Lost in localization: A solution with neuroinformatics 2.0?*

Finn Årup Nielsen
Center for Integrated Molecular Brain Imaging, Copenhagen, Denmark;
DTU Informatics, Technical University of Denmark, Lyngby, Denmark;
Neurobiology Research Unit,
Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark

August 11, 2009

Abstract

The commentary by Derrfuss and Mar (2009) discusses some of the limitations of the present databases and calls for a universal coordinate database. Here I discuss further issues and propose another angle to the solution of a universal coordinate database with the use of wiki technology.

The commentary by Derrfuss and Mar (2009) in NeuroImage shows how far the present coordinate databases lag behind the published literature and the two authors call for a universal coordinate database. For the studies of methodologies for meta-analysis we have undertaken a universal coordinate database has not been necessary. On the other hand, for neuroscience research a universal coordinate database will greatly help to identify relevant studies, in the execution of meta-analyses and to obtain more unbiased interpretations of the prior literature. Several databases exist for coordinates and each has different advantages from 'minimal' approaches to efforts with a high degree of annotation (Hamilton, 2009; Van Essen, 2009; Laird and Fox, 2009). Since 1998 we have gained some experience in the area by working both with the BrainMap and the Brede coordinate databases, and apart from the issues mentioned by Derrfuss and Mar I would like to raise three additional concerns: Ownership, extensibility and community involvement.

One key obstacle identified by Derrfuss and Mar is how new studies can be submitted and included in the database. For some time limitations of the Brede Database have been apparent to us, and recently we have explored wiki-solutions to counter some of them setting up the MediaWiki-based *Brede Wiki*: http://neuro.imm.dtu.dk/wiki/. (MediaWiki is the software that also runs Wikipedia). A wiki presents an open interface for anyone to edit and read where the entered data becomes immediately available. It is even possible to built scripts that automatically add information to a wiki. One interesting project in this domain is *Gene Wiki* where researchers pop-

*Published in *NeuroImage*, **48**:11–13, 2009. DOI: 10.1016/j.neuroimage.2009.05.073

ulate Wikipedia with genetics information (Huss, III et al., 2008). For the Brede Wiki a simple Matlab script allows for formatting coordinate lists from the SPM program in the style suitable for inclusion in the Brede Wiki. Another issue pointed to by the two authors arises when the large amount of information needed to be entered form a contribution barrier. The user needs to learn all aspects of the data entry and go through a time-consuming submission process before the user can contribute. As an important element in recruiting new users a wiki system typically has a low contribution barrier (Bryant et al., 2005)—e.g., it only takes two mouse clicks and a keypress to correct a comma error. In a wiki the data entry task can be broken down as information can be entered incrementally by different editors, e.g., one may start with core information like bibliographic information and coordinates and then later on add, e.g., subject information and imaging modality.

The MediaWiki software has built-in text search facilities and means for categorizing pages. However, it lacks more complex means of query. Recent research effort has gone into semantic or fielded wikis that represent information with types, and the so-called *templates* of MediaWiki enable this. These systems may build taxonomies and ontologies such that computer programs can recognize, e.g., that 'happiness' is an 'emotion' or that a specific number is an x-coordinate rather than a Brodmann area or a page number. Examples of semantic wikis are SNPedia (http://snpedia.com) databasing genetic variations and NeuroLex (http://neurolex.org/) that organizes a neuroscience lexicon. Once data is entered within templates tools can extract it. One large-scale effort is DBpedia (http://dbpedia.org/) that extracts data from Wikipedia and presents online services for queries on the structured data (Auer et al., 2008). We have also been able to extract the structured data from Wikipedia and perform statistical analysis (Nielsen, 2007), and for the Brede Wiki we extract the templates and built an SQL database, that is used for searching after nearby coordinates to a given query. More complex queries can be formed such as 'Find all fMRI papers published between 2003 and 2008 with more than 12 subjects and with coordinates appearing less than 5 mm from $(-40, 0, 30)$'. However, the completeness of the result depends on whether the papers are completely described: If the subject information is not (yet) entered then it cannot be searched, and if the experimental conditions are not defined for the paper it is not possible to search this particular data. MediaWiki templates and categories are defined by the editors rather than the wiki administrator, so this system has inherent extensibility. It is relatively easy to define new templates for, e.g., neuroimaging studies that report their results with respect to brain regions rather than peak coordinates. The MediaWiki software may dump its complete content to an XML file and an application programming interface enables on-line queries. Raw dumps of the Brede Wiki as well as an SQL file with information from its templates are available on the Web. Wikis do typically not support visualization directly. However, extensions to MediaWiki may enable the generation of visualizations, see e.g., the generation of graphs by Dengler et al. (2009). At present, the Brede Wiki calls external Web services, such as ICBM View, for the visualization of individual coordinates.

