

## **Data Management in NeuroMat and the *Neuroscience Experiments System (NES)***

**Kelly Rosa Braghetto**

Department of Computer Science

Institute of Mathematics and Statistics - University of São Paulo

**1st NeuroMat Young Researchers Workshop**

**May, 2015**



"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."



**Data management software**

**Data analysis software**

**FREE SOFTWARE**



**databases**



**OPEN DATA**

# Open Data

- Open data is more than “publicly available data”
- According to the definition of *Open Knowledge Foundation*:  
*“Open means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness).”*
- To be reusable, data must
  - **be “understandable”**
    - with descriptions of their **structure**
  - **have high quality**
    - with information about their **provenance**  
(how, when, where and by who the data were generated)

# Open Scientific Data

- Some scientific publishers started to condition article publication to the public release of study data
    - Ex.: BioMed Central, PLOS One
- <http://www.biomedcentral.com/about/opendata>
- <http://blogs.plos.org/everyone/2014/02/24/plos-new-data-policy-public-access-data-2/>
- Research projects funded by *São Paulo Research Foundation - FAPESP* must make their data and software publicly available as open data and free, open-source software.
  - **Benefits**
    - Possibility of validation and reproduction of the results
    - Science of better quality and greater impact

# Open Data as a Civic Duty

- **UK: Freedom of Information Act – FOIA, and the Environmental Information Regulation – EIR**
  - Guarantee to every citizen the right to access to information that is held by public local institutions - which includes research data from universities and other research institutions funded by public money
- **Brazil, Decree No. 7,724, from May 16, 2012, known as the Law on Access to Information**
  - Governs the access to information produced or held by governmental, federal organizations and agencies, including federal universities and funding agencies for research (eg, CAPES and CNPq)

## **Expected results** →

- Government transparency
- Dissemination of scientific knowledge
- Fostering of science & technology

# Open Data from the Legal Perspective

- Only recently licenses originally conceived for free software and content began to be adapted for use in databases
- The more widely used licenses for sharing open scientific databases are:

- *Creative Commons* (CC) <http://creativecommons.org/>



- *Open Data Commons* (ODC) <http://opendatacommons.org/>



- These licenses allow, for example:
  - to limit the use of data to those who credit authors or providers
  - to establish that data redistributions can only be made with the same or equivalent license as the original data used

# Scientific Datasets Should Be Really Open!

- Copyright laws do not apply to some types of databases.
- In general, **copyright applies to DBs whose content is the result of some intellectual effort or whose compilation/organization consumes significant resources of time or money**
  - DBs with purely factual information, without an original organizational structure, cannot be protected under copyright or licensing
    - This is the case of a large number of DBs in biological domains
- To eliminate legal barriers, facilitate sharing and maximize data reuse, projects that are dedicated to open data sharing recommend that **scientific datasets should be placed in the public domain by waiving all author's and provider rights.**



# For More Information about Open Data ...

- **Digital Curation Centre**

<http://www.dcc.ac.uk/>



- **Open Knowledge Foundation**

<https://okfn.org/>

<http://br.okfn.org/>



- **Brazilian Working Group on Open Science**

<http://www.cienciaaberta.net/>

- Kelly R. Braghetto, “**Open Data in Science: a NeuroMat op-ed**”, NeuroMat Newsletter – July, 2014

<http://neuromat.numec.prp.usp.br/content/open-data-science-neuromat-op-%C2%ADed>

# Open Scientific Data – Main Challenges

- Lack of **standards** for data representation
  - Lack of **consensus** about what should be stored
- Data **provenance** registry
- Scientific community's **resistance**
  - collecting data in experiments is difficult and costly, and often poorly recognized by the peers
- Proper **attribution** to researchers and institutions for the databases they provide; **protection** of their interests regarding the use of data
- Data **curation** (which depends on people and equipments)

