Aalto University
School of Science

Long-term preservation of brain imaging data

Enrico Glerean – web: www.glerean.com – twitter: @eglerean

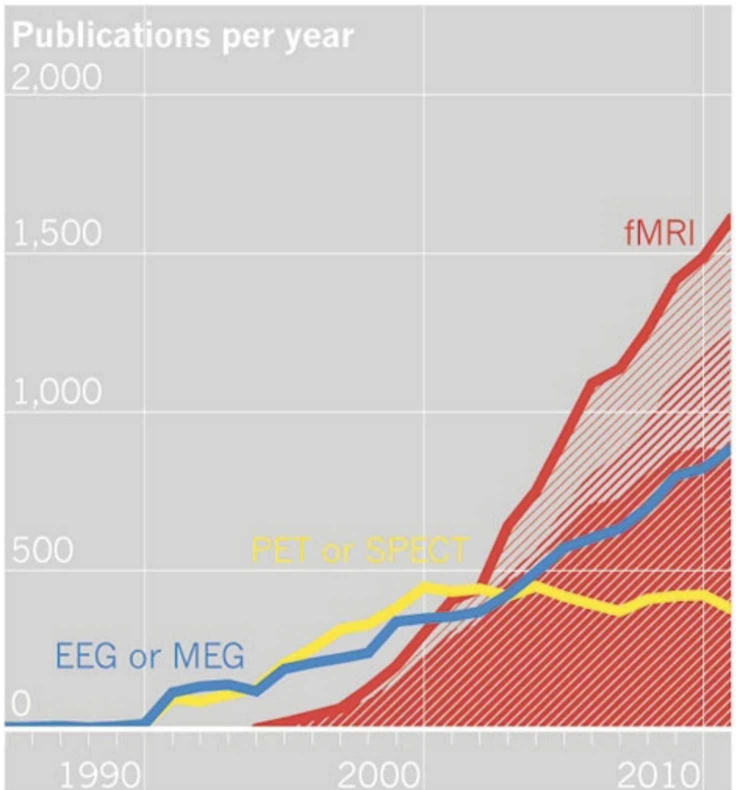
Context

The “big” data hype?

The “big” data hype… not just a hype

- **Data is a valuable asset** both in industry and – as you all know – in **science**
- News outlets topics: **BIG data on a daily basis**
- **BioBanks:** another trending topic often in recent news
  - **http://www.ukbiobank.ac.uk/**
  - **http://www.biopankki.fi/en/**
  - **http://www.healthcapitalhelsinki.fi/en/**

Evergrowing brain imaging data



http://www.nature.com/news/brain-imaging-fmri-2-0-1.10365

- **Larger amounts of data are seen also in neuroscience**
- **Increasing amount of subjects per study** (N > 16)
- **Large-scale projects** with healthy (**Human Connectome Project**, N ≈ 900 with fMRI/DTI/MEG) and clinical population (**ADNI** for Alzheimer's disease N ≈ 2500, **ABIDE** for Autism Spectrum N ≈ 500)
- **http://www.nature.com/sdata/**

SCIENTIFIC DATA

http://www.nature.com/sdata/

SCIENTIFIC DATA | DATA DESCRIPTOR OPEN

A multi-subject, multi-modal human neuroimaging dataset

Daniel G Wakeman & Richard N Henson

SCIENTIFIC DATA | DATA DESCRIPTOR OPEN

A high resolution 7-Tesla resting-state fM
retest dataset with cognitive and physiolo
measures

Krzysztof J Gorgolewski, Natacha Mendes, Domenica Wilfling, Elisabeth Wla
Claudine J Gauthier, Tyler Bonnen, Florence J.M Ruby, Robert Trampel, Pierr
Roberto Cozatl, Jonathan Smallwood & Daniel S Margulies

Affiliations | Contributions | Corresponding author

Scientific Data **2**, Article number: 140054 (2015) | doi:10.1038/sdata.2014.54
Received 29 August 2014 | Accepted 29 October 2014 | Published online 20 January 2015

SCIENTIFIC DATA | DATA DESCRIPTOR OPEN

A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie

Michael Hanke, Florian J. Baumgartner, Pierre Ibe, Falko R. Kaule, Stefan Pollmann, Oliver Speck, Wolf Zinke & Jörg Stadler

