How do we get the data to build computational models? [1]

F. Howell (fwh@anc.ed.ac.uk) [a]
R. Cannon (cannon@anc.ed.ac.uk) [a]
N. Goddard (Nigel.Goddard@ed.ac.uk) [a]

[a] *Institute for Adaptive and Neural Computation, Division of Informatics, University of Edinburgh, 5 Forrest Hill, Edinburgh EH1 2QL, Scotland*

Abstract

We present a new approach to building radically distributed databases of neuroscience data. It aims to make available to modelers the huge amount of useful experimental data and notes which currently sits on experimenters PCs and lab notebooks.

The approach has two components. The initial phase is a user friendly desktop application which experimental neuroscientists can use to markup data and build small catalogs of their data.

The second phase is a server application which acts like a "smarter Google" which is able to combine and index catalogs from multiple researchers and labs so that modelers can download local copies of data relevant to their study.

Key words: Data management, databases, XML, NeuroML

1 Introduction

The great success of bioinformatics has been characterised by huge shared databases of genomic and proteomic data, which are available for automated data mining techniques. But this hasn't yet been repeated for Neuroinformatics, where there have been a number of small database projects, but most of the data is still scattered amongst research labs in spreadsheets and files on PCs and CDs, or published in text form as printed graphs in papers.

[1] Supported by: MRC eScience programme

This makes the task of building models of neural systems hard; modelers would like to obtain as much data as possible concerning their system of interest - ion channel kinetics and distribution, neuron morphology and connectivity, intracellular pathways, electrophysiological recordings, etc. in order to derive or validate their models; but the data is currently scattered amongst papers and labs and is hard to obtain without personal contacts (or doing the experiments yourself).

Experimental scientists are also very poorly served by existing software for managing their data, and typically have experimental data stored as data files on PCs, CDs, with notes written in lab notebooks or typed up in spreadsheets. Very few experimental scientists use relational databases to organise their information.

So how can we address the problem of organising and cataloging such a heterogeneous collection of data, for the dual purposes of improving life for experimentalists and making the modeling task more automated and less labour intensive?

The solution must address both the technical and social issues. Standard databasing solutions were not designed to cope with the heterogeneity of data types and rapidity of change which is endemic in neuroscience. There must also be a reward for researchers putting in the effort to structure and annotate their data for others to use.

2 Motivations for modelers and experimental scientists

In trying to build a model of a particular neuron, network or phenomenon, modelers would like to have access to all known data converning neuron structure, membrane dynamics, channel distribution etc etc for the system of interest.

So modelers would like to have access to all the data, which is currently distibuted across a number of labs, disciplines and perhaps even species. Idealy this would be the raw data; fitting parameters to printed graphs in published papers is error prone.

But experimentalists do not have much motivation for putting huge amounts of effort into making databases of their raw data, as they currently get rewards for writing papers and not for data sharing. So any workable solution must start out being of immediate benefit to the individual researcher for organising their experimental data, and the data sharing/publication process must be a minimal or zero effort side effect of what they choose to do anyway.

3 Current options for metadata

Software tools for dealing with metadata about files are surprisingly primitive, considering how universal the problem of managing complex data and interrelations is. Common approaches include:

- packing all the information into the filename (e.g. an image file named `XSD_jan_pGAL4_rat123.jpg` encodes date, technique, animal id, file type etc into the name) and subdirectory location (e.g. `expts/bob/confocal/*.jpg`).
- including metadata into the header of specific file types (e.g. SWC files include information on digitisation format).
- using a spreadsheet like Excel to enter and analyse structured information about experiments.
- by hand in a lab book.

Curiously *databases* are not in common usage by individual scientists for storing their data. Relational databases (e.g. MS Access, Oracle, DB2) are based on theory from the 1970s [2] but are significantly more complex to set up and use than spreadsheets, and don't handle tree-structured data (such as "all the information related to my experiment") well. They currently require large amounts of programming effort to set up, maintain and curate, so typically contain small amounts of high quality filtered information of a restricted set of data types (e.g. the Senselab project [7]). *Electronic lab notebooks* are not yet in common use, as they lack the convenience of paper.