# An interesting recent reference about the topic...



NATURE | CORRESPONDENCE



## Data curation: Act to staunch loss of research data

Andrew Gonzalez & Pedro R. Peres-Neto

Affiliations | Corresponding author

*Nature* **520**, 436 (23 April 2015) | doi:10.1038/520436c

Published online 22 April 2015

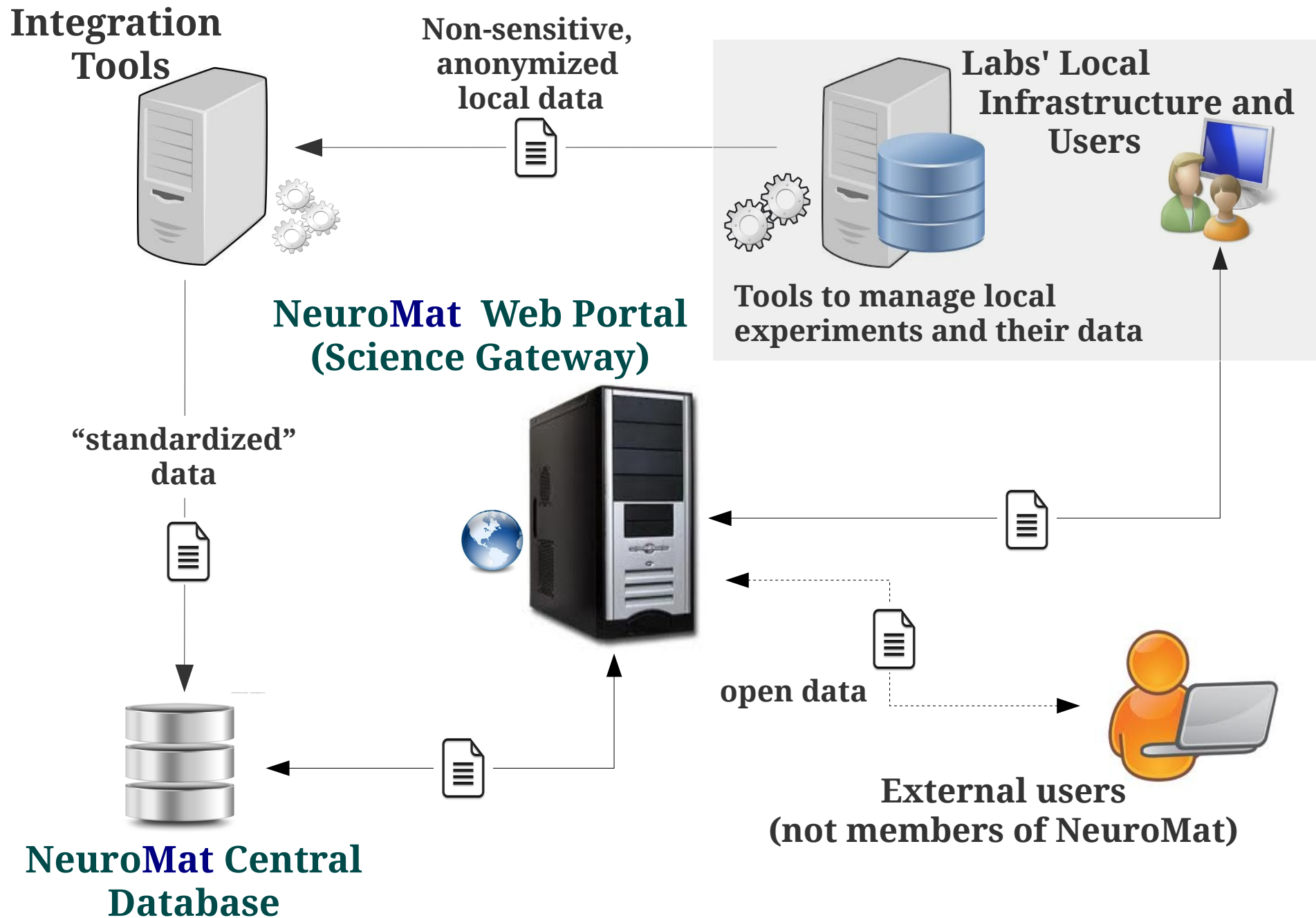


**Subject terms:** [Databases](#) • [Research management](#)

Never before have scientists had the ability to generate and collect so much data — recent estimates suggest that the global scientific output is doubling roughly every decade (see L. Bornmann and R. Mutz, preprint at <http://arxiv.org/abs/1402.4578v3>; 2014, and [go.nature.com/nzejwh](http://go.nature.com/nzejwh)). It is alarming, therefore, that the odds of data being lost are estimated to increase by 17% in every year after publication (T. H. Vines *et al. Curr. Biol.* **24**, 94–97; 2014). And this does not include the 80% or so of research data that are inaccessible or unpublished (B. P. Heidorn *Libr. Trends* **57**, 280–299; 2008).