Affiliations | Contributions | Corresponding author

Scientific Data **1**, Article number: 140003 (2014) | doi:10.1038/sdata.2014.3
Received 04 November 2013 | Accepted 22 January 2014 | Published online 27 May 2014

Big data in science means sharing

- **Excellent examples from animal neuroscience**

  Allen Brain Atlas http://www.brain-map.org/ (mouse)
  CoCoMac http://cocomac.g-node.org/ (macaque)
  FlyCircuit http://www.flycircuit.tw/ (drosophila)
  NeuroData Without Borders http://www.nwb.org/ (rat)
  WormAtlas http://www.wormatlas.org/ (c-elegans)
  ZebraFish Brain Atlas http://zebrafishbrain.org/ (zebrafish)

- **Human neuroscience is somewhat lagging behind**, but great efforts are made with initiatives such as:

  OpenfMRI https://openfmri.org/
  International Neuroimaging Data-sharing Initiative http://fcon_1000.projects.nitrc.org/
  Open MEG Archive https://www.mcgill.ca/bic/resources/omega
  EEGBase https://eegdatabase.kiv.zcu.cz

The problem

With big data comes big responsibility

Issues with larger neuroimaging datasets

- For everyone: **Increasing amount of data → more efficient way of managing data** (i.e. automation)
- For everyone: **do not reinvent the wheel on how to organize your dataset**
- For PIs: **increasing amount of data → less control on data knowledge**
  You don't want to lose data, but you also don't want to lose the knowledge about your data (metadata)
- For university IT / management / grant agencies: **not possible to track data/metadata related to a project** (applies also to costs of data storage)

Data management

Aalto University
School of Science

Data management

Important not only **for good science** but also required in **grants** and **assessments of research quality** by institutions (e.g. recent audit of Aalto)

Three key issues in data management

1. **Storage of new data**
2. **Preservation of current and past data**
3. **Sharing of data** (even within the same lab!)

Data Matters: interview with Russell Poldrack

June 6, 2014 | 1:35 pm | Posted by Andrew Hufton | Category: Data Matters

Russell Poldrack is Professor of Psychology and Neurobiology and Director of the Imaging Research Center at the University of Texas in Austin.

What are the current data preservation practices within your field?

Data preservation practices are really non-existent. If people do anything it's usually saving something to DVDs or tapes, and then sticking it somewhere to rot. I've spoken to my colleagues, trying to find some of the early landmark datasets of fMRI papers to put into OpenfMRI (openfMRI.org), but most of them either say, we can't find the data anymore, or it's on a tape but we don't have the drive that can read it anymore. I have data from 10 years ago on various tape formats that I couldn't get to if I wanted to, though it seems that the technology has stabilised a bit. The other worry is that you put it on a DVD or a hard drive, but those things decay; people often have the assumption that once you put the data onto physical media it will be there as long as you want it, and that is definitely not the case. I think the best strategy is to replicate data geographically across as many different systems as possible so that there's no single point of failure.

Tutkimus-PAS pilot

Pitkäaikais-saatavuuspalvelu

Aalto University
School of Science

Tutkimus-PAS pilot

- Project under the **Avoin Tiede ja Tutkimus** initiative (http://avointiede.fi), financed by the **Ministry of Education and Culture** and coordinated by **CSC** (https://www.csc.fi/)
- First stage done with the **Kansallinen Digitaalinen Kirjasto** (http://www.kdk.fi/, long term storage of books, pictures, films)
- Pilot stage with **three types of scientific data**
  - Aalto University (brain imaging)
  - Jyväskylä University (nuclear physics)
  - Turku University (astronomy)
- Please come to **Open Science Expert Training** 8th/April http://avointiede.fi/osaajakoulutuksen-ohjelma

Tutkimus-PAS pilot outcomes

The pilot project consisted of

- **Questionnaires** regarding the **data and their metadata**
- **Preparation of a data package** for long term storage
- Preparation of XML metadata according to the **Metadata Encoding and Transmission Standard** (METS, http://www.loc.gov/standards/mets/)

Tutkimus-PAS pilot challenges

The most important issues I had to consider, since there are no standards in our field

