HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2018 January 31.

Published in final edited form as:

Nat Methods. 2017 January 31; 14(2): 112–116. doi:10.1038/nmeth.4152.

Win–win data sharing in neuroscience

Giorgio A Ascoli, **Patricia Maraver**, **Sumit Nanda**, **Sridevi Polavaram**, and **Rubén Armañanzas**

Center for Neural Informatics, Structures, and Plasticity Krasnow Institute for Advanced Study, George Mason University, Fairfax, Virginia, USA

Abstract

Most neuroscientists have yet to embrace a culture of data sharing. Using our decade-long experience at NeuroMorpho.Org as an example, we discuss how publicly available repositories may benefit data producers and end-users alike. We outline practical recipes for resource developers to maximize the research impact of data sharing platforms for both contributors and users.

Neuroscience is undergoing a period of considerable societal investments, exciting research breakthroughs, and surging popularity[1]. As 'understanding the brain' rises to the top of the philosophical and technological punch list of humankind, a growing recognition has emerged for the central role of a proper digital infrastructure and the technical requirements necessary and sufficient for storing, representing, and analyzing the expected deluge of data necessary for cracking the neural code[2]. As in other scientific domains, digital data sets are pervasive in nearly all subfields of neuroscience, and computers are deeply integrated in daily laboratory practice. However, unlike their colleagues in other biomedical disciplines, many neuroscientists remain ambivalent with respect to the open sharing of experimental data[3]. As a consequence, in this era of worldwide big science and business analytics, brain science trails behind other scientific disciplines in terms of open data initiatives.

Ten years ago we started NeuroMorpho. Org, a website engineered to facilitate unhindered public access to (and free reuse of) all available three-dimensional reconstructions of neuronal morphology from any imaging modality, animal species, developmental stage, preparation design, brain region, or cell type[4]. The project has continuously grown in data content, community usage, and productive synergy with an ecosystem of related tools. What enabled these accomplishments?

Here we offer a constructive reflection on the challenges and opportunities of data sharing in neuroscience and beyond. Based on our decade of experience, we outline the main benefits of individual and big science data sharing, and we propose pragmatic recipes to ensure a win–win outcome for both researchers giving and receiving data. Although inspired by neuronal morphology, our best-practice principles apply broadly to any scientific field and

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

may be of general interest to policy makers, technical developers planning next-generation resources, and scientists aiming to maximize the impact of their investigations.

Scientific impact of neural data sharing

NeuroMorpho.Org is a freely accessible, centrally curated repository of skeletonized axonal and dendritic morphology traced from light and electron microscopy[5]. The detailed characterization of cellular anatomy is essential for elucidating neuronal computation[6]. Digital reconstructions not only increase the reliability of anatomical quantifications[7], but they also enable biologically realistic computational simulations for investigating the neuronal structure–activity relationship. Until recently, however, sustained collaborations between experimentalists and theoreticians were rare exceptions in neuroscience. The laboratories performing reconstructions remained isolated from each other, typically employing proprietary and mutually incompatible data formats. A decade ago, the few exchanges of three-dimensional tracings largely occurred through peer-to-peer communication.

When it launched in 2006, NeuroMorpho.Org aimed to maximize reuse of labor-intensive morphological reconstructions by organizing and freely distributing in a common format the roughly 1,000 neuron tracings we had collected from known colleagues. Although computing power was then relatively limited, an enthusiastic user community almost immediately coalesced around the shared data. Now, the archive hosts data sets from major international efforts and is recommended for data deposition by leading scientific publishers. Version 7.0, released in September 2016, passed the milestone of 50,000 reconstructions, spanning 36 distinct species, more than 200 brain regions, and over 300 cell types contributed by more than 250 laboratories worldwide from nearly 500 peer-reviewed publications. Almost 6 million tracing files (approximately 50 km of reconstructed axons and dendrites) have been downloaded to date (Fig. 1) over hundreds of thousands of unique visits from 153 countries. The data sets have been employed in community initiatives such as DIADEM challenge. org, inspiring the development of tools for data acquisition, analysis, and modeling[8], tools which can also spread to broader scientific domains. In addition, these data have found their way into educational materials such as textbooks, Massively Open Online Courses[9], the China Applied Math Olympiads, a dedicated testimony to the White House Bioethics commission, and multimedia venues including Neuroscience for Kids and the *Scientific American* blog *Brainwaves.*

