The Brain Imaging Data Structure: a standard for organizing and describing outputs of neuroimaging experiments

Krzysztof J. Gorgolewski (krzysztof.gorgolewski@gmail.com)[1], Tibor Auer (tibor.auer@mrc.cbu.cam.ac.uk)[17], Vince D. Calhoun (vcalhoun@mrn.org)[14,15], R. Cameron Craddock[23,24] (cameron.craddock@childmind.org), Samir Das (samir.das@mcgill.ca)[28], Eugene P. Duff (eugene.duff@ndcn.ox.ac.uk)[25], Guillaume Flandin[22] (g.flandin@ucl.ac.uk), Satrajit S. Ghosh (satra@mit.edu)[4,5], Tristan Glatard (tristan.glatard@mcgill.ca)[28,29], Yaroslav O. Halchenko[8], Daniel A. Handwerker[13], Michael Hanke (michael.hanke@gmail.com)[9,10], David Keator (dbkeator@uci.edu)[11], Xiangrui Li (li.2327@osu.edu)[20], Zachary Michael[16], Camille Maumet[19], B. Nolan Nichols (nolan.nichols@gmail.com)[6,7], Thomas E. Nichols (t.e.nichols@warwick.ac.uk)[12,19], Jean-Baptiste Poline (jbpoline@gmail.com)[27], Ariel Rokem (arokem@gmail.com)[2], Gunnar Schaefer (gsfr@stanford.edu)[1,21], Vanessa Sochat[3], Jessica A. Turner (jturner63@gsu.edu)[14,18], Gaël Varoquaux (gael.varoquaux@inria.fr)[26], and Russell A. Poldrack (poldrack@stanford.edu)[1]

1 Department of Psychology, Stanford University, Stanford, CA, USA
2 The University of Washington eScience Institute, Seattle, WA, USA
3 Program in Biomedical Informatics, Stanford University, Stanford, CA, USA
4 McGovern Institute for Brain Research, MIT, Cambridge, USA
5 Department of Otology and Laryngology, Harvard Medical School, Boston, USA
6 Center for Health Sciences, SRI International, Menlo Park, CA, USA
7 Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA
8 Department of Psychological and Brain Sciences, Dartmouth College, Hanover NH, USA
9 Department of Psychology II, Otto-von-Guericke-University, Magdeburg, Germany
10 Center for Behavioral Brain Sciences, Magdeburg, Germany
11 Department of Psychiatry and Human Behavior, University of California, Irvine, USA
12 Department of Statistics, University of Warwick, Coventry, UK
13 Intramural Research Program, National Institute of Mental Health, Bethesda, MD, USA
14 The Mind Research Network, Albuquerque, NM, USA
15 Department of Electrical and Computer Engineering, The University of New Mexico, Albuquerque, NM, USA
16 Squishymedia, Portland, OR, USA
17 MRC Cognition and Brain Sciences Unit, Cambridge, UK
18 Department of Psychology & the Neuroscience Institute, Georgia State University, Atlanta, GA USA
19 WMG, University of Warwick, Coventry, UK
20 Center for Cognitive and Behavioral Brain Imaging, The Ohio State University, Columbus OH 43210, USA
21 Flywheel Exchange, LLC, Minneapolis, MN, USA
22 Wellcome Trust Centre for Neuroimaging, University College London, UK
23 Computational Neuroimaging Lab, Center for Biomedical Imaging and Neuromodulation, Nathan S. Kline Institute for Psychiatric Research, Orangeburg, NY, USA
24 Center for the Developing Brain, Child Mind Institute, New York, NY, USA
25 FMRIB Centre, University of Oxford, Oxford, UK
26 Parietal team, INRIA Saclay, Palaiseau, FR
27 Henry Wheeler Jr. Brain Imaging Center, Helen Wills Neuroscience Institute, University of California Berkeley, CA, USA
28 McGill Centre for Integrative Neuroscience, Ludmer Centre, Montreal Neurological Institute, Montreal, Quebec, Canada
29 Université de Lyon, CREATIS ; CNRS UMR5220 ; Inserm U1044 ; INSA-Lyon ; Université Claude Bernard Lyon 1, France.