<http://www.nature.com/nature/journal/v520/n7548/full/520436c.html>

# NeuroMat Databases & Software Tools



# Storing and Sharing Neuroscience Resources

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**

**The related initiatives can be classified as:**

- Standards to report experiments in neuroscience
- Computational neuroscience
  - Web portals for resource sharing (*science gateways*)
  - 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 program.
- They share resources such as data sets, software tools, and documents.

www.birncommunity.org









https://www.humanbrainproject.eu



Human Brain Project



# List of Neuroscience Databases (Wikipedia)

Name and External Link	Description	Organism	Level (gene, neuron, macroscopic)	Data (MRI, fMRI, images, descriptive, numerical)	Disorder	Register to view data?	Ref
<b>Allen Brain Atlas</b> [1] 	Atlas, stained sections from brains showing development and gene expression	Mouse, Human	Macroscopic, Gene	Images	Healthy	No	[2]
<b>Alzheimer's Disease Neuroimaging Initiative (ADNI)</b> [2] 	Structural MRI images	Human	Macroscopic	MRI datasets	Healthy and <a href="#">Alzheimer's Disease</a>	Yes	[3]
<b>BIRN fMRI and MRI data</b> [3] 	fMRI, MRI scans and atlases for human and mouse brains	Mouse, Human	Multilevel: brain regions, connections, neurons, gene expression patterns	MRI datasets, fMRI datasets	healthy, <a href="#">elderly</a>	No	
<b>Bipolar Disorder Neuroimaging Database</b> [4] 	Meta-analysis and database of MRI studies	Human	Macroscopic	Descriptive, numerical	<a href="#">Bipolar Disorder</a>	No	[4]
<b>Brain Architecture Management System</b> [5]  [6] 	Online resource for information about neural circuitry	Rat, Mouse, Human	Multilevel: brain regions, connections, neurons, gene expression patterns	Descriptive, numerical	healthy	No	
<b>Brain Cloud</b> [7] 	Gene expression in the human prefrontal cortex	Human	Gene expression patterns	Descriptive, numerical	healthy	No	
<b>Brain-Development.org</b> [8] 	Structural MRI images and Atlases	Human	Macroscopic	MRI datasets	Fetuses, healthy and prematurely born neonates	No	[5]

[http://en.wikipedia.org/wiki/List\\_of\\_neuroscience\\_databases](http://en.wikipedia.org/wiki/List_of_neuroscience_databases)



# Common Problems in Open (Neuroscience) Databases

- Poor quality: inconsistent data, “outdated” data, etc.
- Insufficient documentation; lack of metadata
- Lack of standardization in data representation
- Unstructured data
- Overcomplicated access control to data
- Requirement of specific computer knowledge and additional software installation
- Database as mere data repository (= “federation of datasets”)
  - Datasets with different levels of quality
  - Datasets with different structures
  - Lack of an infrastructure where heterogeneous datasets can be viewed as a unique integrated repository

...



## Do brain image databanks support understanding of normal ageing brain structure? A systematic review

David Alexander Dickie • Dominic E. Job • Ian Poole •  
Trevor S. Ahearn • Roger T. Staff • Alison D. Murray •  
Joanna M. Wardlaw

Received: 25 October 2011 / Revised: 5 December 2011 / Accepted: 29 December 2011 / Published online: 22 February 2012  
© European Society of Radiology 2012

Eur Radiol (2012) 22:1395–1396  
DOI 10.1007/s00330-012-2408-3

## Making better use of our brain MRI research data

Frederik Barkhof

# Standards to Report Experiments in Neuroscience

- **Minimum Information for Biological and Biomedical Investigations – MIBBI project**

