

**5 Suggested "Best Practices" in
Informatics for Data and Resource Discovery in Addiction Research**

Gwen Frishkoff & Team NEMO (last updated 2010 July 5)

1. **BEST PRACTICE #1 (DATA PROVENANCE)**. Use a standardized checklist for *internal documentation* and *reporting of provenance* for each experiment dataset.
 - a. TIME FRAME: During study planning (design phase), data acquisition & analysis, archiving, and publication. *Checklist should take max. 45 mins to complete for a given study.*
 - b. BENEFITS FOR YOU AS AN INDIVIDUAL RESEARCHER:
 - i. In completing the checklist, you will be writing a good part of your Methods section
 - ii. If you store these meta-data with your raw and transformed data at each stage of the experiment, it will be easier to repeat and replicate the study at a later time.
 - c. BENEFITS FOR RESEARCH COMMUNITY:
 - i. Standardized reporting of provenance will facilitate *data sharing* and *meta-analyses*
 - d. IMPLEMENTATION:
 - i. Steps you can take:
 - Adopt existing *checklist* or design a new one within *MIBBI* framework.
 - In completing the checklist for each experiment, select values from *enumerated list* wherever possible (ideally, the list should be use *ontologies* that are linked to MIBBI checklists... OBI/IAO, MC, cogPO).
 - ii. Steps the community can take:
 - Systematic polling to determine MI for research area (cf. published guidelines)
 - Publication of MI checklists via MIBBI portal (http://mibbi.org/index.php/MIBBI_portal)
 - Incorporation of MI standards in journal "Guidelines for Authors" (e.g., <http://www.biomedcentral.com/bmcbioinformatics/ifora/>)

NOTES:

- Everyone should see the utility of this practice
- To see that the practice takes hold in community, need three things: (1) wide consensus or promotion by trusted entity (e.g., NIF, NIH,...); (2) dissemination of checklists that are (3) relatively transparent and fast to complete (<45 mins per study). Cf. BrainMap case study in developing standards for reporting of fMRI data (Laird, et al., 2005).

REFERENCES

- Gibson, F., Overton, P. G., Smulders, T. V., Schultz, S. R., Eglén, S. J., Ingram, C. D., et al. (2008). Minimum Information about a Neuroscience Investigation (MINI): Electrophysiology. *Nature Precedings*.
- Gordon, E., Cooper, N., Rennie, C., Hermens, D., & Williams, L. M. (2005). Integrative neuroscience: The role of a standardized database. *Clin EEG Neurosci*, 36(2), 64-75.
- Laird, A. R., Lancaster, J. L., & Fox, P. T. (2005). Brainmap: The social evolution of a human brain mapping database. *Neuroinformatics*, 3(1), 65-78.

2. **BEST PRACTICE #2 (DATA ANALYSIS— WITHIN DATASETS)**. Apply *data-driven techniques* for analysis of complex, high-dimensional datasets
- a. TIME FRAME: During analysis phase. *May take 1 hr to 1 mo to complete, depending on data dimensionality & preprocessing.*
 - b. BENEFITS FOR YOU AS AN INDIVIDUAL RESEARCHER:
 - i. allow for data discovery
 - ii. greater confidence you haven't missed something in the data
 - iii. compatible with hypothesis-driven approach (see #4 below)
 - c. BENEFITS FOR RESEARCH COMMUNITY:
 - i. Discovery of new patterns
 - ii. Support for data sharing & integration — to the extent that results from different data-driven analysis can be combined (see #5 below).
 - d. IMPLEMENTATION:
 - i. Steps you can take:
 - Use available **MATLAB toolboxes** for data preprocessing & analysis, (EEGLAB, ERP PCA Toolkit, NEMO Toolkit...), data mining (clustering, classification)
 - Record analysis parameters in MIBBI-style checklist (note that NEMO will be representing parameters in EEG/MEG analysis in our MIBBI checklist, and in NEMO_data ontology)
 - ii. Steps the community can take:
 - Work towards integration, interoperability of different data analysis toolkits
 - Research and disseminate recommendations (which analysis methods for which kinds of data — e.g., "best practices" for PCA/ICA for ERPs)
 - Work on representing key concepts (instruments, algorithms, settings, & parameters) in OBI-compatible ontology. See NEMO_data

NOTES:

- I expect this suggested "best practice" to generate more discussion, possibly some controversy. In EEG/ERP community, for example, analysis methods are often divided into hypothesis-driven versus exploratory (or data-driven) methods. This may be a false trade-off. We need **both** to foster good science and data sharing and integration (see Notes under BP#4!)
- Implementation ("best practice" for PCA/ICA) is also likely to generate some controversy, which is fine. Multiple methods should be promoted in research (let a thousand methods bloom...). However, reporting of methods should be standardized, as part of data provenance (BP #1).