Neuroimaging, the study of the brain with medical-imaging devices such as magnetic resonance scanners, is our number one source of quantitative data on brain structure and function. Based on the volume of publication[1], tens of thousands subjects are scanned for research purposes each year. Each study results in complex data involving many files in different formats ranging from simple text files to multidimensional image data, which can be arranged in many different ways. Indeed, a study is typically comprised of multiple imaging protocols and often multiple groups of subjects. To date there has been no consensus about how to organize and share these data, leading researchers, even those working within the same lab, to arrange their data in different and idiosyncratic ways. This is also a major hurdle for neuroimaging data sharing. Lack of consensus leads to misunderstanding and time wasted on rearranging data or rewriting scripts that expect particular file formats and organization, as well as a possible cause for errors. Adoption of a common standard to describe data and its organization on disk can provide multiple benefits:

- **Minimized curation:** Common standards make it possible for researchers who were not directly involved in data collection to understand and work with the data. This is particularly important to ensure that data remain accessible and usable by different researchers over time in the following instances:
  a) within a laboratory over time
  b) between labs facilitating collaboration and making combining data in multi-center studies more clear and less ambiguous
  c) between study public databases (i.e. OpenfMRI) allowing for the quick ingestion of big data organized according to a common scheme.
- **Error reduction:** Errors attributed to the misunderstanding of the meaning of a given datum (e.g., when variable names are not explicitly stated in the data file and standardized across files).
- **Optimized usage of data analysis software** is made possible when the metadata necessary for analysis (i.e. details of the task or imaging protocol) are easily accessible in a standardized and machine-readable way. This enables the application of completely automated analysis workflows, which greatly enhances reproducibility and efficiency.
- **Development of automated tools** for verifying the consistency and completeness of datasets is realized. Such tools make it easier to spot missing metadata that limit how the data could be analyzed in the future.

Previous approaches to neuroimaging data management typically involved complex data management systems[2–8]. However, the challenges associated with installing and maintaining an additional software application, and interacting with one's data primarily through that application, may outweigh the benefits for smaller labs with limited technical resources[9]. In addition, the vast majority of data analysis software requires access to files stored on a hard drive, which is only directly supported by some neuroimaging data management systems. This leads to the need for exporting datasets to a filesystem as the first step of any analysis[10].

The goal of previous databasing approaches has been to efficiently store and manage data rather than creating a standard for describing and standardizing it. In contrast, the XML-based Clinical Experiment Data Exchange schema (XCEDE)[11] attempted to provide a standard for describing results of clinical, including neuroimaging, experiments (independent of any particular databasing system). The approach used by XCEDE employs the eXtensible Markup Language (XML)[12] to provide a hierarchical description of a dataset. This description includes location of every data file along with metadata. Due to the fact that location of files is decoupled from their purpose, XCEDE supports any arbitrary arrangement of files on the hard drive (or even remote locations). In

addition, it does not provide any recommendation on the choice of the file format that the imaging data should be stored in (which puts burden of data conversion on the shoulders of tool developers). Unfortunately, the XCEDE format was not widely adopted, although a number of useful tools were developed[13–15]. We suspect that the combination of extensive use of XML (which is hard to use for scientists without informatics expertise), lack of specification of the file format details, as well as relatively limited support in data analysis packages all may have contributed to the low adoption rate.

In a similar fashion to XCEDE, the OpenfMRI database[16] introduced a dataset description format to fulfill the needs of data curation and dissemination. It relies heavily on specific file naming schemes and paths to convey the functions of files, which allows the application of automated analysis workflows for the entire processing stream. The use of specific filenames and paths can be initially viewed as a limitation in contrast to XCEDE, but it makes it much easier to write software to analyze the data since it does not require consulting additional files (XML descriptions) to understand the purpose of a particular file. In addition, the OpenfMRI standard uses the Gzip compressed version of the Neuroinformatics Informatics Technology Initiative (NIfTI) format[17] to achieve a balance of data analysis software compatibility and file size. This avoids the need to convert between file formats as may have been necessary with XCEDE since it did not specify a particular file format (given that many software tools are limited in the formats that they can read). An important limitation of the OpenfMRI standard is that it had no explicit support for a number of important data types including physiological recordings, diffusion weighted imaging, or field maps, and also had no formal scheme to accommodate longitudinal studies with multiple visits. Despite these limitations and the fact that the OpenfMRI standard was designed to fulfill the needs of one particular repository, it has provided a unified and simple way to organize and describe data. This led to it being adopted (with some modifications) as an internal standard for organizing data in a number of laboratories as well as support by Nipype workflow engine[18].