<https://www.biosharing.org/standards/mibbi>

- “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

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

- **MINEMO** – for event-related potential (ERP)/EEG data

[www.ncbi.nlm.nih.gov/pubmed/22180824](http://www.ncbi.nlm.nih.gov/pubmed/22180824)

- **MifMRI** – for fMRI studies

[http://www.fmrimethods.org/index.php/Main\\_Page](http://www.fmrimethods.org/index.php/Main_Page)

# Example: the MINI Guideline

## Minimum Information about a Neuroscience Investigation (MINI): Electrophysiology

Frank Gibson<sup>\*1</sup>, Paul G Overton<sup>2</sup>, Tom V Smulders<sup>3</sup>, Simon R Schultz<sup>4</sup>, Stephen J Egle<sup>5</sup>, Colin D Ingram<sup>6</sup>, Stefano Panzeri<sup>7</sup>, Phil Bream<sup>4</sup>, Evelyne Sernagor<sup>6</sup>, Mark Cunningham<sup>6</sup>, Christopher Adams<sup>6</sup>, Christoph Echtermeyer<sup>8</sup>, Jennifer Simonotto<sup>1</sup>, Marcus Kaiser<sup>1</sup>, Daniel C Swan<sup>9</sup>, Martyn Fletcher<sup>10</sup>, Phillip Lord<sup>1</sup>

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

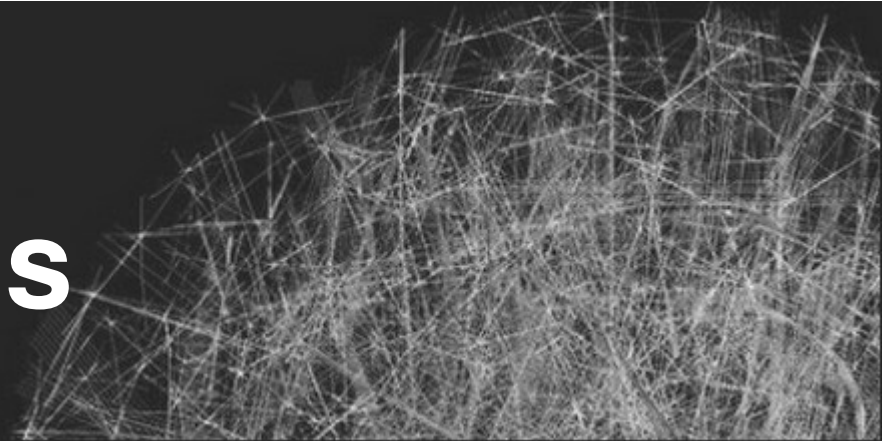
- (a) Date and time
- (b) Responsible person or role
- (c) Experimental context
- (d) Electrophysiology type

### 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.**

# NeuroMat Databases



## Their 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;
- ♦ Gather high quality data which will be made publicly available in a near future.

# Overview of NeuroMat Databases

## Organizational Data

- People and their affiliation
- Laboratories, Projects
- Team members, Working groups

## Experimental Data

- **“Raw” data**
  - Includes provenance data (metadata)
- Derived data
  - Includes information about the process used to derive the data
- Documents (articles, reports, etc.)



# NeuroMat Databases – Development Approach

- Gather data requirements of one laboratory at time
- Consider the (good) work already done in other related initiatives
  - MIBBI guidelines to report experiments
  - Structure of other open databases
- **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/>

# The *Neuroscience Experiments System* (NES)

- **Free, open-source software for the management of clinical and experimental neurophysiological data**  
<https://github.com/neuromat/nas>
- Main functionality: **works as “an access port” to a Lab's local database**
- Under development in NeuroMat since the beginning of 2014
- **Current version: 0.2 (beta) – will be released in May, 2015**
  - Tailored to the INDC-UFRJ needs
  - But easily adaptable to other research labs

- **The currently available modules of NES enable users to manage data from**
  - Patient Record
    - Patient Registration
    - Socio-demographic data and social history
    - Medical records and complementary exams
  - Experiments
    - Subject Groups (with Humans)
    - Experimental Protocols
  - Questionnaire Administrations (integrated to *Lime Survey*)
  - Users and Access Control



