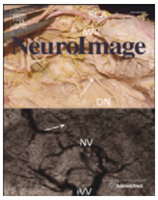
NeuroImage 124 (2016) 1242–1244

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg



NeuroVault.org: A repository for sharing unthresholded statistical maps, parcellations, and atlases of the human brain

Krzysztof J. Gorgolewski [a,j,*], Gael Varoquaux [b], Gabriel Rivera [c], Yannick Schwartz [b], Vanessa V. Sochat [a,d], Satrajit S. Ghosh [e], Camille Maumet [f], Thomas E. Nichols [g], Jean-Baptiste Poline [h], Tal Yarkoni [i], Daniel S. Margulies [j], Russell A. Poldrack [a]

[a] *Department of Psychology, Stanford University, Stanford, CA, USA*
[b] *INRIA Parietal, Neurospin bat 145, CEA, Saclay, 91191 Gif sur Yvette, France*
[c] *InfoCortex UG, Frankfurt am Main, Germany*
[d] *Program in Biomedical Informatics, Stanford University, Stanford, CA, USA*
[e] *McGovern Institute for Brain Research, Massachusetts Institute of Technology, MA, USA*
[f] *Warwick Manufacturing Group, University of Warwick, Coventry CV4 7AL, UK*
[g] *Department of Statistics, University of Warwick, Coventry CV4 7AL, UK*
[h] *Helen Wills Neuroscience Institute, University of California at Berkeley, CA, USA*
[i] *Department of Psychology, University of Texas at Austin, TX, USA*
[j] *Max Planck Research Group for Neuroanatomy and Connectivity, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany*

ARTICLE INFO

Article history:
Accepted 5 April 2015
Available online 11 April 2015

ABSTRACT

NeuroVault.org is dedicated to storing outputs of analyses in the form of statistical maps, parcellations and atlases, a unique strategy that contrasts with most neuroimaging repositories that store raw acquisition data or stereotaxic coordinates. Such maps are indispensable for performing meta-analyses, validating novel methodology, and deciding on precise outlines for regions of interest (ROIs). NeuroVault is open to maps derived from both healthy and clinical populations, as well as from various imaging modalities (sMRI, fMRI, EEG, MEG, PET, etc.). The repository uses modern web technologies such as interactive web-based visualization, cognitive decoding, and comparison with other maps to provide researchers with efficient, intuitive tools to improve the understanding of their results. Each dataset and map is assigned a permanent Universal Resource Locator (URL), and all of the data is accessible through a REST Application Programming Interface (API). Additionally, the repository supports the NIDM-Results standard and has the ability to parse outputs from popular FSL and SPM software packages to automatically extract relevant metadata. This ease of use, modern web-integration, and pioneering functionality holds promise to improve the workflow for making inferences about and sharing whole-brain statistical maps.

Introduction

There is a long history of data sharing in neuroimaging: beginning with precursors such as fMRIDC (Van Horn et al., 2001), transitioning to clinically-focused efforts (ADNI (Weiner et al., 2012), NDAR (Hall et al., 2012), FBIRN (Glover et al., 2012)), and most recently moving to fully open databases (OpenfMRI) (Poldrack et al., 2013) (for review see (Poline et al., 2012)). Sharing data has led to new discoveries (Cai et al., 2014), and has been instrumental in testing new analysis methods (Carp, 2012). Although full, raw datasets provide unprecedented possibilities for analyses that use tools that were not available when the data was initially acquired, sharing comes with a cost. Curation involves precise description of the experimental procedure, requiring significant time and effort especially for task-based fMRI studies. As we argued recently (Poldrack and Gorgolewski, 2014), there is a tradeoff between the amount of effort needed to share a particular type of data and the potential impact the shared data can have. Data types can range from raw datasets that are difficult to share[1] (and subsequently not shared) to their derivatives (peak coordinates), which are included in almost every human brain mapping paper. The challenge of sharing raw data is reflected in our practice: there exist excellent databases dedicated to raw datasets (OpenfMRI and FCP/INDI), however due to various reasons including the time-consuming nature of sharing, these databases capture only a very small fraction of the data from all published research.

* Corresponding author at: Stanford University 450 Serra Mall, Stanford, CA 94305.
E-mail address: krzysztof.gorgolewski@gmail.com (K.J. Gorgolewski).

[1] The issues and barriers of sharing raw neuroimaging data are different for different modalities. For example, task-based fMRI in contrast to resting state fMRI requires additional metadata to describe the stimuli and subject response. Nonetheless, sharing derivatives (such as coordinates or statistical maps) requires less effort than sharing raw data independent of the modality.

http://dx.doi.org/10.1016/j.neuroimage.2015.04.016
1053-8119/© 2015 Elsevier Inc. All rights reserved.