- **How to store brain imaging data, if they were to be accessed 50 years from now?**
- **Is there a documented standard to store human brain imaging data?**
- **Which data and metadata are relevant to fully replicate existing results?**
- **What are the permissions associated to the data?**

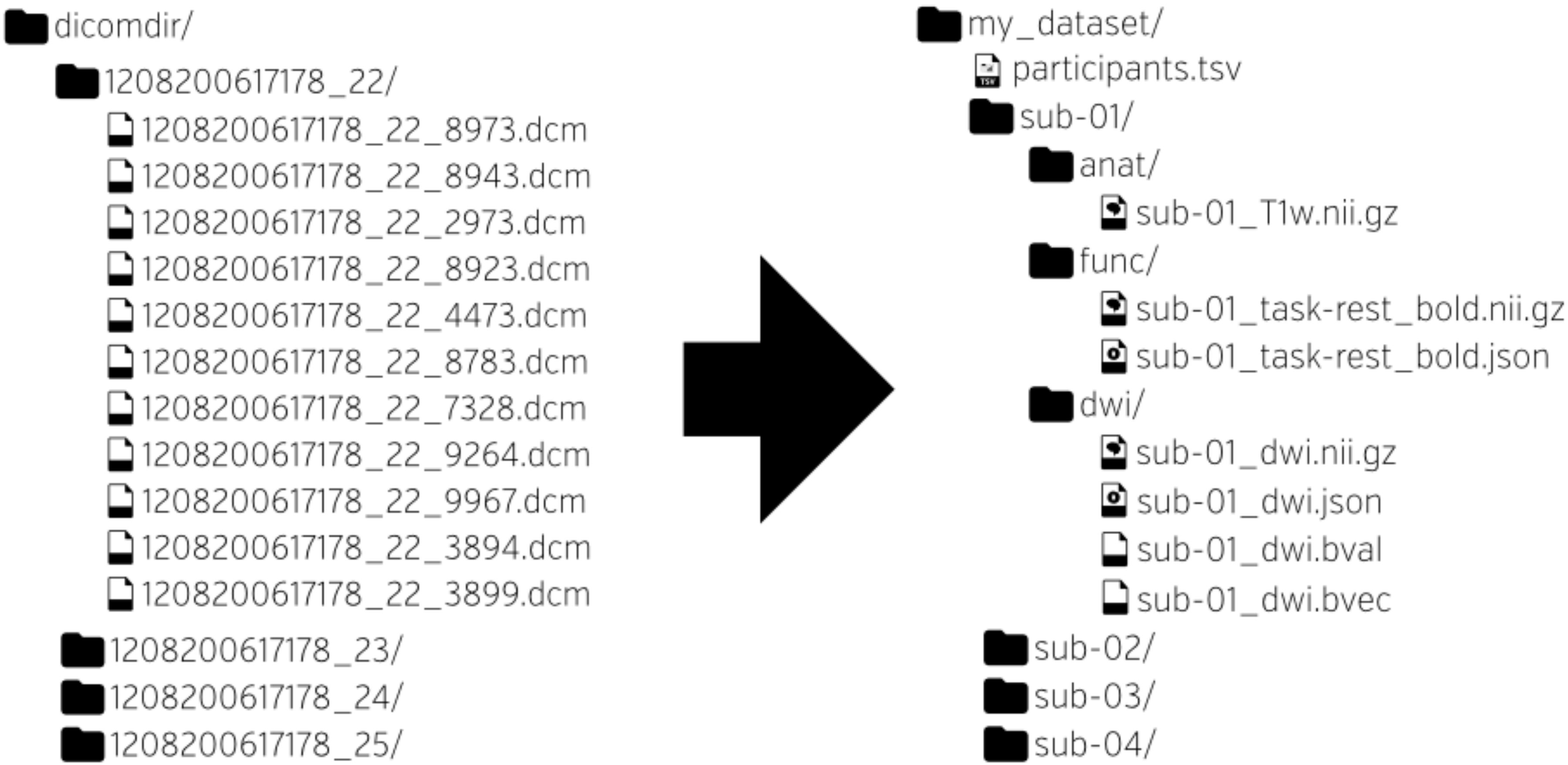
Tutkimus-PAS pilot proposed solutions

- **How to store brain imaging data, if they were to be accessed 50 years from now?**
  – ***Open file forma**t as close as possible to the data*
- **Is there a documented standard to store human brain imaging data?**
  – *So far the only candidate is **BIDS** (fMRI… so far)*
- **Which data and metadata are relevant to fully replicate existing results?**
  – *Storing the data as well as the experimental protocol*
- **What are the permissions associated to the data?**
  – *This is still an open issue*