A more in-depth assessment of data usage through the analysis of cited references demonstrates a multiplicative impact in basic academic research: the number of ‘secondary’ publications already exceeds the number of publications describing the original data sets, and these publications describe a diversity of scientific outcomes. These secondary reconstructions have been used for expected purposes such as comparative quantitative analyses; stereological estimations of potential connectivity; and detailed constraints of biophysical, electrophysiological, or neurodevelopmental models[10]. But unexpected applications have been reported as well, including detailed interpretation of diffusion magnetic resonance brain imaging[11] or estimation of neurological radiation damage[12].

These creative extensions suggest that open data may serve as a catalyst for developing new theories, testable hypotheses, and research designs.

The ‘power of the public’ is also evident in the rising number of breakthrough discoveries enabled by pooled resources that far surpass the capabilities of individual labs or even of institutional initiatives[13]. The potential of these resources can only increase as the amount and diversity of available data becomes statistically representative of the effective knowledge in a given scientific domain.

Despite these encouraging trends, a troubling proportion of authors reporting neuronal reconstructions in peer-reviewed articles flatly decline to share their data upon request. As a consequence, many data sets remain unavailable to the broader research community, causing a waste of time, money, and scientific opportunities.

Recipes for successful data sharing

Effective data sharing platforms constitute the technological infrastructure for delivering information from the source to consumers. At the same time, the resistance to data sharing in many neuroscience subfields gives resource developers the additional responsibility of catalytic mediators. Specifically, database curators should foster ‘virtuous circle’ dynamics in which the benefits of open data availability exceed the costs for both data producers and end-users. Such a crucial task requires a delicate balance of technical work and social engineering. Accordingly, we offer the following recommendations for bioinformatics projects with prominent data sharing components (Box 1).

BOX 1

TIPS FOR WIN–WIN DATA SHARING: THE ROLES OF DATABASE CURATORS

1. Serve the end-users

- Cater to a demonstrated scientific need; complement, rather than duplicate, existing resources
- Maximize interoperability with relevant tools and avoid proprietary formats
- Design and implement intuitive ergonomics and offer simple instructions
- Continuously solicit feedback and progressively improve functionality
- Collect and publish updated statistics on data access, downloads, and reusage

2. Make it easy for data contributors

- Streamline data submission requirements; conversion and standardization are the curators’ job
- Use concise, consistent, specific metadata annotation
- Take responsibility for misunderstandings or suboptimal presentation
- Release the outcomes of all data sharing requests, both fulfilled and declined

Publicly acknowledge the researchers and labs contributing data

3. Keep the eyes on the data

Release all newly available data sets while maintaining accessibility to prior content

Be proactive, persistent, and patient in finding, requesting, and collating data

Establish and maintain appropriate quality standard to maximize research utility

Record and disclose the rationale for and details of any data changes

Diversify your team expertise and plan a realistic budget

The first essential ingredient is serving the community; the reputation of a resource is ultimately linked to its utility, and most users return to a website because they found it useful in the past. Worthwhile endeavors meet clearly identified scientific needs by complementing, not competing with, other resources. Ensuring the interoperability of a repository with relevant tools yields an integrated whole exceeding the sum of its parts. NeuroMorpho.Org synergizes with complementary initiatives such as brain atlases, simulation environments, image repositories, and anatomical knowledge bases[14]. For example, our recently released OntoSearch functionality maps the major metadata dimensions (notably animal species and brain regions) to BioPortal and other relevant sources. This raises the issue of information control, as the database must be kept in sync with those external resources. This process was initially achieved manually at each release, but has been progressively automated via application programming interfaces (APIs).

