Neuroinformatic techniques for provenance & data sharing

Camille Maumet

GlaxoSmithKline - Neurophysics Workshop on Skeptical Neuroimaging

January 14th, 2014
Outline

1. Data sharing: current practice in neuroimaging
2. How to become less skeptical?
3. Neuroinformatics techniques for provenance and data sharing
Outline

1. Data sharing: current practice in neuroimaging
2. How to become less skeptical?
3. Neuroinformatics techniques for provenance and data sharing
Overview of a neuroimaging study

- MR scanner: 1.5T, 3T…
- Type of coils: 32-channels, 16-channels…
- Imaging sequence: TR, TE, FOV…
- fMRI paradigm

- Pre-processing pipeline
- Method employed for each processing
- Parameters for each method, software…

- Model
- (Non-)parametric
- Parameters

- Publication
Neuroimaging and data sharing

Acquisition

Pre-processing

Statistical analysis

Detection

Estimates / Statistic

Publication

Data shared with my collaborators
Sharing data with my collaborators

Acquisition → Pre-processing → Statistical analysis → Publication

Raw data → Pre-processed data → Detection Estimates / Statistic

ZIP

SPM

FSI

XFNI

LORIS

XNAT

Adobe PDF XLS
Neuroimaging and data sharing

Acquisition

Pre-processing

Statistical analysis

Publication

Data shared with my collaborators

Data shared with the whole community
A neuroimaging publication

- **Methods** section: metadata in free-form text.

  General technical implementation
  A single scanner session included the four paradigms separately implemented with the same parameters: a simple block design alternated a rest condition as control and the language task, starting with rest, with a preliminary period of signal acquisition for MRI signal stabilization which was later discarded during data processing. Each paradigm included three 27-s blocks of each condition and had a total duration of 2 min 48 s. The scanner session, including the anatomical acquisition and the four language paradigms, had a duration of about 30-35 min. All subjects performed the tasks in the same order, as during the preparation step, in order to avoid the mix of auditory and visual tasks and the resulting complication for the child. Words required by the tasks were one-to-three syllable words highly frequent in the lexicon of French 8 years old children (Lambert and Chevret, 2001).

  During the rest condition, a red cross was displayed on the instructions screen and children were asked “not to move”, to “think

  Table 2

<table>
<thead>
<tr>
<th>Task comparison (-) and conjunctions (CL)</th>
<th>Auditory language tasks</th>
<th>Visual language tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category</td>
<td>Categ. Def.</td>
<td>Categ. CL.</td>
</tr>
<tr>
<td>Left hemisphere</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Inf. frontal-Operative</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Pre-surgical</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Motor</td>
<td>35-3.66</td>
<td>-</td>
</tr>
<tr>
<td>SMA</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cingulate</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Med. sup. frontal</td>
<td>174-4.60</td>
<td>-</td>
</tr>
<tr>
<td>Bil. spec.</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Insula</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sup. temporal</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sup. parietal</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sup. occipital</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Mid. occipital</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Pseudonome</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

  2D plot(s) of the detections

  Description of the detections

  Table of local maxima

  Table
Outline

1. Data sharing: current practice in neuroimaging
2. How to become less skeptical?
3. Neuroinformatics techniques for provenance and data sharing
Reproducibility

Acquisition

Pre-processing

Statistical analysis

Publication

Acquisition

Pre-processing

Statistical analysis

Publication

New conclusions
Full provenance

Acquisition

Pre-processing

Statistical analysis

Publication

Pre-processing

Statistical analysis

Statistical analysis

Pre-processed data

Raw data

Detection

Estimates / Statistic

Detection

Estimates / Statistic

Detection

Estimates / Statistic
Meta-analysis: analyzing the analyses

- Coordinate-Based Meta-Analysis (CBMA)
- Image-Based Meta-Analysis (IBMA)
How to become less skeptical?

• Reproducibility
  – Confirm results by re-running an analysis

• Provenance
  – Needed for reproducibility
  – Avoid selection bias.

• Meta-analysis
  – Strengthen results by combining studies.

• What do we need?
  – Sharing data, meta-data and provenance.
Data sharing: obstacles

• Psychological
  – “My” data

• Ethical constraints

• Technical: difficulties to share data with enough metadata to be really useful
  – Available data versus usable data.