When developing the Brain Imaging Data Structure (BIDS) standard, proven parts from the aforementioned standards were combined with common laboratory practices to maximize ease of use and adoption. Common practices included encoding the purpose of a file in its filename and reusing already existing and widely recognized file formats (NIfTI, JavaScript Object Notation [JSON], and Tab Separated Value [TSV] text files). The process of defining this standard involved consultations with leading scientists in the field, public calls for comments, and most importantly the generation of example BIDS compatible versions of publicly available MRI datasets. The resulting specification is intentionally based on simple file formats (often text-based) and folder structures. This is done to reflect common lab practices in the community and to make it accessible to a wide range of scientists with limited technical backgrounds. Additional metadata (e.g. acquisition details) are stored in JSON files[19]. JSON is arguably easier to write and comprehend than XML[20], is widely supported by major programming languages, and can be linked to formal ontologies (e.g., Cognitive Atlas[21], Cognitive Paradigm Ontology[22], and NIDM[23]) via JSON-LD[24].

Evolution of the standard

Work on the Brain Imaging Data Structure began at a meeting of the INCF Neuroimaging Data Sharing Task Force (wiki.incf.org/mediawiki/index.php/Neuroimaging_Task_Force) held at Stanford University on January 27-30th 2015. While a flexible solution using the PROV W3C model (http://www.w3.org/TR/prov-overview/) was first investigated, it was acknowledged that this technology would be only viable if tools were in place to write the associated meta data. Since experimental data are obtained from multiple tools, a solution accessible to most

neuroimagers was designed. An initial draft was heavily inspired by the data structure used by the OpenfMRI database, but soon evolved beyond backward compatibility. After the initial draft was formed, a series of discussions and public calls for feedback were conducted. Further refinement of the standard was facilitated by a meeting held during the OHBM conference in Honolulu in June 2015. The discussion over the standard involved domain researchers, computer scientists, MRI physicists, methods developers, data curators and database developers. The first Release Candidate was published on September 21st along with 22 example datasets, online and command line validation tools, and a converter from OpenfMRI standard (https://github.com/INCF/openfmri2bids). The standard became official (version 1.0.0) with the publication of this manuscript and we expect to update and extend it through future releases. We encourage everyone to provide feedback on the standard as well as suggestion for new features and support for more data types. Proposed changes will be discussed publicly trying to accommodate the needs of the community. To facilitate this process we have created the http://bids.neuroimaging.io website.

Basic principles

The following standard describes a way of arranging data (see Figure 1) and specifying metadata for a subset of neuroimaging experiments. It follows a simple but carefully defined terminology. The filenames are formed with a series of key-values and end with a file type, where keys and file types are predefined and values are chosen by the user. Some aspects of the standard are mandatory. For example, each dataset needs to have at least one subject directory. Some aspects are regulated but optional. For example, T1-weighted scans do not need to be included, but when they are available they should be saved under a particular file name pattern specified in the standard.