A recent dispute about data presented by Shmuel and Leopold (2008) has brought forth the issue of ownership to primary neuroimaging data within-laboratory (Fox et al., 2009) but disputes with neuroimaging data sharing between laboratories has a longer history (Aldhous, 2000). The issue of ownership of analysis results and

meta-data, such as bibliographic information, would also be present for a coordinate database. Coordinates themselves are probably not covered by copyright since they do not reach the threshold of originality: Two researchers working independently would get the same set of coordinates given same subject data and the same analysis method. Bibliographic data may be copyrightable. Abstracts can probably be copied as research-based fair use, but they can probably not be copied without permission in commercial contexts, see also the discussion by Dunckley (2009). Other bibliographic data, such as year of publication and page number, can hardly be copyrightable *per se*, however when presented in a database it might gain database rights, — at least in certain jurisdictions. There is no international agreement on the issue of database rights (Sanders, 2006). A 1996 European Union directive protects the maker of a database where the creation constitutes a 'substantial investment' so that a 'substantial part' of the database content cannot be copied without permission. This also covers data that is not copyrightable on its own. In Germany a court ruled that an alphabetic list of just 251 weblinks was protected, and in Denmark a court barred a Web service from deep linking and systematic copying of headlines from news web-sites (Mercado-Kierkegaard, 2006). On the other hand the United States Supreme Court accepted—in the so-called Feist case—that telephone subscriber information could be copied (Sanders, 2006). U.S. National Library of Medicine claims ownership and imposes restrictions on the use of PubMed and may terminate its license (U.S. National Library of Medicine, 2008). Since the NLM license may be terminated NLM data can probably not be merged with data from sources published under non-revokable Creative Commons licenses (http://creativecommons.org). The Brede Database has not yet a formal license but the entire database is available from its homepage for others to include and indeed Hamilton's AMAT database (Hamilton, 2009) has added content from the Brede Database. There has been a tradition in the functional neuroimaging community for relatively open sharing, e.g., of software. The most popular analysis software, SPM, uses a so-called copyleft license. This kind of license encourages mutual sharing and numerous SPM extensions have been written by third-parties. In a database context a copyleft license will ensure that users share their version of and additions to the database if they copy it. This is the notion of share-alike. The only major difference between copyleft and the Creative Commons notion of share-alike is that share-alike may prohibit commercial reuse (the CC-by-nc-sa license) while copyleft licenses always allow it (corresponding to by-sa Creative Commons license). Properly copylefted databases may aid database federation since data can move freely between databases provided they are all copylefted. A private company will not likely use a copyleft license, and users will be left with the search interface that the company can provide, since the data cannot be freely delivered through a third-party search engine. It would be an unfortunate development if neuroimaging result data is hidden behind subscription fees and restrictive licenses, and it will seriously impede the development of novel retrieval and analysis methods. We have setup part of the Brede Database with information from PubMed and included abstracts copyrighted by publishers or authors, so it is questionable if we can put the entire database under a copyleft license. As we have entered the coordinates ourself the database rights for this part belongs to us and we can issue copyleft licenses for that part. When Hamilton uses coordinate data from Brede

Database in her AMAT database and publishes AMAT under a copyleft license her reuse of our database is exactly as we intended. Copyleft and other open licenses have been targeted the development of software code and text. Recently, a licens has been drafted specifically for databases (http://www.opendatacommons.org/-licenses/odbl/) (Miller et al., 2008), and this will be a suitable license for the Brede Wiki.

Wikipedia has shown a tremendous growth. It will be optimistic to think that a specialized wiki such as the Brede Wiki will experience similar growth. However, editing in the Brede Wiki is a big leap forward in comparison with the tedious submission procedure for the Brede Database, which requires downloading a Matlab program, understanding all the data fields, entering the data, storing an XML file and submitting this file.

As neuroscience data are complex it is not entirely clear how neuroinformatics databases should be structured. The original BrainMap database structure as described by Fox et al. (1994) inspired the design of our Brede Database. In many cases the BrainMap framework is sufficient and provides solid performance for standard neuroimaging meta-analysis (Fox et al., 2005). Yet extensions to the framework with the definition of an ontology for brain functions within the Brede Database allowed for automatic coordinate-based mass meta-analysis for all brain functions listed in the ontology (Nielsen, 2005). A brain region ontology allowed us to data mine across all brain regions (Nielsen et al., 2006) as well as link to the CoCoMac database (Kötter, 2004). This latter ontology was not anticipated during our initial design of the Brede Database but was a later addition. The *extensibility* of the database makes such data mining efforts possible.