“Less than a few percents of acquired neuroimaging data is available in public repositories” [Poline 2012]
Outline

1. Data sharing: current practice in neuroimaging
2. How to become less skeptical?
3. Neuroinformatics techniques for provenance and data sharing
Data sharing tools

Acquisition → Pre-processing

Raw data → Pre-processed data

Statistical analysis → Detection

Estimates / Statistic

Publication

OpenfMRI
ADNI
BIRN
HID
LORIS
XNAT

NeuroVault
A public repository of brain activation maps

brainmap.org
SumsDB
neurosynth.org

SPM
Nipype: Neuroimaging in Python Pipelines and Interfaces

LONI PIPELINE
FSL
A standard format for meta-data

• Sharing data across the data sharing tools…
• First attempt of an agnostic format: XML-Based Clinical Experiment Data Exchange Schema (XCEDE): www.xcede.org
  – Describes subject, study, activation
  – Limited provenance encoding
  – Initiative of the BIRN
• NeuroImaging Data Model NI-DM: www.nidm.nidash.org
  – Based on web-semantic tools.
  – Initiative of the BIRN and INCF
Three major players

- Bottom-up approach.
- Lean on existing analysis software (SPM, FSL, AFNI) to disseminate the standard.
Work in progress

- Define a format to represent the results of a neuroimaging study with a focus on meta-analysis.

<table>
<thead>
<tr>
<th>Term name</th>
<th>Definition</th>
<th>Example</th>
<th>BIRNLex or NIDM Concept ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>BonferroniCorrection</td>
<td>Bonferroni correction for multiple statistical tests</td>
<td></td>
<td>nmdm/nidm_80</td>
</tr>
<tr>
<td>chi-squaredStatistic</td>
<td>A statistical parameter drawn from a chi-square statistic</td>
<td></td>
<td>nmdm/nidm_81</td>
</tr>
<tr>
<td>FDR</td>
<td>False Discovery Rate correction</td>
<td></td>
<td>nmdm/nidm_82</td>
</tr>
<tr>
<td>FWER</td>
<td>Family-wise Error Rate correction</td>
<td></td>
<td>nmdm/nidm_83</td>
</tr>
<tr>
<td>Scan Image</td>
<td>An image that is the output of an MRI or CT or PET scan</td>
<td></td>
<td>nmdm/nidm_84</td>
</tr>
<tr>
<td>SliceOrder</td>
<td>The temporal order in which the 2D slices were acquired by the imaging systems</td>
<td></td>
<td>nmdm/nidm_85</td>
</tr>
<tr>
<td>Voxel</td>
<td>Volumetric pixel</td>
<td></td>
<td>nmdm/nidm_86</td>
</tr>
<tr>
<td>Z-Statistic</td>
<td>A statistical parameter drawn from a normal or z distribution</td>
<td></td>
<td>nmdm/nidm_87</td>
</tr>
<tr>
<td>extentThreshold</td>
<td>Minimum cluster size used when thresholding a statistic image</td>
<td>5 voxels</td>
<td>nmdm/nidm_88</td>
</tr>
<tr>
<td>errorDegreesOfFreedom</td>
<td>Degrees of freedom of the error.</td>
<td>73</td>
<td>nmdm/nidm_89</td>
</tr>
<tr>
<td>effectDegreesOfFreedom</td>
<td>Degrees of freedom of the effect.</td>
<td>1</td>
<td>nmdm/nidm_90</td>
</tr>
<tr>
<td>StatisticMap</td>
<td>A map (2D or 3D structured dataset) whose value at each location is a statistic.</td>
<td></td>
<td>nmdm/nidm_91</td>
</tr>
<tr>
<td>voxelSize</td>
<td>3D size of a voxel measured in voxellunits.</td>
<td>(2 2 4)</td>
<td>nmdm/nidm_92</td>
</tr>
<tr>
<td>cluster</td>
<td>A group of neighboring image elements (voxels or vertices)</td>
<td></td>
<td>nmdm/nidm_93</td>
</tr>
<tr>
<td>p-valueFDR</td>
<td>p-value adjusted for the search volume, controlling for the False Discovery Rate</td>
<td>0.000154</td>
<td>nmdm/nidm_94</td>
</tr>
<tr>
<td>p-valueFWE</td>
<td>p-value adjusted for the search volume, controlling for the Familywise Error Rate</td>
<td>0.00056</td>
<td>nmdm/nidm_95</td>
</tr>
<tr>
<td>p-valueUncorrected</td>
<td>Uncorrected p-value</td>
<td>0.0042</td>
<td>nmdm/nidm_96</td>
</tr>
</tbody>
</table>
Neuroimaging terms

- Define a vocabulary to support the format.

