

Published in final edited form as:

Neuroinformatics. 2004 ; 2(3): 327–332. doi:10.1385/NI:2:3:327.

Semi-Automated Population of an Online Database of Neuronal Models (Modeldb) With Citation Information, Using Pubmed for Validation

Andrew P. Davison¹, Thomas M. Morse^{2,*}, Michele Migliore^{2,3}, Gordon M. Shepherd², and Michael L. Hines²

¹Centre National de la Recherche Scientifique, 1 Avenue de la Terrasse (Bat. 33) 91198 Gif-sur-Yvette, France

²Department of Neurobiology, Yale University School of Medicine, 333 Cedar Street, New Haven, CT

³Institute of Biophysics, National Research Council, Palermo, Italy

Index Entries

populating databases; literature-search; neuroscience database; interoperability; data entry

Citations play a valuable role when included in medical and scientific databases. They indicate the association of the data with authoritative research reports or broader review articles. Searches based on citations facilitate further inquiry into the literature. Tools and methodologies (see web resources and below) that populated our database, ModelDB (see below), with citation information, are available for investigators and developers who wish to include citation based tools in their databases.

The main purpose of ModelDB, a database of neuronal models publicly available at <http://senselab.med.yale.edu/senselab/modeldb>, is to make available the computer code that describes a model, the need for which has been addressed previously (Migliore et al., 2003; Davison et al., 2003; Mirsky et al., 1998; Miller et al., 2001). ModelDB has grown to contain over 100 models. The usefulness of any database depends on the database being well populated and on having search methods that help the user find the element(s) of interest within that database. We provide several search methods to find models of interest on the ModelDB home page.

Through these search methods links one can arrive at the web page for a model with just two mouse clicks. The web-page for each model provides information about the model including the source code, and tools for finding related models, among which is the ‘Citation Browser’.

The Citation Browser link loads a page as seen in the top of figure 1. This page in turn allows the user to view an annotated list of the titles of (centered headers) and reference bibliographies from (left column) the one or more papers that are associated with the model, and a list of papers that cite the paper(s) (right column) associated with the model. Each paper in ModelDB that is cited by two or more papers in the database is highlighted in green and has a hyperlink to the Citation Browser page for that paper. Each paper that has a model

* Author to whom all correspondence and reprint requests should be sent. tom.morse@yale.edu.

stored in ModelDB is annotated by a bulleted, pink-highlighted hyperlink to that model. Related papers and models may be found just by following the hyperlinks.

The Citation Browser supplies a convenient method for finding papers and models of interest. In order to implement the Citation Browser, it is necessary for the citation lists of at least model associated papers to be entered into the database. Such population would be tedious if done by hand. The only complete electronic public source of citation information, ISI Web of Knowledge (web resources ISI, 2003a), requires a subscription/license and does not indicate if a paper's model is electronically available. There are research-oriented open citation projects (web resources OpCit, 2003b, c), but they are still in preliminary stages. We therefore developed a web-based interface to automate the entry of citation lists obtained from the electronic versions of the original articles (from PDF or HTML files, or from scanning followed by optical character recognition (OCR) 1998).(web resources Hewlett-Packard, 1998). By leveraging open-source citation-parsing tools from the Open Citation Project (web resources OpCit, 2003a), we were able to produce web form interfaces that automate the process of entering citation data from these electronic versions of journal articles, and that also validate the data when they are also in PubMed (PUBMED, 2003a), providing quality assurance.

Automated reference parsing and reference validation using PubMed

The user interface for entering citations is an HTML form (figure 1 B). In summary, the procedure is as follows (see online supplement for a simplified flow-chart and online supplement appendix for some PERL code excerpts):

- The paper for which the reference list is to be entered is retrieved from ModelDB using either its PubMed ID, or its ModelDB Object ID (the latter is for references that are not in PubMed – mostly books and book chapters; see (Marenco et al., 2003) for more detail on the architecture of ModelDB). If the paper is not in the ModelDB database but has a valid PubMed ID, the paper is entered into the database.
- The list of references is pasted into the text box. The principal formatting requirement is that each reference be on a single line.
- Each reference is parsed to extract the list of authors, the name and volume of the journal, and the starting page of the article.
- References that have been successfully parsed are sent as a batch to PubMed, which returns, for each reference, either a PubMed ID or a message that the reference was not found.
- The PubMed IDs are sent back to PubMed, which returns an XML file containing data for each reference. The XML is parsed, the data entered into ModelDB, and the reference linked to the original paper whose reference list is being entered.
- References that have been successfully entered into ModelDB are removed from the text box. References that have not been successfully parsed or that are not in PubMed are left in the box.
- If there is an error in the reference, e.g. incorrect page number, this can be corrected in the text box and the reference re-submitted to PubMed.
- For references that can be found by manually searching PubMed but cannot be successfully parsed, the PubMed ID can be entered instead of the reference string.
- References that resist parsing and/or are not in PubMed can be transferred to a form that allows manual reference entry. After each submit of the manual entry page, the

papers are searched-for in ModelDB and if found a drop-down list box appears with all the papers from the first author. The user may choose to select the paper if it happens to already be in ModelDB.

