Lost in localization – but found with foci?!

David C. Van Essen
Department of Anatomy & Neurobiology, Washington University in St. Louis, 660 S. Euclid Avenue, St. Louis, MO 63110, 314-362-7043 (tel), 314-747-3436 (fax), vanessen@brainvis.wustl.edu

Abstract

Commentaries by Derffuss and Mar (2009), Nielsen (2009), Hamilton (2009), and Laird and Fox (2009) agree on the need for a comprehensive database of published stereotaxic coordinates but offer diverse views on how best to achieve this objective. Here, I summarize recent enhancements to the SumsDB database that increase its utility and decrease the impediments to data submission, thereby making it attractive as a resource that can approach comprehensive content in a realistic time frame.

In their commentary "*Lost in localization: The need for a universal coordinate database*", Derrfuss and Mar (2009), argue cogently for a comprehensive database that would provide efficient access to the hundreds of thousands of stereotaxic coordinates that summarize key experimental findings in an estimated 10,000 neuroimaging studies. They noted that existing coordinate databases contain only a modest fraction of the relevant data and also that none (at that time) was matching the pace at which new coordinate data are being published. The core problems are that submitting coordinate data requires substantial time and effort and that the benefits from submitting such data have not inspired widespread voluntary participation by the neuroimaging community. The keys to alleviating this bottleneck are to reduce the effort entailed and to increase the benefits of data submission. This is precisely the objective of recent improvements to the SumsDB database (http://sumsdb.wustl.edu/sums/).

Key features of SumsDB

It is useful to summarize key features of SumsDB and the associated visualization software (Caret and WebCaret), especially since many enhancements were implemented after Derrfuss and Mar submitted their commentary. Features that make SumsDB useful for data mining of coordinates ('foci' in our terminology) fall into five main categories.

1. Flexible search options

The 'Quick-Search' repository in SumsDB (April, 2009) currently contains ~40,000 foci from ~1,300 studies (Fig. 1A) and supports searches based on many types of metadata. These include:

- Spatial location (x,y,z coordinates; cortical sulcus, or cortical area)

- Functions or tasks specified in individual tables or other 'experiment-specific metadata'

- Information about the study as a whole (e.g., abstract, title, keywords).

This enables searches that address questions of the following types:

- "What is known about the function of region 'X', such as area MT+ (Fig 1B)?"

- "What brain regions are involved in processing related to function 'Y', such as music (Fig. 1C) or task 'Z' (e.g., a pitch discrimination task)?"

- What brain regions show abnormalities in structure or function in a particular disease or disorder such as autism (Fig. 1D) or schizophrenia?

Each focus is associated with extensive metadata, immediately viewable by clicking on that focus (arrow in Fig. 1E). There are also direct links from each focus to PubMed and to the online article. Thus, while the initial search results typically include many foci that are irrelevant to the primary question posed, information that is close at hand allows screening of extraneous foci and selection of just the relevant foci.

2. Visualization and analysis

SumsDB is linked to online (WebCaret) and offline (Caret) software that enable visualization of search results on a human brain atlas. Important attributes of the atlas and the visualization software include:

- *Surfaces and volumes*. Search results can be viewed on surfaces and/or volumes, thereby capitalizing on the inherent complementarity of these formats (Fig. 1A). Surfaces include the population-average PALS-B12 atlas of cerebral cortex plus the 'Colin' individual atlas of cerebellar cortex (Van Essen and Dierker, 2007b). Volumes include the PALS-B12 average structural MRI.

- *Many stereotaxic spaces*. The PALS-B12 atlas surfaces and volumes are available in numerous stereotaxic spaces that are widely used in neuroimaging studies (AFNI, FLIRT, all SPM versions, and several others). Coordinates originally reported in a diversity of spaces can be automatically projected to any desired target space while preserving access to the original reported coordinates (Van Essen and Dierker, 2007a). The ability to switch easily from one space to another facilitates accurate comparisons with fMRI activations and other volumetric data reported in different stereotaxic spaces.

- *Reference maps*. Extensive reference data aid in localizing coordinates relative to cortical areas (defined by any of several partitioning schemes), cortical sulci, cerebellar lobules, and various other regions of interest. Reference maps can be updated as additional information becomes available from newly published studies.

- *Online and offline data mining*. WebCaret software allows SumsDB search results to be viewed immediately, without downloading software or data. Specific search results (or the entire Quick-Search Repository) can be downloaded for offline analysis using Caret. This facilitates comparisons with other studies and with experimental data from the investigator's own lab, whether these are viewed as additional foci or as data mapped onto surfaces or volumes.