We propose that a workable solution has to (a) provide an immediate reward for the researcher adding metadata, by providing software tools which are as convenient and useful as spreadsheets, and (b) be as radically distributed as the Web and peer-peer applications like Napster.

4 Solution part 1: "Catalyzer"

The first part of our solution is a desktop application which is intended to be used by an individual scientist for organising and marking up all their data. It is both a data entry tool and a browser which provides sorting and searching of all the disparate types of data and metadata. Figure 1 shows a screenshot of the main window.

There is not yet a standard name for this type of application; it has import functions from spreadsheets like Excel, the hierarchical file system and databases. One way to think of it is as a version of Excel which handles tree structures naturally (rather than just 2D grids); another way is to think of it

as a file manager which lets you add extra fields other than just the file name and directory to files (e.g. you could add Species, Solution and Temperature fields to every data file).

Its native file format is XML, but it differs from most applications in that there is *not a fixed schema for data*; whenever you discover that the experiment really needs an extra field to be added (like "RatWeight", "LengthOfTail") you can add this field to the existing form.

Once you've entered data (or imported it from the file system or existing spreadsheets), you can browse/search/sort it by any of the data you've entered. This gives you all the power of a relational database, but lets you manage the data in a tree structure (like a combination of the file system hierarchy with spreadsheets). This is a much simpler and more natural way to structure complex heterogeneous data than forcing you to work with the low level relational model of the 1970s.

If you want to share the data with other members of the lab, or the world, the software will generate a structured website containing the metadata, and optionally upload the raw data files.

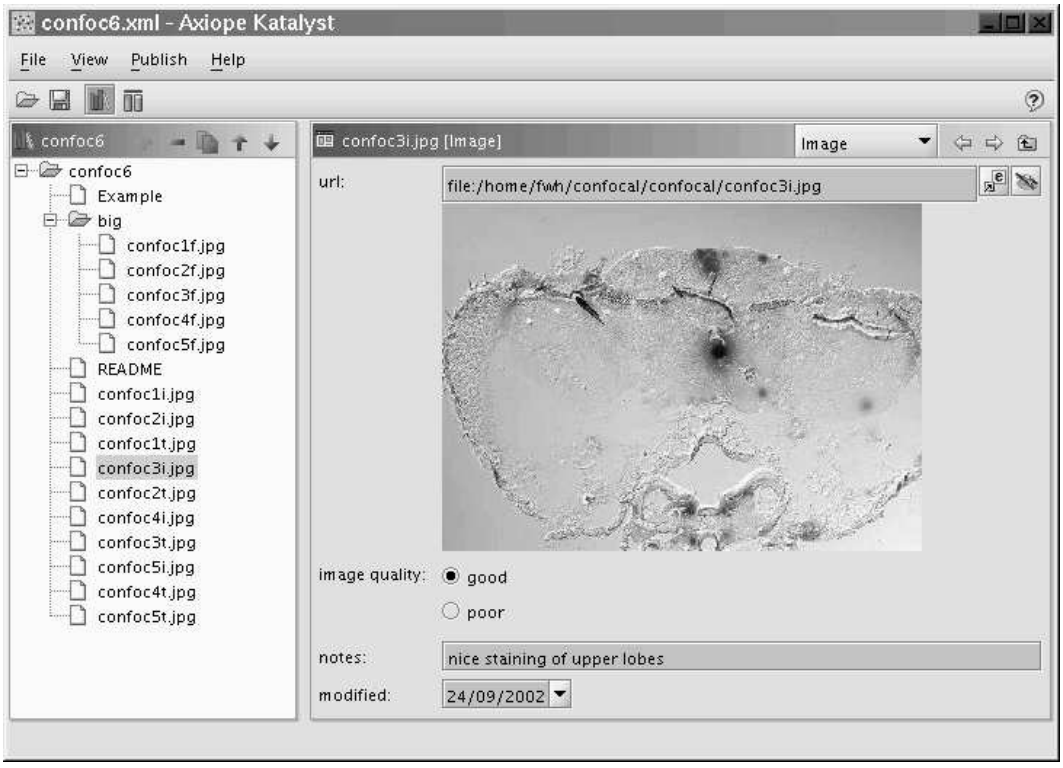

Fig. 1. A screenshot of the Catalyzer window. At the left is a tree view of the catalog structure, and on the right is a view of a single record. Note that the form itself can be edited at any time by adding and removing fields.