REFERENCES

Frishkoff, G.A., Frank, R., Rong, J., Dou, D., Dien, J., & Halderman, L. (2007). A framework to support automated ERP pattern classification and labeling. *Computational Intelligence and Neuroscience*, vol. 2007, Article ID 14567, 13 pages.

3. **BEST PRACTICE #3 (DATA METRICS)**. Describe patterns in complex, high-dimensional data using multiple metrics that fully characterize key dimensions of the data (e.g., temporal/spectral, spatial/topographic, individual conditions as well as contrasts where possible). Report full set of *data metrics*, along with *data provenance* (e.g., as Supplement). Note this "best practice" complements BP#1-2.
- a. TIME FRAME: During analysis and reporting phase. *Measure generation is very fast with appropriate tools. See (for example) NEMO_ERP_Metric_Extraction & RDF generation tools.*
 - b. BENEFITS FOR YOU AS AN INDIVIDUAL RESEARCHER:
 - i. allow for multiple exploratory analyses (which may be based on single workflow, rather than multiple repeated analyses)
 - ii. compatible with hypothesis-driven approach (see #4 below)
 - c. BENEFITS FOR RESEARCH COMMUNITY:
 - i. Discovery of new patterns
 - ii. Support for data sharing & integration — to the extent that results from different data-driven analysis can be combined (see #5 below).
 - d. IMPLEMENTATION:
 - i. Steps you can take:
 - Use available tools for statistical extraction
 - Record data metrics in MIBBI-style checklist (note that NEMO represents spatial, temporal/spectral, and experiment metrics in our checklist and ontology)
 - ii. Steps the community can take:
 - Work towards integration, interoperability of different measure generation tools
 - Work on developing checklists to represent standard (in OBI-compatible ontology). See NEMO_data.

NOTES:

- This recommendation may also stimulate some discussion. Like BP#2, this practice is **not** intended as an alternative to hypothesis-driven research. Scientific report can (and usually should) focus on one or two a priori hypotheses.
- However, **full set of results can be generated, archived**, and made available with little or no additional effort. For scalp-recorded EEG and MEG data, I would argue that data sharing and integration cannot be accomplished without BP #2-3.

REFERENCES

- Jakobovits, R., Soderland, S. G., Taira, R. K., & Brinkley, J. F. (2000). Requirements of a web-based experiment management system. *Proc AMIA Symp*, 374-378.
- Small, S. L., Wilde, M., Kenny, S., Andric, M., & Hasson, U. (2009). Database-managed grid-enabled analysis of neuroimaging data: the CNARI framework. *Int J Psychophysiol*, 73(1), 62-72.

4. **BEST PRACTICE #4 (DATA CLASSIFICATION & LABELING).** Classify and label patterns using **pattern definitions** from the literature ("knowledge-driven" approach)
- a. **TIME FRAME: *Difficult to quantify.*** Classification and labeling is at the heart of science research. Known definitions, classification methods can be applied to a single dataset to generate results for publication. Ideally, though, the raw (unlabeled) data metrics & provenance will be archived and shared to permit large-scale meta-analyses, the results of which may suggest a new taxonomy of patterns (and hence, re-classification of legacy data).
 - b. **BENEFITS FOR YOU AS AN INDIVIDUAL RESEARCHER:**
 - i. Most cognitive-behavioral paradigms are variations on prior studies. It is therefore important to link your findings to prior literature. To do so, it is often necessary to apply pattern classification (definitions) that have been used in previous reports
 - ii. Linking your data to data from the same and similar paradigms is (typically) necessary for publication.
 - iii. Linking your data to data from disparate paradigms (e.g., making links between research on visual attention, research on reading acquisition) leads to deeper understanding, synthesis across research areas. Researchers live for this kind of synthesis!
 - c. **BENEFITS FOR RESEARCH COMMUNITY:**
 - i. Support for data sharing & integration — to the extent that results from different data-driven analysis can be combined (see #5 below).
 - d. **IMPLEMENTATION:**
 - i. **Steps you can take:**
 - Read the literature (but with a critical eye -- given that authors may not provide sufficient details about the data (i.e., full set of metrics) or data provenance to draw parallels across datasets)
 - Make sure that your pattern definitions are complete (e.g., that they make use of spatial as well as temporal/spectral criteria)
 - ii. **Steps the community can take:**
 - Develop and vet domain ontologies
 - Use ontologies for data classification
 - Revise ontologies to reflect new science