Accordingly, open community standards are always preferable to proprietary formats, which hamper productive research exchanges. The best standard is that which works in practice for most users (e.g., the SWC format in the case of neuronal morphology), not necessarily the theoretically optimal design. It may be wise to select the most broadly adopted options to encourage convergence on a common standard; if the majority of users adopt the same data format from among the available options, the rest of the community often follows, to everyone’s advantage. Thanks to this priming effect, most new reconstruction and analysis tools are now ‘born compliant’ with this de facto standard. Similar phenomena of ‘success begetting success’ have also benefited other subdomains of neuroscience, such as functional imaging and neuroanatomical registration (as exemplified by the NIfTI standard and Waxholm space, respectively).

Intuitive ergonomics requiring only minimal instructions facilitate widespread adoption. Good designs germinate from simple implementations and progressively build their functionality. NeuroMorpho. Org began with frugal browsing capabilities, a succinct ‘quick-start’ guide, and a list of frequently asked questions. Continuously soliciting external advice is useful both for prioritizing improvements and establishing an invaluable layer of distributed quality checks; even in the best infrastructures, it is not unusual for a user query to point out missing or ambiguously presented data. Most feedback comes from the user community, data providers, and software developers commenting on search, display, and analysis functionalities through mailing lists, conference presentations, and social media. At

the same time, it is also necessary to collect and publish updated statistics on content access, downloads, and reusage. These return-on-investment metrics quantify the communal utility of data sharing for funding agencies and data providers alike. The prospect of demonstrating enhanced impact is a powerful incentive for authors to share their data sets.

The second key ingredient is removal of the critical barriers to sharing. The foremost impediment for potential data contributors is lack of time. Indeed, they have already done their part by producing the data. The process for those willing to share a data set should be as easy as emailing an attachment or clicking an upload link. All necessary conversion, standardization, mapping, and interpretation are the repository curators' job. This is especially important with respect to time-consuming metadata annotation, a problem which may be alleviated by consistent adherence to specific templates[15]. Nevertheless, full credit must go solely to the data producers; resource developers are the intermediaries, not the main actors. The terms of use must unambiguously require anyone using a data set in a publication to cite the original primary reference pertaining to those data, facilitating impact tracking.

Contributors should also have the prerogative to embargo data for a reasonable period (e.g., up to one year) to allow completion of ongoing analyses and follow-up studies. Furthermore, scientists entrusting their data to a repository rightfully expect the opportunity to preview their data sets on a limited-access website before public release and to request revisions or corrections until they are satisfied. The responsibility for misunderstandings or suboptimal presentation rests with the curator. NeuroMorpho.Org contributors most often agree with the presentation and express gratitude for the added value. Occasionally, they disagree with the standardization and enter a constructive mediation with us curators, resulting in further data edits to everyone's satisfaction. Although authors have the option to veto the standardization and withdraw the data any time during this process, in practice this has never occurred in the ten years we have been operating NeuroMorpho.Org.

The last, but not least, ingredient in the recipe for win–win data sharing is an uncompromising focus on the data sets themselves; release of newly acquired content and maintenance of already available data ought to remain the top priorities for curators. To maximize its impact in a given domain, a repository needs to achieve dense coverage for the relevant type of data[16], meaning that if a data set is available, it should be found within the repository. Sadly, data do not spontaneously knock on the door just because a database exists! Good curators cultivate three 'p-virtues': they are proactive, persistent, and patient in finding, requesting, and collating all data sets that investigators are willing to share. We regularly search the literature for any new publications describing neuronal morphology, promptly inviting the corresponding authors to deposit their digital reconstructions[17]. Our communications follow a systematic protocol with a set number of reminders. When data owners express interest in sharing, we work with their schedule to agree on a practical time-line. Furthermore, we release online the outcome of every data sharing request—be it fulfilled, ongoing, or declined. Such transparency may stimulate broader societal adoption of open science practices[18].