- **The new modules will be related to:**
  - Experiments
    - Electrophysiological data acquisition
    - Neuroimaging data acquisition
    - Behavioral data acquisition
    - Subject groups with non-humans
  - Data recovery & visualization
  - Laboratory Organizational Structure
  - Derived Data
    - Analysis processes (= scientific workflows)
- **Other kinds of neuroscience resources that we should consider to store:**
  - Neuron models (as suggested by Prof. Dr. Antonio Roque yesterday!)
  - [Do you have other suggestions? ;) ]

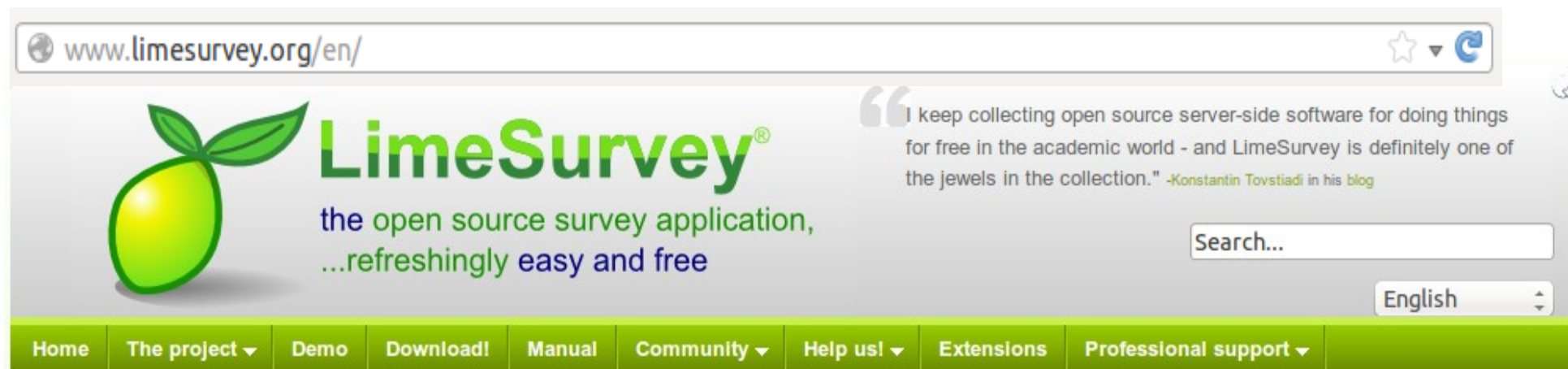
# NES and Study-Specific Data

- 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 a set of experiments
  - Data described in terms of a “standardized” structure defined by a database model
- **Problem:** Experiments may result in other kinds of data, with variable structure
  - Example: data collected by means of questionnaires in the Brachial Plexus Injury study at INDC-UFRJ
- Computational solution proposed to digitalize study-specific data:

## Electronic Questionnaires

# Integration of NES and LimeSurvey

- There are software systems which enable users to create electronic 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 the free, open-source software **LimeSurvey**  
<http://limesurvey.org/>
- NES has a special module that interfaces the access to questionnaires created with LimeSurvey → “centralized” management of experimental data



## Summary

- Document the experiments, assure reproducibility
- Store project data in an efficient and secure manner
- Automatize Lab's routine tasks
- Enable data reuse
- Facilitate the sharing of project resources
- Promote scientific collaboration
- Promote open science

## **Current Team**

- Carlos Ribas (NUMEC – USP)
- Evandro Santos Rocha (NUMEC - USP)
- Dr. Diogo de Carvalho Pedrosa (NUMEC – USP)
- Prof. Dr. Fabio Kon (IME – USP)
- Prof. Dr. Kelly R. Braghetto (IME – USP)

## **Main collaborators**

- Prof. Dr. Claudia D. Vargas (INDC – UFRJ) & her team
- Prof. Dr André F. Helene (IB – USP) & his team

# Data Management in NeuroMat and the Neuroscience Experiments System (NES)

Thank you for your attention.

Kelly Rosa Braghetto

kellyrb@ime.usp.br

This presentation was produced as part of the activities of FAPESP Research, Innovation and Dissemination Center for Neuromathematics (grant #2013/07699-0, S. Paulo Research Foundation)