Mandatory submission could be a goal, but in my opinion mandatory submission should not be undertaken before a system has been setup that is sufficiently easy to use and where the data can move freely between different databases. The Journal of Cogntive Neuroscience at one point required the submission of imaging data to the fMRI Data Center for studies published in the journal. This is no longer required, and it provides an illustration of the issues with mandatory submission of neuroscience data.

Wikipedia has shown how powerful commons-based peer production can be when Web-based technology enables it. Furthermore, collaborative information aggregation and collective prediction can be quite effective and in certain cases better than that of experts: We have experienced it in such diverse cases as humans in a web game and an ensemble of mathematical models (Pennock et al., 2001; Hansen et al., 2001). The neuroimaging community should embrace the notions of open knowledge and collective intelligence for community involvement in managing neuroinformatics data.

Acknowledgment

Thanks to the Lundbeck Foundation for funding and Lars Kai Hansen and Daniela Balslev for comments.

References

Aldhous, P., 2000. Prospect of data sharing gives brain mappers a headache. Nature 406, 445–446.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z., 2008. DBpedia: A nucleus for a web of open data. In: The Semantic Web, Volume 4825 of *Lecture Notes in Computer*

Science. Springer, Heidelberg/Berlin, pp. 722–735.

Bryant, S. L., Forte, A., Bruckman, A., 2005. Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. In: Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work. ACM, New York, NY, USA, pp. 1–10.

Derrfuss, J., Mar, R. A., 2009. Lost in localization: The need for a universal coordinate database. NeuroImage, doi:10.1016/j.neuroimage.2009.01.053.

Fox, P. T., Bullmore, E., Bandettini, P. A., Lancaster, J. L., 2009. Protecting peer review: Correspondence chronology and ethical analysis regarding Logothesis vs. Schmuel and Leopold. Human Brain Mapping 30, 347–354.

Fox, P. T., Laird, A. R., Fox, S. P., Fox, P. M., Uecker, A. M., Crank, M., Koenig, S. F., Lancaster, J. L., 2005. BrainMap taxonomy of experimental design: Description and evaluation. Human Brain Mapping 25, 185–198.

Fox, P. T., Mikiten, S., Davis, G., Lancaster, J. L., 1994. BrainMap: A database of human function brain mapping. In: Thatcher, R. W., Hallett, M., Zeffiro, T., John, E. R., and Huerta, M. (Eds.), Functional Neuroimaging: Technical Foundations. Academic Press, San Diego, California, pp. 95–105.

Hansen, L. K., Nielsen, F. Å., Strother, S. C., Lange, N., 2001. Consensus inference in neuroimaging. NeuroImage 13, 1212–1218.

Huss, III, J. W., Orozco, C., Goodale, J., Chunlei, Batalov, S., Vickers, T. J., Valafar, F., Su, A. I., 2008. A gene wiki for community annotation of gene function. PLoS Biology 6, e175.

Kötter, R., 2004. Online retrieval, processing, and visualization of primate connectivity data from the CoCoMac database. Neuroinformatics 2, 127–144.

Nielsen, F. Å., 2005. Mass meta-analysis in Talairach space. In: Saul, L. K., Weiss, Y., and Bottou, L. (Eds.), Advances in Neural Information Processing Systems 17. MIT Press, Cambridge, MA, pp. 985–992.

Nielsen, F. Å., 2007. Scientific citations in *Wikipedia*. First Monday 12, http://firstmonday.org/issues/-issue12_8/nielsen/

Nielsen, F. Å., Balslev, D., Hansen, L. K., 2006. Data mining a functional neuroimaging database for functional segregation in brain regions. In: Olsen, S. I. (Ed.), Den 15. Danske Konference i Mønstergenkendelse og Billedanalyse. The Department of Computer Science, University of Copenhagen, Copenhagen, Denmark.

Pennock, D. M., Lawrence, S., Nielsen, F. Å., Giles, C. L., 2001. Extracting collective probabilistic forecasts from web games. In: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. SIGKDD, Association for Computing Machinery, ACM, New York, NY, USA, pp. 174–183.

Shmuel, A., Leopold, D. A., 2008. Neuronal correlates of spontaneous fluctuations in fMRI signals in monkey visual cortex: implications for functional connectivity at rest. Human Brain Mapping 29, 751–761.

U.S. National Library of Medicine, 2008. License agreement for NLM data. http://www.nlm.nih.gov/databases-/license/license.pdf.