Once received, data are processed through a pipeline ensuring format uniformity and compliance with pre-established quality standards. Optimal curation must balance contrasting criteria; a benchmark that is too loose could subsequently backfire into data misinterpretation and erroneous analyses; excessively strict filters would slow down editing and reduce the amount of viable data. Maximization of research utility requires a practical compromise between overly loose and excessively strict criteria. To paraphrase an ancient Italian proverb ("Il meglio è nemico del bene"), better or perfect (perhaps, tomorrow) are the nemesis of good enough (for sure, today). NeuroMorpho.Org never rejects published data as long as they meet the inclusion criteria (only defining data type, not perceived quality). The rationale is that if a data set was judged scientifically useful by peer reviewers, it may be useful again for reanalysis after sharing. This implies a substantial curation investment in standardization and annotation, though the exact effort involvement varies for each data set. Such diversity of sources points to the importance of encouraging the adoption of standards for both data and metadata already at the stage of experimental design. To ensure traceability of data provenance, both original and standardized data remain available along with detailed change logs and explanation of the editing rationale (neuromorpho.org/StdSwc1.21.jsp).

Lessons learned

Although Box 1 distils the key ingredients of our strategy, the social and technological complexity of neuroscience data sharing cannot be reduced to a simple list. The long-term success of NeuroMorpho.Org to date was built on a complex infrastructure for literature mining, data processing, and state-of-the-art information technology. Every element of this workflow relies on two parallel lines of effort. On one side, several skilled computer programmers constantly write or revise code to automate and improve the most repetitive and error-prone steps. These include, but are not limited to, keyword-based literature screening, metadata management, format conversion, systematic quality control, graphic generation, measurement extraction, data ingestion, traffic monitoring, statistics updates, and server deployment. On the other side, a large team of supervised students manually deal with all aspects that require human intervention. Among other tasks, these include article-by-article literature evaluation, contextual interpretation of metadata, interactive visual checking of all morphologies, and editing of idiosyncratic and/or accidental errors. In more general terms, properly combining all ingredients described above into a productive data sharing initiative requires broadly diversified expertise, outstanding personnel commitment, and constant adaptation to continuous changes.

These considerations raise the essential issue of sustainability, since centralized management requires continuous grant support. We have regularly 'probed' the possibility of distributing the effort among contributing labs akin to the GitHub and Apache models in other communities, but up until now the cultural resistance in neuroscience has been insurmountable. Even with the added value of data curation service, and no additional workload for the contributors, still only half of the authors we contact share their data. Bypassing this societal impasse will require a gradual transition toward a distributed model. We are designing a web-based metadata annotation form for matching textual strings extracted from peer-reviewed publications to controlled vocabularies. To ensure terminological consistency, this tool needs to be sufficiently user friendly for data

contributors to adopt systematically. Furthermore, we are adapting the software for quality control (format conversion, error detection, etc.) for online deployment, which would allow data providers to perform in-house self checks before sending the files.

Cascading replication of specific technical infrastructures across neuroscience domains is often impeded by the notorious diversity of data structures in this field, each with its own requirements. Nevertheless, other resources have adopted or adapted similar high-level strategies. For example, ModelDB[19] (a widely employed database of computational neuroscience models) also systematically mines the literature to identify data of interest, then it explicitly invites authors to share, followed by in-house testing and curation before release. Many other data sharing tenets proposed here were also implemented often, though not always fruitfully, in other neuroscience initiatives. Among the best known success stories are big science whole-brain databases such as the Allen Brain Atlas[20] and the Human Connectome Project[21]. Although these enterprises differ in scale, resolution, privacy considerations, funding mechanisms, and support availability, they adopt a centralized curation model. Other early endeavors struggled to reach critical mass in cellular microscopy, functional magnetic resonance imaging, and electro-physiological time series. Despite the initial difficulties, attempts in these subdomains have continued to evolve, productively seeding subsequent developments. Citizen Science crowdsourcing may also provide a suitable alternative model[22], opening the prospect of creating a community of curators. Our specific training protocol of neuroscience and bioinformatics students, piloted as a project-specific intern-ship program, gradually expanded over the years into a formal academic course. The continuous proliferation of big science neuroscience operations[23] is rapidly creating a new recruiting market for these particular skills[24].