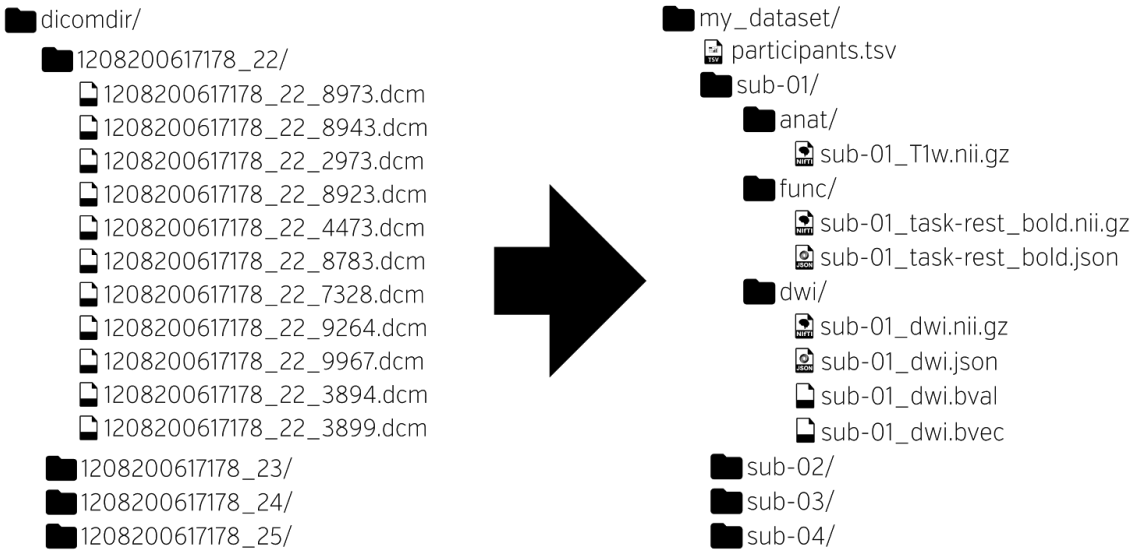

Figure 1. BIDS is a protocol for standardizing and describing outputs of neuroimaging experiments (left) in a way that is intuitive to understand and easy to use with existing analysis tools (right).

This standard aspires to describe a majority of datasets, but acknowledges that there will be cases that do not fit the present version (1.0.0) of BIDS. In such cases one can include additional files and subfolders to the existing folder structure following a set of general naming guidelines and common sense. For example, one may want to include eye tracking data in a vendor specific format that is not covered by this standard. A sensible place to put it

is next to the continuous recording file with the same naming scheme but different extensions. The solutions will vary from case to case and publicly available datasets will be reviewed to include common data types in the future releases of the BIDS specification.

Raw vs. derived data

BIDS in its current form is designed to standardize (convert to a common file format) and describe raw data. During analysis, such data will be processed and intermediate as well as final results will be saved. Derivatives of the raw data[1], should be kept separate from the raw data. This clearly separates raw from processed data, makes sharing of raw data easier, and prevents accidental changes to the raw data. Even though BIDS specification currently does not contain a particular naming scheme for different data derivatives (correlation maps, brain masks, contrasts maps, etc.) we recommend keeping them in a separate "derivatives" folder with a similar folder structure as presented below for the raw data. For example: derivatives/sub-01/ses-pre/mask.nii.gz. In the future releases of BIDS we plan to provide more detailed recommendations on how to organize and describe various derivatives.

The Inheritance Principle

Any metadata file (e.g., files ending with: .json, .bvec, _events.tsv, _physio.tsv.gz, and _stim.tsv) may be defined at one of four levels (in hierarchical order): MRI acquisition, session, subject, or dataset. Values from the top level are inherited by all lower levels unless they are overridden by a file at the lower level. For example, /task-nback_bold.json may be specified at the dataset level to set Time of Repetition (TR) for all subjects, sessions and runs. If one of the runs has a different TR than the one specified in the dataset level file, a /sub-<subject_id>/sub-<subject_id>_task-nback_bold.json file can be used to specify the TR for that specific run.

File Formats

Imaging files

Imaging data are stored using the NIfTI file format, preferably Gzip compressed NIfTI files (.nii.gz), of either version 1.1 or 2.0. We recommend converting imaging data to the NIfTI format using a tool that provides as much of the NIfTI header information (such as orientation and slice timing information) as possible. Since the NIfTI standard offers limited support for the various image acquisition parameters available in DICOM files, we also encourage users to provide additional meta information extracted from DICOM files in a sidecar JSON file (with the same filename as the .nii.gz file, but with a .json extension). Extraction of a minimal set of BIDS compatible metadata can be performed using dcm2niix (https://www.nitrc.org/projects/dcm2nii/) and dicm2nii (http://www.mathworks.com/matlabcentral/fileexchange/42997) DICOM to NIfTI converters. We encourage combining those tools with other software to extract additional metadata. A provided validator (https://github.com/INCF/bids-validator) will check for conflicts between the JSON file and the data recorded in the NIfTI header.

Tabular files

Other files are saved as tab-delimited values (.tsv) files, similar to comma-separated value (CSV) files where commas are replaced by tabs. Tabs must be true tab characters and not a series of space characters. Each TSV file needs to start with a header line listing the names of all columns (with the exception for physiological and other

[1] For the purpose of this standard we consider NifTi files as "raw data" even though they are obtained by converting the data directly produced by the scanner, even though some metadata may be removed.

continuous acquisition data - see full specification for details). String values containing tabs should be escaped using double quotes.
Missing values should be coded as “n/a”.