3. Links to online publications

Rapid access to the original publication is frequently important, in order to assess the relevance of various foci to whatever question the investigator has in mind or to critically evaluate exactly what experiments were done and how they were analyzed and interpreted. To facilitate this process, SumsDB, WebCaret, and Caret all provide links from specific search results to the corresponding online publications (via PubMed, and also directly when possible). For most foci, the table, figure, or page number is specified, thereby allowing the relevant section of the study to be found quickly.

4. Study collections and meta-analyses

SumsDB contains a 'Study Collection Library', in which each study collection points to a thematically unified group of studies and associated foci. Some study collections represent a formal meta-analysis linked to a published review or research article on that topic, such as 'deception' (Christ et al., 2008). Others represent informal, non-comprehensive collections that provide useful entrees to various topics of interest (e.g., 'face perception'). Study collections can be kept private or made public, and they can be easily modified (e.g., to add newly identified studies related to the study collection topic).

5. Federation with other resources

The Neuroscience Information Framework portal (NIF, http://neuinfo.org/) allows efficient searching of a wide range of neuroscience-related resources available on the web. For SumsDB and many other resources, NIF supports 'deep' data mining, wherein relevant database contents (not just the home page) can be directly accessed by queries initiated in NIF. For SumsDB, NIF-initiated queries report specific search results and in addition allow users to immediately link out and view the results using WebCaret.

Efficient data submission

Data submission to SumsDB entails entering three types of data into two curated 'libraries'.

- 'Core' metadata for each study are entered into the Master Study Library and are mainly extracted automatically from PubMed.

  This includes the authors, title, citation, abstract, stereotaxic space, species, and data type (e.g., fMRI or PET).

- 'Experiment-specific' metadata, generally including succinct task characterizations extracted from published tables (and their subheaders), figures and page references, are also entered into the Master Study Library.

- The x,y,z coordinates of each focus are entered into the Foci Library, along with 'focus-specific' metadata (e.g., cortical area or sulcus) and the table (and subheader), figure, or page number.

  Besides the assignments extracted from the original publication, a standard set of assignments derived from reference maps on the PALS atlas are added by a post-submission process carried out by the curators.

For any given data entry, the Master Study Library and Foci Library support multiple versions that differ in their metadata content, providing useful flexibility and updating capabilities. The Quick-Search repository used to expedite routine searches is a distillation that includes a single entry for each focus and each study.

Tutorial and instruction documents (accessible via 'Foci Data Mining' on the SumsDB home page) show how to enter coordinate-related data into SumsDB. Training takes 5-10 hours, depending on initial familiarity with Caret software. After training, data submission typically takes 30 – 60 minutes per study. This is only modestly slower than the ~20 minutes needed for the AMAT database with its 'minimal' metadata requirements (Hamilton, 2009) and is much faster than the extensive task characterizations required by the BrainMap database (Fox et al., 2005; Laird and Fox, 2009). Thus, SumsDB provides an important middle ground, with large value added for a modest data entry effort.

Data submission to SumsDB offers multiple benefits to the submitter:

- submitting foci from publications of your own lab will increase their visibility, through data mining initiated in SumsDB or NIF;

- submitting relevant studies from your research subfield will facilitate cross-study comparisons and promote broader awareness of research in that area;

- individual contributors are recognized by 'provenance' assignments for each study (or version) entered into SumsDB.

- SumsDB libraries can also be used to store foci and study collections for ongoing projects that are not yet published. (Data in these libraries are not made public until requested by the submitter and then vetted by a curator in the Van Essen lab to insure conformance to basic metadata description standards.)

Nielsen (2009) has proposed a wiki-based approach to submitting coordinates and metadata. As noted by Hamilton (2009) and Laird and Fox (2009), this approach may encounter difficulties if it lacks an enforced, coherent metadata structure. Indeed, our experience with SumsDB is that a robust and carefully designed infrastructure is necessary for dealing with various technical complexities. These include avoiding unwanted duplication of foci and studies already in the database while allowing multiple versions of foci and studies when they differ meaningfully in metadata content (e.g., in the description of behavioral tasks). Providing for these and other important needs adds a modest overhead to the data submission process, but yields major benefits in making this a user-friendly resource.

Scaling up and catching up – volunteers needed!