NOTES:

- This best practice will **require ontologies** (not just C.V.?) and in an interesting way.
- NEMO approach is to catalog all ERP pattern definitions and encode these in formal semantics (using DL, not just text definitions). Then use ontologies to classify data. Data that fall into undefined_ERP_pattern class (which is define as a complement to defined ERP pattern classes) provide stimulus for revising ontologies. Hence, a **happy marriage between bottom-up (data-driven) and top-down (hypotheses-driven) methods!**

REFERENCES

Frishkoff, G.F., LePendu, P., Frank, R.M., Liu, H., and Dou, D. (2009). Development of Neural Electromagnetic Ontologies (NEMO): Ontology-based Tools for Representation and Integration of Event-related Brain Potentials. *Nature Precedings*, <http://dx.doi.org/10.1038/npre.2009.3458.1>.

5. **BEST PRACTICE #5 (DATA ANALYSIS — ACROSS DATASETS)**. Classify and label patterns using **data mining** from the literature ("knowledge-driven" approach)
- a. TIME FRAME: Could take 1 mos or 1 year, depending on domain and available datasets and tools
 - b. BENEFITS FOR YOU AS AN INDIVIDUAL RESEARCHER:
 - i. Test robustness of effects and how they may generalize (or not) across research paradigms.
 - ii. Stimulate ideas for future research studies based on meta-analysis results
 - c. BENEFITS FOR RESEARCH COMMUNITY:
 - i. Establish robustness of previously reported findings
 - ii. Discover new findings
 - d. IMPLEMENTATION:
 - i. Steps you can take:
 - Learn about data mining tools and techniques (check out WEKA for free clustering & classification tools)
 - Contribute data to projects such as NIF, BrainMap/cogPO, NEMO, and HeadIT
 - ii. Steps the community can take:
 - Support projects such as NIF, BrainMap/cogPO, NEMO, and HeadIT!

REFERENCES

- Dou, D., Frishkoff, G., Rong, J., Frank, R., Malony, A., and Tucker, D. (2007). Development of NeuroElectroMagnetic Ontologies (NEMO): A framework for mining brain wave ontologies, *Proceedings of the Thirteenth International Conference on Knowledge Discovery and Data Mining (KDD2007)*, pp. 270-279, San Jose, CA.
- Gardner, D., Knuth, K. H., Abato, M., Erde, S. M., White, T., DeBellis, R., et al. (2001). Common data model for neuroscience data and data model exchange. *J Am Med Inform Assoc*, 8(1), 17-33.
- Hasson, U., Skipper, J. I., Wilde, M. J., Nusbaum, H. C., & Small, S. L. (2008). Improving the analysis, storage and sharing of neuroimaging data using relational databases and distributed computing. *Neuroimage*, 39(2), 693-706.
- LePendu, P., Dou, D., Frishkoff, G. , and Rong, J. (2008). Semantic data modeling: Methods for ontology-based queries and an application to brainwave data., *Proceedings of the 20th International Conference on Scientific and Statistical Database Management (SSDBM-08)*, Hong Kong, China.
- Small, S. L., Wilde, M., Kenny, S., Andric, M., & Hasson, U. (2009). Database-managed grid-enabled analysis of neuroimaging data: the CNARI framework. *Int J Psychophysiol*, 73(1), 62-72.
- Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*, 25 (11), 1251-55.