Dynamic publication model for neurophysiology databases

**Daniel Gardner^{*}, Michael Abato, Kevin H. Knuth, Robert DeBellis
and Steven M. Erde**

Department of Physiology, Weill Medical College of Cornell University, 1300 York Avenue, New York, NY 10021-4805, USA

We have implemented a pair of database projects, one serving cortical electrophysiology and the other invertebrate neurones and recordings. The design for each combines aspects of two proven schemes for information interchange. The journal article metaphor determined the type, scope, organization and quantity of data to comprise each submission. Sequence databases encouraged intuitive tools for data viewing, capture, and direct submission by authors. Neurophysiology required transcending these models with new datatypes. Time-series, histogram and bivariate datatypes, including illustration-like wrappers, were selected by their utility to the community of investigators. As interpretation of neurophysiological recordings depends on context supplied by metadata attributes, searches are via visual interfaces to sets of controlled-vocabulary metadata trees. Neurones, for example, can be specified by metadata describing functional and anatomical characteristics. Permanence is advanced by data model and data formats largely independent of contemporary technology or implementation, including Java and the XML standard. All user tools, including dynamic data viewers that serve as a virtual oscilloscope, are Java-based, free, multiplatform, and distributed by our application servers to any contemporary networked computer. Copyright is retained by submitters; viewer displays are dynamic and do not violate copyright of related journal figures. Panels of neurophysiologists view and test schemas and tools, enhancing community support.

Keywords: networked databases; electrophysiology; neurophysiology; Internet

1. INTRODUCTION

Community data resources require user interoperability: design, implementation and support that facilitate access to and utilization of data. We here report on the specification and implementation of two neurophysiology databases, each designed for interoperability. One component of user interoperability is technical; availability should ideally be multiplatform, require no new client software, and provide open access to data. However, much of scientific computing has evolved in a highly pragmatic fashion, focusing on meeting data needs on a case-by-case basis using the best tools or methods for data exchange then current. Applications often specify unique file formats, generally with significant implementation-specific data, making interoperability difficult.

The Java language and Java Virtual Machine (Gosling & McGilton 1996) aid technical interoperability by providing a uniform environment within contemporary disparate hardware and software environments, facilitating the design of tools for scientific computing that can be used—and perform identically—on IBM-compatible personal computers (PCs), Macintoshes and many UNIX or Linux workstations. Java viewers have been developed for our databases to deliver neurophysiological datasets

and accompanying metadata that place them in a specific experimental context.

Similarly platform-independent is the world wide web, its underlying hypertext transfer protocol (HTTP) and hypertext mark up language (HTML) standards, and browsers such as Netscape Communicator and Microsoft Internet Explorer. These browsers display formatted text, graphics, sounds and movies. To link project-designed Java viewers and neurophysiological datasets to our databases, we have used the extensible mark up language (XML). XML is a multi-platform language with links to both the web and Java that permits semantic and syntactic specification of not only datasets but also models and specifications for data exchange.

Beyond hardware–software compatibility, user interoperability additionally derives from designs conforming to the standards and practices and common core of knowledge of a domain of science. In our design, we have tried to achieve both technical and collegial components of user interoperability by emulating two proven schemes for information interchange: the journal article and the sequence database. We emulate the utility—and interoperability—of journals by focusing on figures of characteristic or defining datasets accompanied by textual metadata as the delivery interface for database users. As in sequence databases, we have designed a data model based on conventions used by investigators in the field to serve as a master repository for author-originated

^{*}Author for correspondence (dan@aplysia.med.cornell.edu).

submissions. We provide universal access to the actual data underlying figures in a uniform format to users with PCs, Macintoshes and several types of UNIX workstations. Additionally, we minimize technical barriers by shielding users from internals, and by enabling authors to submit their data directly.

(a) Sequence databases present an open model

Sequence databases represent a highly successful example of scientific data interoperability. (Baldi & Bruniak 1998; Baxevanis, 2000; Benson *et al.* 1999). This success derives in part from the nature of nucleic acid and protein sequence data, which are represented by a data type—text—readily grasped by people using western alphabets and incorporated in all general purpose computer systems. The datasets themselves are self-describing, with high inherent significance, and with corresponding low dependence upon metadata for description and selection. Sequences are in addition exemplars of re-analysable datasets, whose value is enhanced with subsequent comparisons and meta-analyses. The utility of the information and of methodology for data access, search, comparison and display is high: homologues can be found, genomes compared with cDNA or EST, sequence–structure comparisons can be made, and pattern analysis can reveal data-derived functional correlates such as consensus sequences and putative transmembrane regions.

The utility of sequence databases also derives from designs that maximize user interoperability. Standards and tool development arising from the community have enhanced ease of use and promoted open access: non-exclusionary, technologically accepting of many different hardware and software platforms, and allowing direct (moderated or non-moderated) author submission of data (Altschul *et al.* 1990, 1997; Schuler *et al.* 1996). Some of this utility can be emulated in the design of neuroscience databases, including open access to actual datasets, central repositories, ease-of-use and direct author submission.

(b) Neurophysiology data transcends sequence data

There are, however, major differences between sequence data and the range of data to be represented in neurophysiology databases. Sequences are self-defining. To a first approximation, the sequence is the information—everything else is statistical, inferred, or comparative. Additionally, sequences can be represented simply as text files selected from four-letter or 20-letter sets. Finally, ASCII has provided a vehicle for encoding text files that is universally implemented on all contemporary computers.