SumsDB libraries have nearly tripled in content over the last 16 months, from 14,000 (~500 studies) in January 2008 to ~40,000 foci (~1,300 studies) in April 2009. This approaches the rate at which new studies reporting coordinates are published and makes SumsDB the fastest-growing of the existing coordinate databases. We anticipate being able to sustain this pace through ongoing curation efforts in the Van Essen lab. However, to accelerate the process and substantially reduce the large backlog, it is vital to enlist volunteers from the neuroimaging community. An attractive and feasible model is for one or two individuals (students, postdocs, or knowledgeable technicians) from each of many laboratories to enter data published by their own laboratory plus selected topics related to that lab's research interests. For example, if 50 volunteers each added ~20 studies per year (15-30 hours per volunteer, including training), the current rate of submission would approximately double, and about half of the relevant literature would be covered in ~5 years.

Psychology and neuroscience courses offer another way to promote data submission. For example, a classroom project to explore a specific aspect of brain function might include analysis of existing studies in SumsDB plus addition of relevant studies from the literature that are not yet in SumsDB. Bearing in mind that our central curation process prevents flawed or invalid entries from becoming public, such efforts can be undertaken by graduate students or even undergraduates in a supervised instructional setting.

Datasets underlying published meta-analyses constitute a particularly attractive source for database submissions. A PubMed search for 'neuroimaging AND meta-analysis' reports 180 meta-analyses, half of which were published since 2006. Most of these meta-analyses (based on inspection of the 20 most recent ones) involved extraction of coordinate data directly from the literature, not from any database. Instructions for handling existing meta-analyses are included in the documentation for submitting coordinate-related data into SumsDB, so as to capitalize on the work already done in extracting coordinate data and to generate a study collection that can be linked to the published meta-analysis.

Looking around the bend

Several developments could further accelerate the pace at which coordinate data are made accessible and useful to the community.

Accelerating foci submission

Greater sharing of data across existing coordinate databases would reduce duplication of effort. Already, SumsDB contains over 5,000 foci (228 studies) extracted from the AMAT database (with permission from and acknowledgment to A. Hamilton); similarly, AMAT includes many entries initially entered in the BREDE database (Nielsen, 2003). Data from the SumsDB Foci Library and Master Study Library are freely available for data mining or incorporation into other databases, with the expectation that usage of SumsDB will be appropriately acknowledged. Open sharing of data would allow each database to capitalize on its unique capabilities. The two largest databases (Brain Map and SumsDB) each have extensive visualization and analysis capabilities that differ greatly, making these resources more complementary than competing. Volunteers would presumably be more willing to contribute coordinate data to their preferred database if they anticipate that the data will soon populate other databases and thus be of broader utility. SumsDB allows credit to be allocated to the database of origin as well as to individual contributors, thus sharing the credit appropriately. Because databases differ in data format and metadata content, sharing can be expedited by providing a schema that characterizes the database structure, as we learned through the process of federating SumsDB with NIF. More generally, the recommendation that the neuroimaging community embrace open knowledge for data mining (Nielsen, 2009) is timely and meritorious.

Greater standardization of how results are tabulated and reported in journal articles could increase the efficiency of extracting coordinate data and metadata. This can begin with modest, incremental steps, though the ultimate objective should be to enable automatic or semi-automatic data extraction. However, without buy-in from the neuroimaging community, journal editors are unlikely to impose steps that might be even modestly burdensome to authors. On the other hand, systematization of how coordinate data are reported and described would benefit authors and journal readers alike (Poldrack et al., 2008 – see 'Figures and tables should stand on their own' section). In short, the potential benefits of data standardization are large, but the timing and the process for building community support need careful attention. Organizations such as the Organization for Human Brain Mapping, and Society for Neuroscience, and the International Neuroinformatics Coordinating Facility could help catalyze this process in response to inputs from their membership.

The iceberg beneath the tip

The conciseness of x,y,z coordinate data is both a strength and a limitation of this data format. Obviously, much greater information about complex spatial and temporal patterns of brain activation is available in the volume and surface data from which foci are extracted. In an explicit comparison of approaches, Salimi-Kharshidi et al. (2009) demonstrated that image-based meta-analyses (IBMA) are more sensitive than coordinate-based meta-analyses (CBMA) in revealing patterns of activation that are consistent across studies. For IBMA to become widely used for meta-analyses, a searchable database repository for image (volume) data from a large number of studies would be extremely useful. SumsDB can already handle volume as well as surface data and thus provides an existing option for this purpose. Indeed, a growing number of studies directly link to datasets in SumsDB, providing access to the underlying data and enabling WebCaret visualization of scenes that replicate the published figures (see 'Publications with links to SumsDB' on the home page). To facilitate efficient data mining, we intend to establish a Surface Library and a Volume Library in SumsDB, analogous to the

existing Foci Library, that will house curated sets of published surface and volume data and associated metadata.