<table>
<thead>
<tr>
<th>Term name</th>
<th>Definition</th>
<th>Example</th>
<th>BIRNLex or NIDM Concept ID</th>
<th>synonyms and related urls</th>
<th>Parent term</th>
</tr>
</thead>
<tbody>
<tr>
<td>cluster</td>
<td>A group of neighboring image elements (voxels or vertices)</td>
<td></td>
<td>nidm:nidm_93</td>
<td><a href="http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C43">http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C43</a></td>
<td></td>
</tr>
<tr>
<td>qValueFDR</td>
<td>p-value adjusted for the search volume, controlling for the False Discovery Rate</td>
<td>0.000154</td>
<td>nidm:nidm_94</td>
<td><a href="http://purl.obolibrary.org/obo/OBI_0000251">http://purl.obolibrary.org/obo/OBI_0000251</a></td>
<td></td>
</tr>
<tr>
<td>pValueUncorrected</td>
<td>Uncorrected p-value</td>
<td>0.0542</td>
<td>nidm:nidm_96</td>
<td><a href="http://purl.obolibrary.org/p-value">http://purl.obolibrary.org/p-value</a> i.e. nidm:nidm_0011</td>
<td></td>
</tr>
<tr>
<td>SelectionProcedure</td>
<td>Procedure to select the values that are being reported</td>
<td></td>
<td>nidm:nidm_97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>clusterSizeInVoxels</td>
<td>Number of voxels contained in a cluster.</td>
<td>18</td>
<td>nidm:nidm_98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>softwareVersion</td>
<td>Name and Number specifying software version.</td>
<td>SPM99, SPM2, SPM5, SPM8, SPM12b, FSL5.0.0</td>
<td>nidm:nidm_99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>clusterSizeInVertices</td>
<td>Software revision number</td>
<td>v5417</td>
<td>nidm:nidm_100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>clusterSizeInResels</td>
<td>Number of resels contained in a cluster.</td>
<td>10</td>
<td>nidm:nidm_101</td>
<td></td>
<td></td>
</tr>
<tr>
<td>voxelUnits</td>
<td>Units associated to each dimensions of some N-dimensional data.</td>
<td>[mm 'mm' 's']</td>
<td>nidm:nidm_102</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ReselSizeInWorldUnits</td>
<td>Volume of a resel, a resolution element, expressed in units. It expresses the smoothness of the noise, with smoother images having larger resels.</td>
<td></td>
<td>nidm:nidm_103</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Map</td>
<td>2D or 3D structured dataset.</td>
<td></td>
<td>nidm:nidm_104</td>
<td><a href="http://en.wikipedia.org/wiki/Resel">http://en.wikipedia.org/wiki/Resel</a></td>
<td></td>
</tr>
<tr>
<td>fileName</td>
<td>Name associated with a file (without path).</td>
<td></td>
<td>nidm:nidm_107</td>
<td></td>
<td></td>
</tr>
<tr>
<td>numberOfDimensions</td>
<td>Number of Dimensions of some N-dimensional data.</td>
<td>3</td>
<td>nidm:nidm_108</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dimensions</td>
<td>Dimensions of some N-dimensional data.</td>
<td>[64, 64, 20]</td>
<td>nidm:nidm_109</td>
<td></td>
<td></td>
</tr>
<tr>
<td>coordinateSystem</td>
<td>Type of coordinate system.</td>
<td></td>
<td>nidm:nidm_110</td>
<td></td>
<td></td>
</tr>
<tr>
<td>searchVolumenInVoxels</td>
<td>Total number of voxels within the search volume.</td>
<td>66866</td>
<td>nidm:nidm_111</td>
<td>Synonyms of nidm:volumenInVoxels</td>
<td></td>
</tr>
<tr>
<td>searchVolumenInResels</td>
<td>Total number of resels within the search volume.</td>
<td>151.3</td>
<td>nidm:nidm_112</td>
<td>Synonyms of nidm:volumenInResels</td>
<td></td>
</tr>
</tbody>
</table>
Data model

- Based on PROV-DM a W3C recommendation to encode provenance. [www.w3.org/TR/prov-dm/]
Data model
Data model: activities
Data model: agent
Data model: entities
Data model: entities
Conclusion

- Data sharing is one key to reduce skepticism.
- There is already a number of technical solutions for data sharing in neuroimaging.
- A meta-data standard would benefit all of these efforts
  - NI-DM: [http://nidm.nidash.org](http://nidm.nidash.org)
Q & A

This work is supported by the Wellcome Trust

Warwick