Data sharing can range from commonly shared coordinate data to rarely shared raw datasets, and there are significant drawbacks associated with each of the two. Peaks of statistically significant clusters of activation that are reported in papers have been successfully used to perform meta-analyses (Laird et al., 2005; Yarkoni et al., 2011), however it is easy to imagine a scenario in which a coordinate-based strategy misses subthreshold effects. Further, discarding information that is below threshold is akin to not publishing null results, a dangerous practice that generates publication bias to skew our perception of accumulated knowledge (Rosenthal, 1979). Clearly, there is an opportunity to use an intermediate form of the data between these two extremes that might optimize these constraints.

Using unthresholded statistical maps, an intermediate between raw data and peak coordinates, would provide a significant advance in meta-analytic power. Coordinate-based meta-analysis (CBMA) methods, although more widely used due to the availability of coordinate data, are substantially less powerful than image-based meta-analysis methods (IBMA; meta-analysis based on unthresholded statistical maps; see (Salimi-Khorshidi et al., 2009)). The extended use of CBMA methods can be attributed to the ease of which peak coordinates can be shared in publication tables. Unthresholded statistical maps, in stark contrast, can only be shared by way of ad hoc means (an author's web site or server), and common infrastructure is needed to propagate this as standard practice. While some have advocated for the sharing of statistical maps in the past (Salimi-Khorshidi et al., 2009; Van Essen, 2009; Van Horn, 2003), such an infrastructure has unfortunately not emerged.

NeuroVault is a recently introduced (Gorgolewski et al., n. d.) database that aims to solve these problems. It is a web-based repository that makes it easy to deposit and share statistical maps, parcellations, and atlases of the human brain. It provides attractive visualization and cognitive decoding of the maps that can improve collaborations and readability of the results. At the same time, it also provides an API for researchers to download the data, perform powerful analyses, and build new tools.

Database description

Purpose of the database

NeuroVault was designed to be an easy-to-use repository for statistical maps, parcellations, and atlases of the human brain. It allows users to quickly upload the results of their statistical analyses and share them publicly or with selected colleagues. The focus of the repository is on capturing as many studies as possible, and therefore the submission process does not involve curation, and annotation of the dataset is optional. We justify this by the assumption that the best description of a statistical map is the related paper, a comprehensive document of the methods and metadata associated with the data. Therefore, linking the map to the document object identifier (DOI) of the paper is an efficient way to preserve information that can later be extracted using other methods. One of the reasons why NeuroVault does not focus on data annotation is that there are competing approaches how this should be performed. Some argue that only manual annotation done by experts can lead to meaningful results (Laird et al., 2005). Others prefer to automate the process and extract annotations from papers via machine learning algorithms (Yarkoni et al., 2011). There is also a middle ground where annotations are crowd sourced (http://brainspell.org). As long as NeuroVault provides DOIs of corresponding papers all these methods can be applied.

Current status of the database

NeuroVault is rapidly growing and currently hosts 1356 images spread over 201 collections (47% public) from 280 registered users. Forty of the collections are linked to publications via a DOI (for a detailed description of annotation coverage; see (Gorgolewski et al., n.d.)). The majority of data currently in NeuroVault consists of unthresholded statistical maps (T, F or Z) in MNI space, and the remaining include parcellation maps, atlases, multivariate analysis weight maps, and source localized MEG/EEG maps. All statistical maps uploaded to NeuroVault undergo automated quality control, checking if they are unthresholded or if they roughly align with the MNI template. The results of these checks are reported to submitters and are being exposed through the API.

Data types available

NeuroVault is agnostic to the modality used when computing maps and therefore accepts results from fMRI (task-based and resting state), structural imaging (for example T1 derived voxel based morphometry maps) as well as MEG/EEG (after source localization). In addition, NeuroVault supports NIDM-Results format (http://nidm.nidash.org/specs/nidm-results_020.html) — a bundle that includes contrast maps, masks, design matrix, residuals and other metadata. All data are available in compressed NIFTI format, all images are in MNI space, and each collection and image has a permanent citable URL. In some cases (for example in the case of data derived from OpenfMRI) images are linked to the publicly available raw datasets. Each collection and image can be optionally annotated with population, acquisition, processing and statistical details as well as terms form the Cognitive Atlas (Poldrack et al., 2011). Availability of this metadata depends on the amount of work committed by the researcher submitting the data.