In conclusion, it is useful to place this and other emerging neuroimaging-related efforts in neuroinformatics into perspective relative to the burgeoning field of bioinformatics. Powerful data mining tools for analyzing genome and protein sequencing data have dramatically advanced our understanding of genetics and molecular biology over the past two decades. Progress has been slower in making neuroinformatics approaches useful to the neuroscience community for a variety of technical and sociological reasons (Gardner et al., 2008). Coordinate data from human neuroimaging studies represent 'low-hanging fruit', now ripe for exploitation using increasingly powerful neuroinformatics tools that can accelerate progress in elucidating brain function in health and disease.
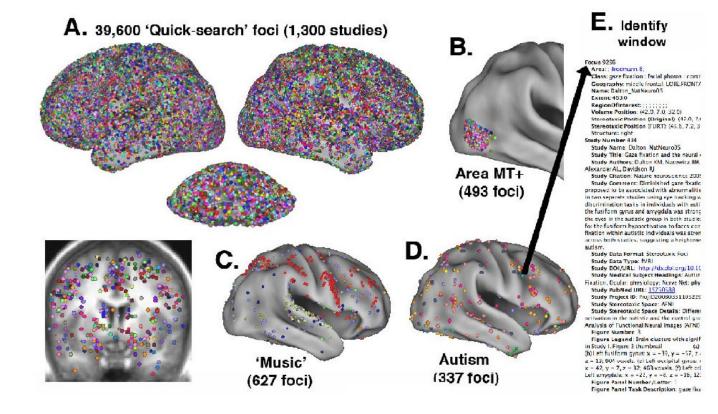
Acknowledgments

References

Christ SE, Van Essen DC, Watson JM, Brubaker LE, McDermott KB. The contributions of prefrontal cortex and executive control to deception: Evidence from activation likelihood estimate meta-analysis. Cerebral Cortex. 2008doi 10.1093. epub

Derffuss J, Mar RA. Lost in localization: The need for a universal coordinate database. Neuroimage. 200910.1016/j.neuroimage.2009.01.053

Gardner D, Akil H, Ascoli GA, Bowden DM, Bug W, Donohue DE, Goldberg DH, Grafstein B, Grethe JS, Gupta A, Halavi M, Kennedy DN, Marenco L, Martone ME, Miller P, Muller HM, Robert A, Shepherd GM, Sternberg PW, Van Essen DC, Williams RW. The neuroscience information framework: A data and knowledge environment for neuroscience. Neuroinformatics 2008;6:149–160. [PubMed: 18946742]

Hamilton A. Lost in localization: a minimal middle way. Neuroimage. 2009in press

Laird AR, Fox PT. Lost in localization? The focus is meta-analysis. Neuroimage. 2009in press

Nielsen, FA. The Brede database: A small database for functional neuroimaging. 9th International Confernece of Functioanl Mapping of the Human Brain; June 19-22, 2003; 2003.

Nielsen FA. Lost in localization: A solution with neuroinformatics 2.0? Neuroimage. 2009in press

Poldrack RA, Fletcher PC, Henson RN, Worsley KJ, Brett M, Nichols TE. Guidelines for reporting an fMRI study. Neuroimage 2008;40:409–414. [PubMed: 18191585]

Salimi-Khorshidi G, Smith SM, Keltner JR, Wager TD, Nichols TE. Meta-analysis of neuroimaging data: a comparison of image-based and coordinate-based pooling of studies. Neuroimage 2009;45:810–823. [PubMed: 19166944]

Van Essen DC, Dierker D. On navigating the human cerebral cortex: response to 'in praise of tedious anatomy'. Neuroimage 2007a;37:1050–1054. [PubMed: 17766148]discussion 1066-1058. PMCID: PMC2045137

Van Essen DC, Dierker DL. Surface-based and probabilistic atlases of primate cerebral cortex. Neuron 2007b;56:209–225. [PubMed: 17964241]

Figure 1.
A. 39,600 foci in the SumsDB Quick-Search repository, projected onto atlas surfaces (inflated PALS-B12 cerebral cortex and Colin cerebellum) and a coronal slice through the atlas volume. Each focus is colored according to the study from which it originated. B. 493 foci localized to area MT+, displayed on the PALS-B12 right hemisphere. C. 627 foci associated with studies containing the term 'music' in the abstract, keywords, or other metadata. D. 337 foci associated with studies referring to autism. E. Extensive metadata from a particular focus related to autism reported in the WebCaret 'Identify window'.