Please see supplemental online material for implementation and paper scanning details.

Discussion

The task of manually entering a reference to a paper and its citations is tedious. We developed a web-based interface to automate the process of reference entry, interfacing with PubMed for validation and standardization of citation data whenever available in PubMed. When unavailable in PubMed the citations are validated manually using a further web interface.

ModelDB's semi-automated citation entry software has enabled us to populate ModelDB with 400 reference lists consisting of over 10,000 papers and 11,000 distinct author names (Note that the bulk of computational paper references are to the experimental literature). Including the citations in ModelDB allows us to construct a citation database that is focused on the computational neuroscience domain. This modeling citation database allows investigators to use our web application tool, the Citation Browser, to explore how models were used by other researchers. The Citation Browser displays a list of references from the bibliography of the paper(s) associated with a model; each reference is an annotated hyperlink to a paper or a computational neuroscience model if that reference is cited by more than one paper in or has a model available in ModelDB. The Citation Browser displays papers that cited a modeling paper and indicates if the model from any paper is already in ModelDB. This resource is enhanced by our Citation Browser's unique capability to aid researchers in finding models and papers of interest. We hope that this article will help database developers create similar tools to assist the importation of relevant references to the developer's own context. Citations provide an invaluable resource for medical and scientific databases, and indeed for any database that maintains data for which there is a substantial associated literature. The research reports and review articles contain the authoritative description of the item(s) stored in the database and provide the larger context in which the usefulness of the item can be understood. The references in the ModelDB papers also provide a starting place for searches based on citation indices, useful for finding and understanding the work that the paper built upon, and also for seeing which work potentially built upon the work under consideration (which cited the paper).

Citations, because of their varying degrees of relatedness to the article of interest (web resources Garfield, 1994a), are frequently the objects of search and retrieval engines. PubMed is premised on this concept, as are online versions of most journal articles, and indeed, most of modern Library Science. The Institute for Scientific Information (ISI) supplies notable citation search tools. They offer complete lists of citation relationships for multidisciplinary analysis. A common use of citations is the ranking of the importance of papers, individuals, institutions, countries, and journals (web resources ISI, 2003b; Garfield, 1994b) by the number of citations that point to each of these entities. Our citation database is by comparison narrower in focus being limited to publications in the field of computational neuroscience and the references contained therein. We have added value to our Citation Browser's references lists by annotating them with a color coding that indicates if a publication has a model available in ModelDB. This can help investigators who are interested in modeling to know that this paper may have special interest to them.

Future Directions

We are exploring tools that can identify model papers automatically from electronically available journals (Craστο et al., 2003). These same tools will enhance the use of citations by attempting to annotate them automatically with the keywords that we currently use to search for models. Researchers could then more easily find papers on neuronal and network modeling topics.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to acknowledge Luis Marengo and Chiquito Craστο for helping with technical issues, discussions on past and future directions, and proof reading. We gratefully acknowledge the support of the National Institute for Deafness and Other Communicative Disorders, National Institute of Mental Health, the National Institute of Neurological Disorders and Stroke, National Science Foundation, and the National Cancer Institute from their combined Human Brain Project (grant number 5P01DC004732-04).

Web resources

- Garfield, E. Expected Citation Rates, Half-Life, and Impact Ratios: Comparing Apples to Apples in Evaluation Research. 1994a. Available from: <http://www.isinet.com/essays/citationanalysis/10.html> on December 22, 2003
- Garfield, E. The Relationship Between Citing and Cited Publications: A Question of Relatedness. 1994b. Available from: <http://www.isinet.com/essays/useofcitationdatabases/5.html/> on December 22, 2003
- Hewlett-Packard. OCR bundled software: HP OfficeJet G Series Scan version 2.0. 1998.
- ISI. ISI Essential Science Indicators. 2003. Available from: <http://www.isinet.com/media/presentrep/tspdf/sem-esi-1-0-0702.pdf> on December 22, 2003
- OpCit. The Open Citation Project. 2003a. Available from: <http://opcit.eprints.org/> on December 22, 2003
- OpCit. Citebase Search. 2003b. Available from: <http://citebase.eprints.org/cgi-bin/search> on December 22, 2003
- OpCit. ParaCite. 2003c. Available from: <http://paracite.eprints.org/> on December 22, 2003
- PUBMED. National Center for Biotechnology Information, National Library of Medicine (US). 2003a. Available from: <http://www.pubmed.gov> on December 22, 2003
- PUBMED. Pubmed Batch Citation Parser. 2003b. Available from: <http://www.ncbi.nlm.nih.gov/entrez/getids.cgi> on December 22, 2003