Peering around the next decade

With the ongoing global democratization of science, data sharing propels knowledge sharing. Vast sectors of neuroscience are still anchored to a competitive mentality emphasizing data possession. Yet the price of information in a knowledge economy is dictated by the societal value of the information, not by the associated cost of production. This is especially important when considering low-income countries, which are rich in human talent, but largely rely on secondary data on account of limited experimental resources. Therefore, delaying data release reduces the data's potential to lead to discovery, while the first shared data sets carry the most impact.

The complexity of the brain makes public access to digital data particularly crucial in neuroscience. The parallel needs for establishing a comprehensive theoretical framework and multidimensional data repositories might be met by a novel data publishing model. The foundational status of axonal and dendritic morphology in understanding nervous system function suggests a central role for neuronal reconstructions in the impending biomedical and cybernetic revolutions. We thus find it interesting to extrapolate recent trends into the foreseeable future. A linear rate of progress would predict the availability of 2.5 million neuron tracings by 2026, though advances in imaging and reconstruction automation[25] could further accelerate the pace of data collection. This figure falls between the number of neurons in the most popular invertebrate and vertebrate animal models (fruit fly; ~10^5,

mouse; ~10^8). The scientific impact of these data, however, will likely depend as much on quantity as on quality, diversity, and completeness, underscoring the complementary roles of big science international initiatives and individual investigator-initiated projects.

Acknowledgments

We gratefully acknowledge grants R01NS039600 (NIH), R01NS086082 (NIH), and DBI-1546335 (NSF) to G.A.A.

References

1. BRAIN Initiative Working Group. Brain research through advancing innovative neurotechnologies (BRAIN) working group report to the advisory committee to the director. National Institutes of Health; 2014. BRAIN 2025: A scientific vision.
2. Tiesinga P, Bakker R, Hill S, Bjaalie JG. Curr Opin Neurobiol. 2015; 32:107–114. [PubMed: 25725212]
3. Wiener M, Sommer FT, Ives ZG, Poldrack RA, Litt B. Neuron. 2016; 92:617–621. [PubMed: 27810004]
4. Ascoli GA. Nat Rev Neurosci. 2006; 7:318–324. [PubMed: 16552417]
5. Ascoli GA, Donohue DE, Halavi M. J Neurosci. 2007; 27:9247–9251. [PubMed: 17728438]
6. Denk W, Briggman KL, Helmstaedter M. Nat Rev Neurosci. 2012; 13:351–358. [PubMed: 22353782]
7. Scorcioni R, Polavaram S, Ascoli GA. Nat Protoc. 2008; 3:866–876. [PubMed: 18451794]
8. Peng H, et al. Neuron. 2015; 87:252–256. [PubMed: 26182412]
9. Chu P, Peck J, Brumberg JC. J Undergrad Neurosci Educ. 2015; 13:A95–A100. [PubMed: 25838808]
10. Teeters JL, Harris KD, Millman KJ, Olshausen BA, Sommer FT. Neuroinformatics. 2008; 6:47–55. [PubMed: 18259695]
11. Van Nguyen D, Grebenkov D, Le Bihan D, Li JR. J Magn Reson. 2015; 252:103–113. [PubMed: 25681802]
12. Alp M, Parihar VK, Limoli CL, Cucinotta FA. PLoS Comput Biol. 2015; 11:e1004428. [PubMed: 26252394]
13. Parekh R, Ascoli GA. Neuroscientist. 2015; 21:241–254. [PubMed: 24972604]
14. Parekh R, Ascoli GA. Neuron. 2013; 77:1017–1038. [PubMed: 23522039]
15. Parekh R, Armañanzas R, Ascoli GA. Cell Tissue Res. 2015; 360:121–127. [PubMed: 25653123]
16. Halavi M, et al. Neuroinformatics. 2008; 6:241–252. [PubMed: 18949582]
17. Halavi M, Hamilton KA, Parekh R, Ascoli GA. Front Neurosci. 2012; 6:49. [PubMed: 22536169]
18. Ascoli GA. PLoS Biol. 2015; 13:e1002275. [PubMed: 26447712]
19. McDougal, RA., et al. J Comput Neurosci. 2016. http://dx.doi.org/10.1007/s10827-016-0623-7
20. Sunkin SM, et al. Nucleic Acids Res. 2013; 41:D996–D1008. [PubMed: 23193282]
21. Glasser MF, et al. Nat Neurosci. 2016; 19:1175–1187. [PubMed: 27571196]
22. Roskams J, Popovi Z. Neuron. 2016; 92:658–664. [PubMed: 27810012]
23. Huang ZJ, Luo L. Science. 2015; 350:42–44. [PubMed: 26430110]
24. Grisham W, Lom B, Lanyon L, Ramos RL. Front Neuroinform. 2016; 10:28. [PubMed: 27486398]
25. Acciai L, Soda P, Iannello G. Neuroinformatics. 2016; 14:353–367. [PubMed: 27447185]