5 Solution part 2: the server

The "Catalyzer" application is targeted at handling the needs of an individual researcher in cataloging and publishing their data in a structured manner. One output of the data is as a set of linked web pages which contain links back to the raw data files. But the catalyzer also publishes its catalog as a machine readable XML file, for use by the server.

The task of the server is to construct a distributed database which indexes all the individual catalogs, and presents a convenient interface to search, browse and download data coming from multiple sites. It is similar in concept to the original Napster file sharing index, which contained a directory of which mp3 files were stored where; the index can be centralised even if the data remains distributed (see [5] for a Napster-style index of models. The NeuroSys semistructured database project [6] adopts a similar solution.). But it is more general than Napster; the server has to be able to index all the heterogeneous data formats of all the catalogs so that it can respond to a request like "show me all the electrophysiology experiments conducted on CA1 hippocampal pyramidal cells". It is important to note that the primary copy of the data and metadata remains under the control of the researcher who performed the experiments; the server just maintains indexes and pointers back to the original source, just as Google maintains indexes of webpages.

The server presents a dynamic view of a single or multiple catalogs, and provides all the sort/search facilities expected of a dynamic website such as Amazon's. In effect it acts as a "data browser"; just as a web browser browsers HTML pages, the server can browse XML catalogs of data. But because it is XML and the server has access to the structure of the data, it can combine several, or all known, catalogs together.

6 The vision

The aim is to enable neuroscience data to be as available on the web in structured databases as genomic and proteomic data are now. The increased diversity of neuroscience data makes this a much harder problem than that of databasing the genome, and a problem which is not suited to traditional databasing techniques. It is likely that bioinformatics will reach similar issues as they start to integrate large numbers of proteomic experimental techniques and data, and in this respect will start to look more like neuroinformatics.

One potential issue is that our approach does not attempt to enforce standards about how to describe data, and different labs are likely to develop

different ways of describing similar experiments. Would it not be better to fix a standard structure and force people to conform? We initially attempted standardisation using NeuroML [4], for model descriptions but rapidly discovered (as did Gardner et al [3]) that the "standard data format for neuroscience" is too huge to construct, and it is better to focus on generic tools which can cope with mapping between diverse structures.

7 Conclusions

Convenient, automated access to the large volume of experimental data which exists is crucial to building realistic models. The recent NIH policy requiring data sharing should help provide a stick to encourage this process, but the largest barrier to overcome is still the lack of software tools for experimentalists to mark up and organise their raw data. In future, modeling ought to move towards more automated techniques: model fitting to known experiments on channel kinetics: network connectivity statistics derived automatically from measurements: etc.

Our software (as part of the Axiope project (`www.axiope.com`)) is being developed as one possible solution towards making the data available and searchable - initially to other members of a single lab, then to collaborators, and eventually to form a global database of neuroscience data.

References

[1] R.C. Cannon, F.W. Howell, N.E. Goddard, E. de Schutter, Non-Curated distributed databases for experimental data and models in neuroscience, (*Network: Comput. Neural Syst. 13 No 3 (August 2002) 415-428*)

[2] E.F. Codd, "A relational model for large shared databanks", CACM 13(6) pp 377-387

[3] D. Gardner, "BrainML", `brainml.org`

[4] N. Goddard et al, Towards NeuroML: Model description methods for collaborative modeling in neuroscience", Phil. Trans. Roy. Soc, series B, (Aug 2001), vol 356, issue 1412 1209-1228.

[5] F. Howell. "The Neuroml/Napster project", `www.neuroml.org/napster`

[6] S. Pittendrigh, G. Jacobs "Neurosys", `cns.montana.edu/research/neurosys`

[7] G. Shepherd et al, "Senselab", `senselab.med.yale.edu/senselab`