BIDS – Brain Imaging Data Structure

- **Developed at Poldrack Lab @ Stanford** **https://poldracklab.stanford.edu/**
  (same people behind http://openfmri.org/ http://www.neurovault.org/)
- **Website http://bids.neuroimaging.io/**
- Supported by **http://incf.org/** (behind other data sharing initiative such as Neurodata Without Borders)
- Version 1.0 from **October 2015**, received feedback from neuroscience community (you can still give comments!)
- **Focused on human MRI experiments, currently extending to PET**, but it’s clear that the same data structure would work for MEG/EEG/etc.

BIDS in one picture



Example of metadata file

```
{
    "RepetitionTime": 3.0,
    "EchoTime": 0.03,
    "FlipAngle": 78,
    "SliceTiming": [0.0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2,
1.4, 1.6, 1.8, 2.0, 2.2, 2.4, 2.6, 2.8],
    "MultibandAccellerationFactor": 4,
    "ParallelReductionFactorInPlane": 2,
    "InPlanePhaseEncodingDirection": "AP"
}
```

Why BIDS?

- **Simple**: not relying on external databases or complicated formats.
- Some **metadata encoded in the folder structure and in the filename**
- Use of **tab-separated-value files**
- **NIFTI** format for imaging data
- **JSON** for metadata
- **Allows customization**

See also: **http://slideshare.net/chrisfilo1/brain-imaging-data-structure**

Advantages

- for PIs: **easy sharing** of one dataset from one student/ postdoc to another
- for those who collect the data: **less effort in documenting how the data is stored**
- for workflow developers: writing **pipelines expecting a specific file** organization. **Validating** the completeness of a dataset **automatically.**
- for database curators: **no need to re-define ad hoc structure** for input datasets.