Key/value files (dictionaries)

JSON files will be used for storing key/value pairs. Extensive documentation of the format can be found at http://json.org. Several editors have built-in support for JSON syntax highlighting that aids manual creation and editing of such files. An online editor for JSON with built-in validation is available at http://jsoneditoronline.org. JSON files need to be encoded in in ASCII or UTF-8. The order of keys is arbitrary and should does not convey any meaning.

Limitations, future work and extensions

Since BIDS was designed to maximize adoption, it heavily relies on established file formats such as NIfTI and bvec/bval (see the protocol for details). This decision was made because those file formats are widely supported by neuroimaging software. Using other file formats (such as DICOM which is closer to the scanner output or HDF5 which is much more flexible and allows for storing all metadata) would result in a more concise and robust data structure, albeit at the cost of additional software development necessary to adapt existing software to the new file format. Storing metadata in JSON files has advantages of accessibility, but can be error prone because data and metadata do not live in the same file. In future revisions of BIDS we will explore the possibility of storing metadata as a JSON text extension of the NIfTI header.

While BIDS enforces NIfTI as the imaging format, we also recognize that specific tools or communities use other neuroimaging file formats such as MINC or NRRD for both technical and historical reasons. Different flavors of BIDS can be designed to support such formats, which would additionally require (1) to identify the metadata fields that should be included in the sidecar JSON file (as opposed to the data file headers), and (2) to modify the validator to read the new file format and check for required and optional metadata. For example, the MINC community is currently working on mBIDS specification that is based on BIDS, but uses MINC file format instead of NIfTI. Even though having multiple flavors of one standard can be problematic and confusing for software developers, those flavors are also necessary to meet the specific needs of some communities. To avoid confusion, any future derivatives of the BIDS specification that are not compatible with the original should be clearly marked in the dataset metadata.

The current release of BIDS does not include support for Electroencephalography (EEG) and Magnetoencephalography (MEG) data, because, at present, there is no single commonly accepted data exchange file format for such data (akin to NIfTI in neuroimaging). However, we plan to extend the standard with support for EEG/MEG in a future release. Similarly the current version of the standard does not cover Positron Emission Tomography (PET), Arterial Spin Labeling (ASL) and spectroscopy, but those extension will be considered in the future.

Our major focus in the near future will be on extending the software ecosystem supporting BIDS to provide incentives for researchers to use it. So far SciTran (database)[25] and Quality Assurance Protocol (QA toolbox) support BIDS compatible datasets. Work is underway to include BIDS support within XNAT (database), heudiconv

(data organizer), PyMVPA (statistical learning toolbox), automatic analysis (framework for analysing multimodal datasets), C-PAC (resting state analysis toolbox), and CBRAIN (data analysis platform). Furthermore, XNAT, COINS and LORIS databases are planning to support BIDS as an export and/or import option. We also plan to build tools to facilitate conversion to the NIMH Data Archive and ISA-TAB (supported by the journals Scientific Data and Gigascience) format. Wide support from different databases can lead to BIDS becoming a "glue layer" - a common exchange format for moving data across databases as NIfTI format became for data files in neuroimaging research, thus facilitating data sharing.

This commentary serves only as an introduction to the BIDS standard. For a more comprehensive description of all aspects of the specification, example datasets, a list of resources, and pointers on how to give feedback on future releases please visit http://bids.neuroimaging.io.

ACKNOWLEDGEMENTS

This work has been supported by the International Neuroinformatics Coordinating Facility (INCF) and by the Laura and John Arnold Foundation. VDC has been supported by in part by NIH P20GM103472. TEN and CM have been supported by the Wellcome Trust. NN has been supported by NIH NIAAA [1 U01 AA021697]. JBP, BNN, and RAP have been supported by NIH NIAAA OD [1 U01 AA021697-04S1].