References

- Craστο CJ, Marengo L, Migliore M, Mao B, Nadkarni PM, Miller PL, Shepherd GM. Text Mining Neuroscience Journal Articles to Populate Neuroscience Databases. *Neuroinformatics*. 2003; 3:215–238. [PubMed: 15046245]
- Davison, AP.; Morse, TM.; Migliore, M.; Marengo, L.; Shepherd, GM.; Hines, ML. ModelDB: A Resource for Neuronal and Network Modeling, in *Neuroscience databases: a practical guide*. Kotter, R., editor. Kluwer Academic; Boston: 2003. p. xviii-310.[2] p. of plates
- Marengo L, Tosches N, Craστο C, Shepherd G, Miller PL, Nadkarni PM. Achieving evolvable Web-database bioscience applications using the EAV/CR framework: recent advances. *J Am Med Inform Assoc*. 2003; 10(5):444–53. [PubMed: 12807806]
- Migliore M, Morse TM, Davison AP, Marengo L, Shepherd GM, Hines ML. ModelDB Making models publicly accessible to support computational neuroscience. *Neuroinformatics*. 2003; 1:131–34. [PubMed: 15055398]

- Miller PL, Nadkarni P, Singer M, Marengo L, Hines M, Shepherd G. Integration of multidisciplinary sensory data: a pilot model of the human brain project approach. *J Am Med Inform Assoc.* 2001; 8(1):34–48. [PubMed: 11141511]
- Mirsky JS, Nadkarni PM, Healy MD, Miller PL, Shepherd GM. Database tools for integrating and searching membrane property data correlated with neuronal morphology. *J Neurosci Methods.* 1998; 82(1):105–21. [PubMed: 10223520]

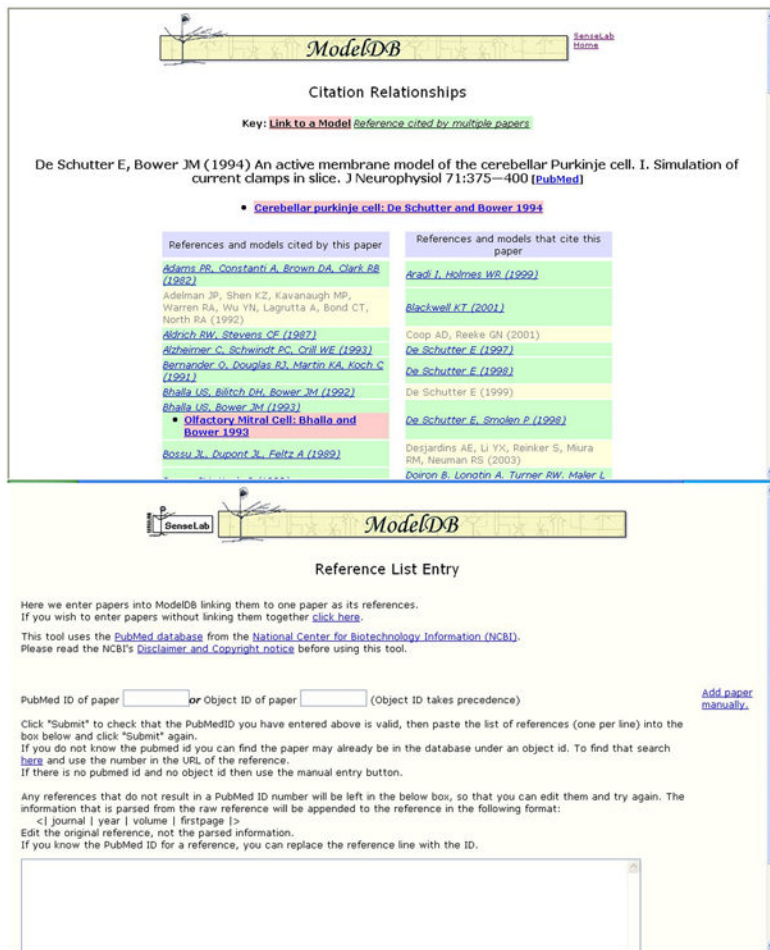


Fig. 1. A. Example of the Citation Browser page. B. The ‘Reference List Entry’ page uses existing PubMed IDs and reference lists from papers to automatically extract PubMed data and enter it into our (Citation Browser) database.