BIDS implementation for the PAS pilot

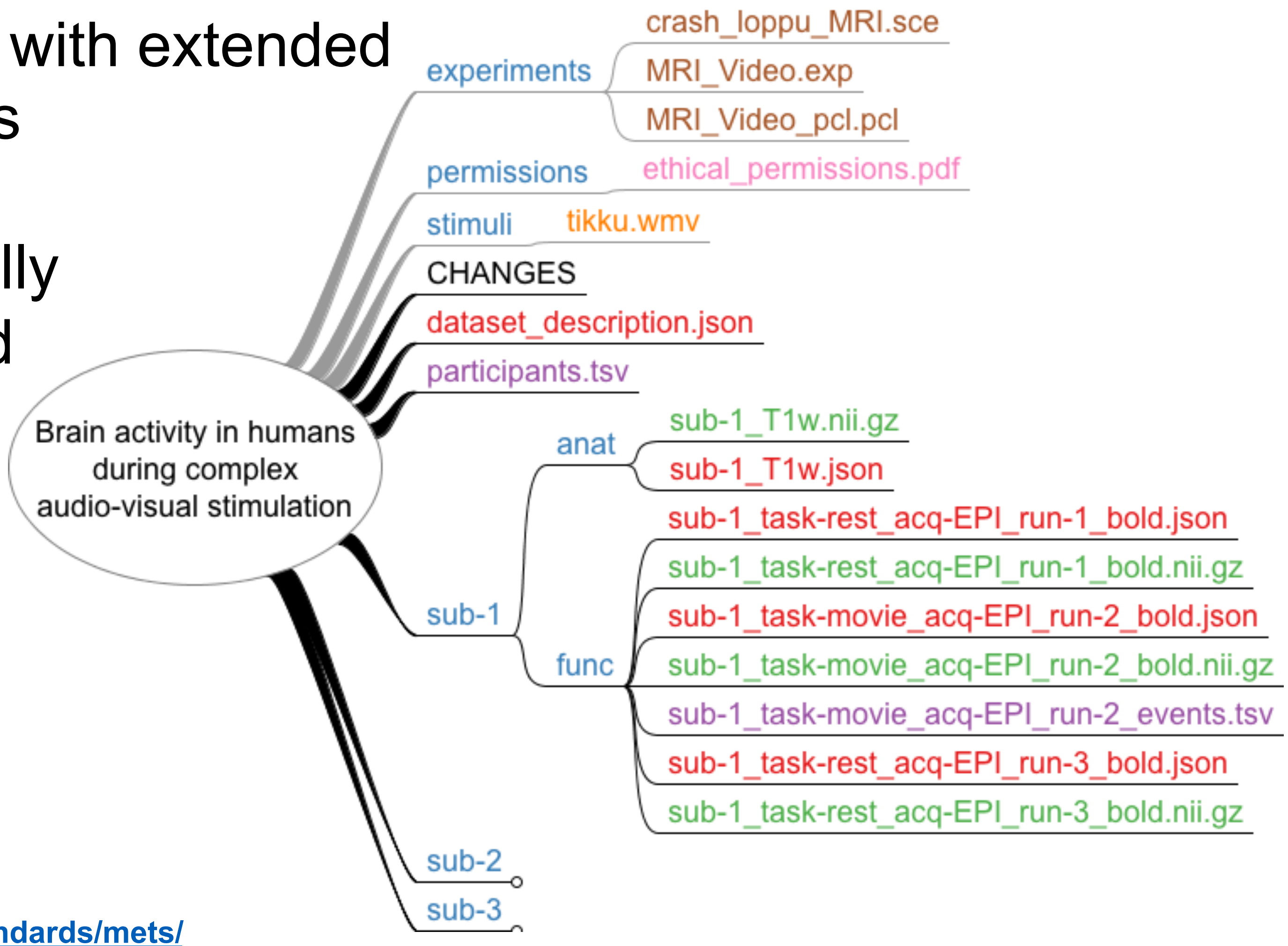
- **Dicom to NIFTI** and storing relevant DICOM header information (see github dcm2niix)
- **Anonymizing** (de-facing) MRI anatomicals (using mri_deface)
- **Reorganizing files following folder structure**

Full pipeline code for our lab at https://github.com/eglerean/bramila_dcm2bids (I am happy to help if you want to get started). See also BIDS website for more tools: http://bids.neuroimaging.io/

Submitted package for the PAS pilot

– Valid **BIDS** with extended subfolders

– Automatically generated **XML** file following METS specs*



* http://www.loc.gov/standards/mets/

Challenging questions from PAS pilot

- **How the dataset is permitted to use?**
  The dataset cannot be shared publicly. Collaborators can obtain access to the data by contacting the authors or owner of the dataset.
- **How to link to the dataset?**
  The dataset doesn't currently have a digital object identifier.
- **From where the dataset can be found? [e.g. catalog, link]**
  This dataset is available only on request

Broader implications of research data management

Aalto University
School of Science

Implications of data management

1. **How to deal with data sharing and data policies in journals?**
   Using data-sharing initiatives/databases? Hosting own data and a committee to approve data access?
2. **How to deal with long-term preservation of data?**
   Data expiry date? How long will an open database exist?
3. **Ethical implications of shared human data**
   Make data anonymous from day zero?
4. **How to establish a dialogue between scientists, IT and management personnel?**
   Monitoring tools to track data? (e.g. ownership, permissions, storage, expiry date, etc)
5. **Lack of training**
   How to motivate PIs to better manage data?

Take home messages

Three take home messages

1. **PIs: enforce best practice of data management**
   It will save your time, your students time and will make it easier to go back to old dataset, share new one and in general be in control of large projects. **BIDS** seems like a promising solution (I am happy to help).
2. **Standardized data and metadata formats** will make it also easier for **IT/management/grant agencies** to monitor projects and costs.
3. **Unsolved issues: what/where to store for long term preservation? Data expiry date? Open sharing of data? Ethical implications? Training to become efficient data managers?**

Aalto University
School of Science

Long-term preservation of brain imaging data

Enrico Glerean – web: www.glerean.com – twitter: @eglerean