REFERENCES

1. Smith, K. Brain imaging: fMRI 2.0. *Nature* **484,** 24–26 (2012).
2. Marcus, D. S., Olsen, T. R., Ramaratnam, M. & Buckner, R. L. The extensible neuroimaging archive toolkit. *Neuroinformatics* **5,** 11–33 (2007).
3. Das, S., Zijdenbos, A. P., Harlap, J., Vins, D. & Evans, A. C. LORIS: a web-based data management system for multi-center studies. *Front. Neuroinform.* **5,** 37 (2011).
4. Scott, A. *et al.* COINS: An Innovative Informatics and Neuroimaging Tool Suite Built for Large Heterogeneous Datasets. *Front. Neuroinform.* **5,** 33 (2011).
5. Book, G. A. *et al.* Neuroinformatics Database (NiDB)--a modular, portable database for the storage, analysis, and sharing of neuroimaging data. *Neuroinformatics* **11,** 495–505 (2013).
6. Van Horn, J. D. & Toga, A. W. Is it time to re-prioritize neuroimaging databases and digital repositories? *Neuroimage* **47,** 1720–1734 (2009).
7. Ozyurt, I. B. *et al.* Federated web-accessible clinical data management within an extensible neuroimaging database. *Neuroinformatics* **8,** 231–249 (2010).
8. Keator, D. B. *et al.* A National Human Neuroimaging Collaboratory Enabled by the Biomedical Informatics Research. *IEEE Trans. Inf. Technol. Biomed.* **12,** 162–172 (2008).
9. Nichols, B. N. & Pohl, K. M. Neuroinformatics Software Applications Supporting Electronic Data Capture, Management, and Sharing for the Neuroimaging Community. *Neuropsychol. Rev.* **25,** 356–368 (2015).
10. Schwartz, Y. *et al.* PyXNAT: XNAT in Python. *Front. Neuroinform.* **6,** 1–11 (2012).
11. Gadde, S. *et al.* XCEDE: An Extensible Schema for Biomedical Data. *Neuroinformatics* **10,** 19–32 (2012).
12. Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E. & Yergeau, F. Extensible markup language (XML). *World Wide Web Consortium Recommendation REC-xml-19980210. http://www. w3. org/TR/1998/REC-xml-19980210* **16,** (1998).
13. Grethe, J., Taylor, D., Potkin, S. & Birn, F. A general XML schema and SPM toolbox for storage of neuro-imaging results and anatomical labels. *Neuroinformatics* (2006). at <http://www.springerlink.com/index/qh4223m0h175672x.pdf>
14. NITRC: BXH/XCEDE Tools: Tool/Resource Info. at <http://www.nitrc.org/projects/bxh_xcede_tools/>
15. Keator, D. B. *et al.* A General XML Schema and SPM Toolbox for Storage of Neuro-Imaging Results and Anatomical Labels. *Neuroinformatics* **4,** 199–211 (2006).
16. Poldrack, R. A. *et al.* Toward open sharing of task-based fMRI data: the OpenfMRI project. *Front. Neuroinform.* **7,** 1–12 (2013).
17. Cox, R. W. *et al.* A (Sort of) New Image Data Format Standard: NIfTI-1. in *Proceedings of the 10th Annual Meeting of Organisation of Human Brain Mapping* (2003). at <http://nifti.nimh.nih.gov/nifti-1/documentation/hbm_nifti_2004.pdf>
18. Gorgolewski, K. *et al.* Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Front. Neuroinform.* **5,** 13 (2011).
19. Crockford, D. JSON: The fat-free alternative to XML. in *Proc. of XML* **2006,** (2006).
20. Nurseitov, N., Paulson, M., Reynolds, R. & Izurieta, C. Comparison of JSON and XML Data Interchange Formats: A Case Study. *Caine* **9,** 157–162 (2009).
21. Poldrack, R. a. *et al.* The Cognitive Atlas: Toward a Knowledge Foundation for Cognitive Neuroscience. *Front. Neuroinform.* **5,** 1–11 (2011).
22. Turner, J. A. & Laird, A. R. The Cognitive Paradigm Ontology: Design and Application. *Neuroinformatics* **10,** 57–66 (2011).
23. Keator, D. B. *et al.* Towards structured sharing of raw and derived neuroimaging data across existing resources. *Neuroimage* **82,** 647–661 (2013).
24. Sporny, M., Kellogg, G., Lanthaler, M., Group, W. R. W. & Others. Json-ld 1.0-a json-based serialization for linked data. *W3C*

Working Draft (2013).

25. Wandell, B. A., Rokem, A., Perry, L. M., Schaefer, G. & Dougherty, R. F. Data management to support reproducible research. *arXiv [q-bio.QM]* (2015). at <http://arxiv.org/abs/1502.06900>