Non-curated distributed databases for experimental data and models

R.C. Cannon, F. W. Howell, N.E. Goddard and E. De Schutter

ABSTRACT

Most databases currently under development in neuroscience are heavily curated, in that substantial effort is required by database builders to read the literature and enter the distilled information in a database. This provides a valuable resource, but may be too labor intensive to scale up to the huge volumes of data being generated at present. Here we consider the opposite end of the databasing spectrum at which there is no curation, minimal standardization, and where data remains under the control (intellectual and physical) of those who collected it. The Web in conjunction with a search engine such as Google takes non-curated distributed databasing to the extreme. Our approach is to impose slightly stricter constraints, specifically the requirement for some form of machine readable records for each data file, in order to improve the efficiency of cataloging and searching. The first goal is to find a level of standardization which is adequate for automated cataloging and searching, without being so strict as to discourage potential data providers. The second goal is to design the system in such a way that intellectual property remains clearly in the hands of the data providers, and is not transferred to database maintainers.
Most of the features commonly associated with curated databases, such as quality control and security can also be developed for non-curated databases though by quite different mechanisms. Here we describe the design and implementation of a pilot scheme focussed on models, code modules and already standardized data such as neuronal morphology files, but with applications to much more heterogeneous data sets.

SUMMARY

In considering the effort involved in creating a database, the dominant issues for the academic research community are sociological issues of who does the work and why. This is different from a commercial environment where the rewards are primarily financial and can be used to direct massive labor investment in tightly curated database systems (e.g. Physiome sciences, {\it www.physiome.com}).
Successful curated neuroscience databases in the public domain are often the work of a small group of dedicated individuals creating a resource which is perceived as their intellectual property, not that of the data providers, normally using information which is already in the public domain in the form of published papers. This immediately explains why the groups remain small - because others have very little interest in contributing to a project for which all the credit will go elsewhere, and there is no accepted method for the redistribution of intellectual property among database contributors other than joint publications. This also gives an indication of how a successful large scale database system might be structured - as a conglomerate of many small units, each controlled by the providers of the data it contains.

Distributed, or ‘‘grass roots’’ databases
would look quite different from existing databases because they

require a modularization of the database functionality to match the facilities and competences of the participants. For example, individual sites need not run a database server or a search engine: these could be provided in a small number of sites run by informaticians. Likewise, analysis and visualization tools need not be part of any one site. It is sufficient that they are cataloged and available at at least one place, and that information about what forms of data are compatible with which tool is also available (though perhaps on a different site controlled by the software users). A major goal in designing a successful distributed system for the academic environment is that it should be composed of distinct interoperable units, where the perception of intellectual property of the whole is distributed to the units commensurately with their real intellectual investment.

After intellectual control, probably the most important issue for distributed databases is the minimization of redundant effort. This amounts to making software do as much as possible of the work - certainly all the work of web site creation, cataloging and cross referencing, but also that of using knowledge bases to resolve differences in the way data is documented and to prompt users to supply the required information when it is not yet available.

Database Architecture

The database architecture being developed is rather similar to the original Napster music sharing system which was shut down in 2001 for copyright reasons. The main difference from other neuroscience databases is that the two tasks - the provision of data and the application of IT expertise - are widely separated, to the extent of even being located in different places. Each data provider uses freely available software to create a local archive, here called a ''data site'' which is hosted on a server of their choice, and remains completely under their control. Such sites are no more than conventional web sites, and require no database engines or search facilities. They simply contain data and corresponding documentation in machine-readable formats.

All the cataloging and searching facilities for collating and cross-referencing data from different sites are provided in one or more IT-centric sites. These sites maintain a list of data providers which is kept up to date by
the data-site building software which sends in the address of each new site that is available. The catalog sites do not copy any data, but only the concise descriptions, or records, associated with data files. From these they can build up a searchable catalog of available data and cross reference it to software or visualization modules from other sites. A user of the database would first visit the catalog sites to search or browse, but would be directed to the original source sites to retrieve any actual data.