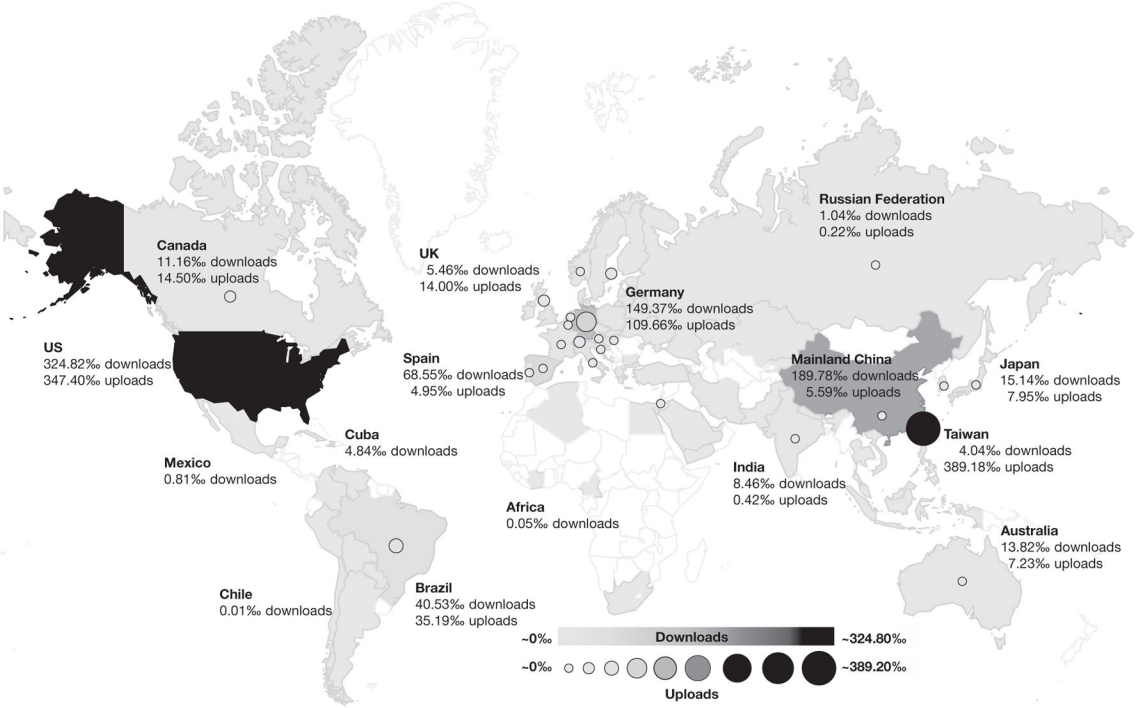

Figure 1.

Worldwide downloads and uploads of digital tracings. Country grayscale shows the fraction (per thousand) of total downloads, bubble size and grayscale indicate the fraction (per thousand) of total uploads, to date. The labels report the per-thousand information numerically for selected countries. Publ. Note: Springer Nature is neutral about jurisdictional claims in maps.