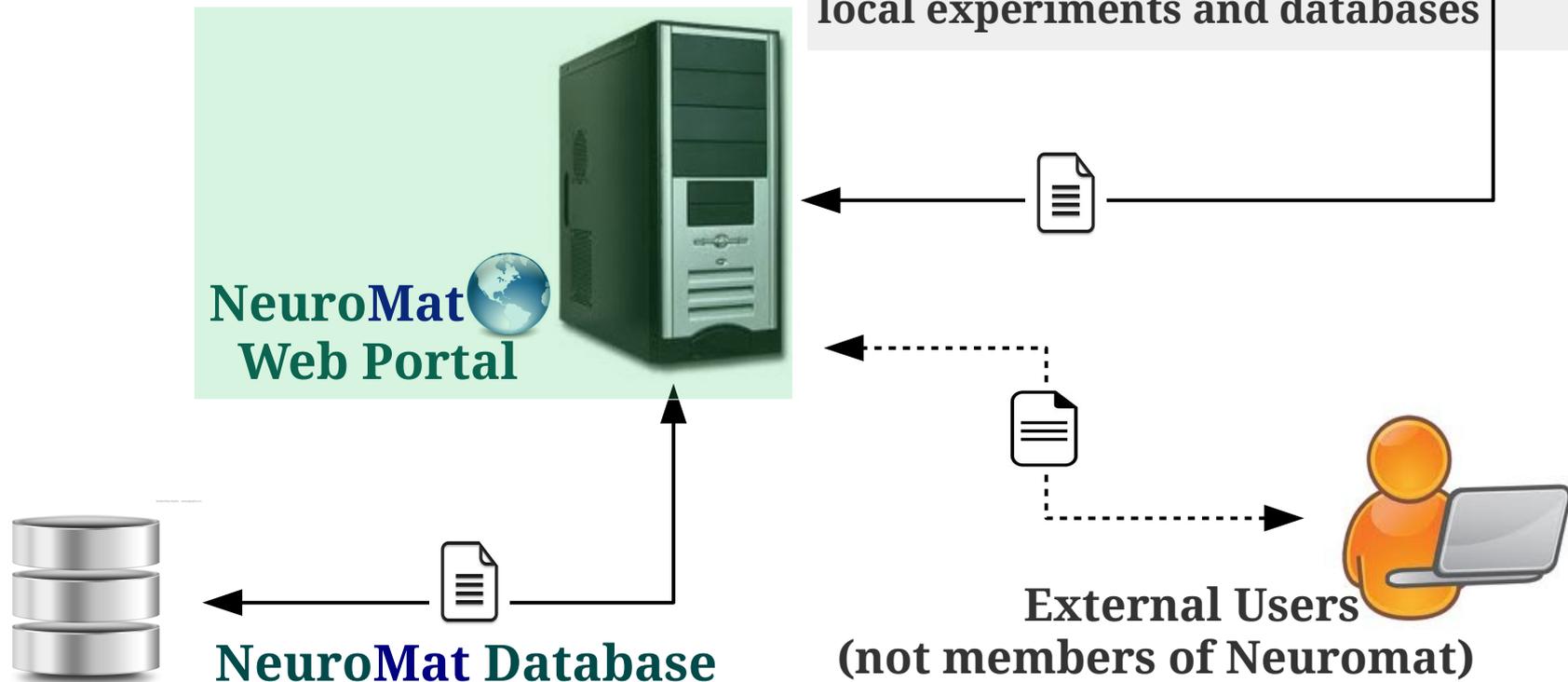


Integration Tools



NeuroMat Web Portal

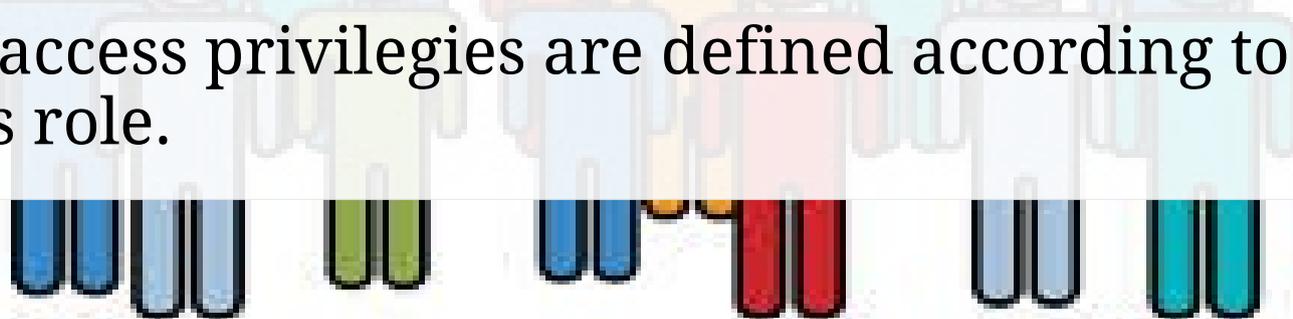
- Search engine, to support complex queries over data
- Main display for the project results



Authorization Strategy (Access Control) for the NeuroMat Web Portal



- ♦ At a first moment, the NeuroMat database will be used mainly to facilitate the **interaction between the project members**.
- ♦ Different classes of users, each one representing a different role in the project.
- ♦ Data access privileges are defined according to the user's role.



Design and Development – **NeuroMat** Database

Team

- ♦ Amanda S. Nascimento (NUMEC – USP)
- ♦ Ana Carolina Q. Simões (CECS – UFABC)
- ♦ Carlos Ribas (NUMEC – USP)
- ♦ Fabio Kon (IME – USP)
- ♦ Kelly R. Braghetto (IME – USP)

Collaborators (up to now)

- ♦ Claudia D. Vargas (INDC – UFRJ) & her team
- ♦ André F. Helene (IB – USP) & his team
- ♦ Gilberto Xavier (IB – USP) & his team

This project is supported by:



