Neuroinform (2012) 10:115–117
DOI 10.1007/s12021-012-9143-4

EDITORIAL

Twenty Questions for Neuroscience Metadata

Giorgio A. Ascoli

Published online: 22 February 2012

Finding relevant data in the biomedical literature can be difficult sometime. To select a few neuroscience examples, suppose you would like to know (a) which serotonin receptor subunits are expressed in dentate gyrus mossy cells; (b) the average volume of the amygdala in adult male chimpanzees; (c) whether there is an EEG signature of Creutzfeld-Jacob disease; or (d) studies reporting bilateral fMRI activity in Brodmann area 38. These are fairly simple questions, for which standard search engines should fare relatively well. Yet even for these kinds of questions, securing the relevant information takes much longer than googling up the local weather forecast for the week-end or tomorrow's commuter train schedule.

The actual data required to build biologically realistic computational models are often more detailed: what is the time constant of the excitatory synaptic current from a specified pair of neuron types? Finding the answer in this case might require many hours or even days of queries over multiple search engines. Most importantly, the results of these queries must be typically followed by at least cursory reading of dozens of papers. When the graduate student triumphantly brings to the lab the needed reference, the adviser could mumble without lifting the eyes from the keyboard "that's in young animals, and it was recorded at room temperature". Another unfortunate major limitation is that, until and unless a definitive answer is found, it is usually impossible to know whether the information is available or not. In other words, existing biomedical search engines are ill-equipped to inform users that something is not yet known.

Standard search algorithms such as PubMed are less than ideal to deal with data identification, because they are ultimately based on matching strings or concepts that appear in the title, abstract, and the keywords. These texts, however, are written with narrow scientific agendas in mind. The authors cannot possibly provide a list of keywords that would encompass all research projects for which some data in their articles might be relevant. If the topic of a report is the molecular phenotyping of a new genetic model of schizophrenia, the technical details of the deconvolution algorithm to deblur the optical micrographs would be nearly impossible to pick up through keyword searches. Could we devise a procedure to interrogate the scientific literature so as to extract accurately and efficiently most if not all of the relevant data? Is there a literature mining protocol that could give us the confidence that, if the query returns a blank, it means that the sought data is not yet available?

Although common to all of biomedical science, this issue is particularly critical in neuroscience because of its unmatched diversity of dimensions, scales, questions, approaches, and techniques. Thus, effective tagging of publications with relevant metadata remains an outstanding neuroinformatics challenge. Full text searches provide half of the solution, in that they eliminate many of the issues related to false negatives. Many of the helpful terms to identify relevant data, for example, appear in the *Materials and Methods* sections of published articles rather than in their titles, abstracts, and keywords. Search engines scanning through the entire main text of publications include early visionary projects such as Textpresso[1], which started within the limited domain of C. elegans, then expanded to a

G. A. Ascoli (✉)
Krasnow Institute for Advanced Study, George Mason University, Fairfax, VA, USA
e-mail: ascoli@gmu.edu

[1] Müller, H. M., Kenny, E. E., & Sternberg, P. W. (2004). Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol., 2*(11), e309.

neuroscience-wide corpus (textpresso.org/neuroscience), and now added drosophila, arabidopsis, and pilot bibliographies for the mouse and the amoeba *dictyostelium discoideum*.

The National Library of Medicine's PubMed Central, a resource of the U.S. National Institutes of Health (NIH), and Google Scholar search the full text of freely available publications, including articles published in open access journals, open access articles in other journals, the E-Print ArXiv archives, and the final peer-reviewed journal manuscripts that arise from NIH funds, available under the NIH Public Access Policy (subject of a previous editorial[2] and commentary[3] in this journal). The majority of neuroscience and biomedical data, however, is not available in open access. Recently, publishers have begun to provide complementary (and complimentary) full text search services for all their publications. In particular, Scirus (scirus.com) is an umbrella engine searching publications from Springer (the publisher of Neuroinformatics), Elsevier, and Nature Publishing Group among many others. Other sources are also indexed under the same consortium, including patents, conference material, and doctoral theses. The service is free and does not require licenses to any of the indexed sources. Users are provided with just the citations and abstracts of all positive hits, and must pay to access non-open source full-texts.

Full-text search is an extremely useful advancement for biomedical research, but a few issues remain. First, not all full-text sources and search engines cover article sections that might contain essential information, such as figure legends, footnotes, and supplementary materials. Second, and most importantly, comprehensive access to full text without proper contextual constraints, background knowledge, and expert curation, often causes a data flood. In other words, a query can result in a huge number of hits, most of which are unavoidably false positives. The best solution to date for this problem has been provided by sophisticated page ranking algorithms, including those implemented by Google and by Scirus. Nevertheless, the complexity of these algorithms makes them extremely challenging for practicing researchers to understand, control, and adapt to specific needs.

A complementary solution could be to ask authors to annotate their articles with relevant metadata through online forms. Leveraging modern web technology, a combination of user friendly drop-down menus, radio buttons, and auto-completing free text could enable authors to conveniently enter all information related to any dimensions germane to their report. Pre-populating entries with available thesauri[4, 5, 6] whenever possible would ensure adherence to controlled vocabularies while at the same time minimizing the annotator's effort. The incentive for the authors to undergo this additional step upon acceptance of the paper would be greater exposure, impact, and citations empowered by the enhanced search engines.

The main challenge of such a system is to determine the appropriate metadata dimensions for each article. Few categories of information, such as the animal species and brain region, are likely to be applicable to many different types of studies. Others, such as a particular visualization method or analysis technique, might only be pertinent for a minute fraction of the published literature. In fact, knowing which types of metadata are applicable to a given article is the hardest part. For example, once it is understood that the slicing orientation is relevant to an article, it becomes relatively easy to determine from its full text whether it was sagittal or coronal. The solution to this difficult problem could take the form of a sequence of dynamic entry forms. The first entry form would be the same for all authors. The subsequent entry forms would depend on the answer provided on the previous form(s). For example, the first form could identify the species. If the study involved human subjects, the second form might ask whether it was primarily a behavioral, neuroimaging, clinical, genetic, or post-mortem investigation. If the study was performed on mice, the second form might ask whether it was in vivo or in vitro, or using specific mutants. The third form would be even more specific, and so forth.

Querying authors with a sequence of dynamic entry forms has a charming resemblance with the "twenty question game". The basic set-up is for a player to think of anything, to fix that idea in his/her mind without revealing it to the others, and to be ready to respond truthfully to a series of yes/no question. The other player(s) have twenty questions to guess the "secret" idea of the first player. Even if it might seem hard at first to read someone's mind with a few yes/no questions, just a bit of experience with the game is enough for most players to start guessing right in less than twenty questions. The first few questions are usually general: does the idea correspond to an abstract concept or a

[2] Ascoli, G. A. (2005). Looking forward to open access. *Neuroinformatics., 3*, 1–4.

[3] Bug, W. (2005). The impact of the NIH public access policy on literature informatics: What role can the neuroinformaticists play? *Neuroinformatics., 3*, 81–91.

[4] Gardner, D., Goldberg, D. H., Grafstein, B., Robert, A., & Gardner, E. P. (2008). Terminology for neuroscience data discovery: Multi-tree syntax and investigator-derived semantics. *Neuroinformatics, 6*, 161–174.

[5] Bowden, D. M., Song, E., Kosheleva, J., & Dubach, M. F. (2012). NeuroNames: An ontology for the BrainInfo portal to neuroscience on the web. *Neuroinformatics., 10*, 97–114.

[6] Bug, W. J., Ascoli, G. A., Grethe, J. S., Gupta, A., Fennema-Notestine, C., Laird, A. R., Larson, S. D., Rubin, D., Shepherd, G. M., Turner, J. A., & Martone, M. E. (2008). The NIFSTD and BIRN-Lex vocabularies: building comprehensive ontologies for neuroscience. *Neuroinformatics., 6*, 175–194.

concrete object? In the latter case, is it bigger or smaller than a spoon, or than a house? After a dozen of questions, the hypotheses are invariably narrowed down to very specific possibilities: a four-legged wild carnivore animal living on high mountains, or a hand-held sterile blunt metal tool used in medical practice.

The twenty question game could serve as more than just mere inspiration to design dynamic entry forms for neuroscience metadata. Powerful machine learning techniques have been developed to optimize the sequence of questions based on the combination of answers. More importantly, these algorithms can be designed to learn and evolve from past experience. The popular online web site 20q.net, for example, started as an Artificial Intelligence experiment and is powered by a neural network. After more than 80 million games played, and with approximately 50,000 hits a day, 20q.net outperforms most twenty question connoisseurs, correctly guessing the human player's hidden thought the vast majority of times, often in as few as 15 questions.

The 20q.net site offers the classic game in 22 different languages as well as 16 themed games (the Name game, sports, movies, etc.). It is intriguing to envision a somewhat specialized "PubMed" twenty question game. A human player would select a target article, and the algorithm would ask a series of metadata questions to guess the correct PMID. Since there are more than 21 million records in PubMed, the trivial numeric strategy of asking whether the PMID is higher or lower than the median, quartile, etc. would require around 24 questions. On the basis of the 20q.net performance to date, it is expected that the real solution would use definitely less than 20 questions. Most importantly, the mature algorithm could be reverse-engineered to help solve the metadata selection problem. In particular, the winning strategy would constitute an excellent initial design in a dynamic entry form system for authors to annotate new articles.

Realistically, author-entered metadata are unlikely to completely solve the challenge of exposing *all* the data in a paper to search engines. Annotating every last detail that is less important in the context of the given article, but which might nonetheless be relevant in other scientific projects, would possibly require hundreds of terms in very specific combinations. A systematic and comprehensive solution will ultimately necessitate nearly-full automation. This will certainly demand considerable advances in text mining, semantic analysis, and deep cross-integration of data, knowledge, and information. Initial efforts such as the Neuroscience Information Framework[7, 8] and the BioLexicon[9] planted promising seeds in these directions, but much progress is still needed. For the time being, and in the foreseeable future, author annotation is still important and its integration with full-text searches and dynamic entry forms might prove to be especially valuable in neuroscience.

[7] Gardner, D., Akil, H., Ascoli, G. A., Bowden, D. M., Bug, W., Donohue, D. E., Goldberg, D. H., Grafstein, B., Grethe, J. S., Gupta, A., Halavi, M., Kennedy, D. N., Marenco, L., Martone, M. E., Miller, P. L., Müller, H. M., Robert, A., Shepherd, G. M., Sternberg, P. W., Van Essen, D. C., & Williams, R. W. (2008). The neuroscience information framework: A data and knowledge environment for neuroscience. *Neuroinformatics., 6*, 149–160.

[8] Gupta, A., Bug, W., Marenco, L., Qian, X., Condit, C., Rangarajan, A., Müller, H. M., Miller, P. L., Sanders, B., Grethe, J. S., Astakhov, V., Shepherd, G., Sternberg, P. W., & Martone, M. E. (2008). Federated access to heterogeneous information resources in the neuroscience information framework (NIF). *Neuroinformatics., 6*, 205–217.

[9] Thompson, P., McNaught, J., Montemagni, S., Calzolari, N., del Gratta, R., Lee, V., Marchi, S., Monachini, M., Pezik, P., Quochi, V., Rupp, C. J., Sasaki, Y., Venturi, G., Rebholz-Schuhmann, D., & Ananiadou, S. (2011). The BioLexicon: a large-scale terminological resource for biomedical text mining. *BMC Bioinformatics, 12*, 397.