Accessibility

There are two access levels in NeuroVault: public and private. All public collections are accessible (read-only) through the web interface as well as our JSON/XML API, which can be queried by DOI of associated papers. Access to public collections is completely unconstrained and does not require signing agreements or creating an account. Data from public collections are distributed under a Creative Commons (CC0) license (https://creativecommons.org/publicdomain/zero/1.0/). In contrast, private collections are given a randomly generated URL that is first only known to the collection owner, and sharing this address is at his or her discretion. Knowledge of the secret URL is sufficient to access the collection (collaborators do not need to create accounts in NeuroVault to be able to see a shared private collection). Data from private collections are not exposed through the API.

The process of selecting and downloading data is facilitated by the API, which provides both metadata and download URLs for all maps. Researchers can use the API to select an arbitrary subset of the database and programmatically download maps one-by-one. Exposing the maps in this piecewise fashion ensures that if the transfer is interrupted, the download can be resumed from where it left off. All maps are available as compressed NIFTI files to reduce download times. Additionally, we provide a Python package to help with the interaction with the API and data downloads (https://github.com/NeuroVault/pyneurovault).

Anonymization and sensitive data

At the moment, NeuroVault is focused on the collection of group level maps in MNI space. Such data intrinsically do not include details of individual subjects and therefore do not pose a risk of identity disclosure. Therefore, no anonymization procedures are required before depositing group maps into NeuroVault. In the case of researchers submitting single-subject statistical maps, care must be taken on the part of the researcher to use fully anonymous names to describe the individual maps. In contrast to anatomical scans, statistical maps in MNI space do not provide enough spatial details to recognize individuals via their facial features, and therefore do not require defacing procedures.

Long-term sustainability

Although we cannot make any certain claims about the future, we have designed the NeuroVault service to ensure extendibility and robustness to hosting large-scale data. NeuroVault is an open-source project (the code is available at https://github.com/NeuroVault/NeuroVault) that is dependent only on free and open source components (web servers, content management systems, databases, etc.). This means that NeuroVault can be set up by anyone to run on a new server at any given moment. However, software is not the most important part of the project. To preserve the data, we are performing daily offsite backups that are later copied to other locations. Further, we have automated and heavily tested the process of using the code and data to completely restore the service from scratch (https://github.com/NeuroVault/neurovault_puppet). The last component of the service reliability is hardware. It is worth noting that statistical maps take considerably less space than other types of brain imaging data (e.g., raw fMRI datasets). A hard drive (available for $50) can store almost 500,000 statistical maps. Furthermore, the cost of server maintenance and connection to the Internet can easily be leveraged by utilizing existing academic institutions' infrastructures. In short, we argue that despite not being able to guarantee long-term availability of NeuroVault, the nature of its design and data makes the repository easy and cheap to move a new location if the need arises.[2]

Feedback and communication with users

Users can send comments on their experience using NeuroVault by way of a “UserVoice” widget available on every page of NeuroVault.org. The more technically inclined user can post bug reports and feature suggestions on GitHub. For communicating outages, and new features and submissions, we use Twitter and Google +. There is currently no notification system for updating users if data is withdrawn/revised/added to, but we plan to investigate using RSS and mailing lists for that purpose in the future.

Openness to new contributions

NeuroVault is open to contributions of statistical maps, parcellations, and atlases from studies of healthy and clinical populations using almost all neuroimaging techniques. Due to the small file size and simple nature of statistical maps, it only takes 5 min to upload a new dataset and link it to a recent publication. We hope that our efforts in streamlining the interface, providing useful features, and cooperating with neuroimaging organizations and journals have continued and will continue to make NeuroVault grow in the future. The open source model we adopted in developing the repository allows researchers to contribute not only data, but also code (see https://github.com/NeuroVault/NeuroVault for details). More details about the database can be found in (Gorgolewski et al., n.d.).

Acknowledgments

This work was partially funded by the National Institutes of Health (NIH), R01MH096906 [TY] and NSF OCI-1131441 [RP], International Neuroinformatics Coordinating Facility (INCF) and the Max Planck Society [KJG and DSM].

We thank the INCF Neuroimaging Data Sharing task force members for their input during several discussions.

References

Cai, W., Ryali, S., Chen, T., Li, C.-S.R., Menon, V., 2014. Dissociable roles of right inferior frontal cortex and anterior insula in inhibitory control: evidence from intrinsic and task-related functional parcellation, connectivity, and response profile analyses across multiple datasets. J. Neurosci. 34, 14652–14667.

Carp, J., 2012. On the plurality of (methodological) worlds: estimating the analytic flexibility of FMRI experiments. Front. Neurosci. 6, 149.

Glover, G.H., Mueller, B.A., Turner, J.A., van Erp, T.G.M., Liu, T.T., Greve, D.N., Voyvodic, J.T., Rasmussen, J., Brown, G.G., Keator, D.B., Calhoun, V.D., Lee, H.J., Ford, J.M., Mathalon, D.H., Diaz, M., O'Leary, D.S., Gadde, S., Preda, A., Lim, K.O., Wible, C.G., Stern, H.S., Belger, A., McCarthy, G., Ozyurt, B., Potkin, S.G., 2012. Function biomedical informatics research network recommendations for prospective multicenter functional MRI studies. J. Magn. Reson. Imaging 000, 1–16.

Gorgolewski, K.J., Varoquaux, G., Rivera, G., Schwartz, Y., Ghosh, S.S., Maumet, C., Vanessa V Nichols Thomas, Sochat, Poldrack, R.A., Poline, J.-B., Yarkoni, T., Margulies, D.S. 2015 n.d. NeuroVault.org: A web-based repository for collecting and sharing unthresholded statistical maps of the human brain. Front. Neuroinformatics 9. http://dx.doi.org/10.3389/fninf.2015.00008.

Hall, D., Huerta, M.F., McAuliffe, M.J., Farber, G.K., 2012. Sharing heterogeneous data: the national database for autism research. Neuroinformatics 10, 331–339.

Laird, A.R., Lancaster, J.J., Fox, P.T., 2005. BrainMap. Neuroinformatics 3, 65–77.

Poldrack, R.A., Gorgolewski, K.J., 2014. Making big data open: data sharing in neuroimaging. Nat. Neurosci. 17, 1510–1517.

Poldrack, R.A., Kittur, A., Kalar, D., Miller, E., Seppa, C., Gil, Y., Parker, D.S., Sabb, F.W., Bilder, R.M., 2011. The cognitive atlas: toward a knowledge foundation for cognitive neuroscience. Front. Neuroinformatics 5, 1–11.

Poldrack, R.A., Barch, D.M., Mitchell, J.P., Wager, T.D., Wagner, A.D., Devlin, J.T., Cumba, C., Koyejo, O., Milham, M.P., 2013. Toward open sharing of task-based fMRI data: the OpenfMRI project. Front. Neuroinformatics 7, 1–12.

Poline, J.-B., Breeze, J.L., Ghosh, S., Gorgolewski, K., Halchenko, Y.O., Hanke, M., Haselgrove, C., Helmer, K.G., Keator, D.B., Marcus, D.S., Poldrack, R.A., Schwartz, Y., Ashburner, J., Kennedy, D.N., 2012. Data sharing in neuroimaging research. Front. Neuroinformatics 6, 9.

Rosenthal, R., 1979. The file drawer problem and tolerance for null results. Psychol. Bull. 86, 638.

Salimi-Khorshidi, G., Smith, S.M., Keltner, J.R., Wager, T.D., Nichols, T.E., 2009. Meta-analysis of neuroimaging data: a comparison of image-based and coordinate-based pooling of studies. NeuroImage 45, 810–823.

Van Essen, D.C., 2009. Lost in localization — but found with foci?! NeuroImage 48, 14–17.

Van Horn, J.D., 2003. Online availability of fMRI results images. J. Cogn. Neurosci. 15, 769–770.

Van Horn, J.D., Grethe, J.S., Kostelec, P., Woodward, J.B., Aslam, J.A., Rus, D., Rockmore, D., Gazzaniga, M.S., 2001. The Functional Magnetic Resonance Imaging Data Center (fMRIDC): the challenges and rewards of large-scale databasing of neuroimaging studies. Philos. Trans. R. Soc. Lond. B Biol. Sci. 356, 1323–1339.

Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., Harvey, D., Jack, C.R., Jagust, W., Liu, E., Morris, J.C., Petersen, R.C., Saykin, A.J., Schmidt, M.E., Shaw, L., Siuciak, J.A., Soares, H., Toga, A.W., Trojanowski, J.Q., 2012. The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. Alzheimers Dement. 8, S1–S68.

Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., Wager, T.D., 2011. Large-scale automated synthesis of human functional neuroimaging data. Nat. Methods 8, 665–670.

[2] In the event that the database will be hosted by a new institution and managed by a different set of individuals we would either set up a redirection service or transfer the rights to the domain name so existing URLs would point to the right datasets.