Unlike genome data, many neuroscience data types are context- and experiment-dependent. They represent a view of the cell or organism that is closer to that provided by expression data than to the genome. Data from the nervous system often depend in complex ways upon recording technology and experimental protocols. The interpretation of such records usually requires information describing the individual neurone sampled, and pharmacological-, behavioural-, time-, or event-dependent factors, each of which has to be encoded in metadata. In addition, electrophysiology datasets consist of multiple

time-series or other instrument-derived data that are non-linear in several senses and encompass a wide range of experimental techniques and community conventions.

The methodological constraints and diversity of neurophysiological data types therefore impose requirements for access, storage, query and data entry techniques that are very different from those of genomic databases. The nature of time-series and other data, and the subtle distinctions among spike trains or synaptic currents, favour schemas that search the metadata tags rather than the datasets themselves.

(c) Journal article as exemplar for database content and scope

We have utilized the journal metaphor for determining the type, scope and amount of data to comprise entries to our neurophysiology databases. A major design goal has been to make available characteristic or defining datasets that, like a journal figure, inform a community of investigators about a central fact or hypothesis test. Data storage and metadata design are thus different from those needed for alternative models such as an exhaustive digital laboratory notebook or local ‘collaboratory’. As described herein, database query results and views of data are organized following a model derived from journals’ parsing of data into distinct figures. Such databases and the datasets they supply should be viewed not as substitutes for the literature but as adjuncts, with data and literature coexisting, each informing and adding value to the other.

As for journal articles, all database entries are attributed and time-stamped, making submission analogous to authorship. Publications provide a conventional entity for organizing related data, figures and metadata—the database equivalent is the experiment wrapper, a uniform but open framework that organizes without restriction.

Such publication-model databases can transcend as well as complement papers. The data-driven database model minimizes textual hypotheses, tests, interpretations and commentary. Display and delivery of neurophysiology data and controlled vocabulary (CV) hierarchies require new methods beyond those currently used for paper-based or page-oriented Adobe Acrobat (.pdf) articles. Unlike static illustrations in these journal articles, database viewer tools can be designed to have access to the actual datasets. Using this capacity, we have designed such tools to resemble an oscilloscope, with dynamic control over magnification and extent of data views, as well as the developing capability to download actual datasets, permitting reanalysis. Such dynamic viewers allow presentation and exploration of datasets without infringement of journal copyrights on static figures presenting similar data. Submitters, rather than database maintainers, retain copyright to their data. Also unlike other forms of presentation, the metadata defining and distinguishing submissions and datasets can be designed to be searchable, allowing targeted exploration of database contents. The archiving and delivery of data from multiple laboratories in a common format additionally facilitates comparison of related data.

In addition, data entry methods can follow models previously established by journals and sequence databases. Both provide open reading, but restricted and reviewed

submission. A useful design goal is to minimize the extra work required of submitters to enter data so that the total effort is no more than that needed to construct a comparable figure for publication. Review can consist of moderators filtering submissions to determine suitability for posting, as in journals or some databases. An alternative model for submission, non-curved and non-moderated, involves a registration process verifying the identity of submitters and providing minimal credentialling, perhaps through membership in a recognized scientific society. In such an open system, investigators themselves take responsibility for the accuracy of their entries. A conjoint responsibility of the database is to verify submitters' identities, by requiring a login and password, and generating an audit trail for establishing precedence, by appending to the entry the submitter's name, IP address and a date and time-stamp. As in sequence databases, data submission can mandate accompanying metadata sufficient to describe the dataset.

(d) *Data-driven databases*

Journals present not only data and metadata but also related hypotheses, tests and interpretations. In contrast, sequence databases are largely data-driven. Both the nature of neurophysiology data and the complexity of their interpretation favour standards and methods for archiving and exchanging not knowledge but data. Acknowledging Hamming's (1962) dictum—'the purpose of computing is insight, not numbers'—we also recognize that realizable technology will find it easier to provide researchers with numbers and graphical representations of those numbers, from which insight can be derived, rather than to attempt to encode, index and deliver the insight directly. Consequently, we propose standards for archiving, searching, browsing, displaying, and exchanging not knowledge but datasets, with ancillary attention to recording sites, methodology and allied models.

Unlike interpretations drawn from them, data are persistent and model-independent. In neurophysiology the quality or figure-of-merit of many classes of data is largely determined by the recording methodology and protocols, and is relatively observer-independent. We highlight classes of datasets for which archiving and exchange are both feasible and beneficial—such datasets gain value from comparison, *post hoc* analysis and experimenter–modeller exchange. This reanalysability of data further justifies a data-driven schema, but we recognize that the model is more easily justified for some domains of neuroscience, such as neurophysiology, than for others. Our focus on data does not prevent linking with knowledge-based databases such as NeuroScholar (see Burns, this issue), via interfaces that recognize and mediate compatibilities among schemas.

(e) *Goals for data-driven neuroscience databases*

Our two neurophysiology databases thus follow many components of the journal and sequence exemplars, adapted to the nature of neuroscience data and its use. As for sequence data, neuroscience data are accumulated by multiple individual investigators. Unlike sequences, this volume of raw local data produced by individual investigators is great, so the scheme is for characteristic rather than exhaustive datasets. Also, unlike sequences, there

are no uniform standards for data representation in software or even for the scope of data acquired and analysed.

We selected datatypes to encompass a broad range of electrophysiological data. Initially concentrating on data with a high utility for reinterpretability by *post hoc* analysis and collaboration, datatypes are extensible to accommodate the more general requirements of neurophysiologists using a wide range of preparations as well as new methods. Following sequence databases, actual datasets are made available for *post hoc* analysis and collaboration. Unlike text-based self-describing sequences, these datatypes require extensive development of methods for archiving and access.

User interoperability—the ease with which users can submit and can read these datasets—benefits from a data model that should be intuitive to investigators in a targeted area of neuroscience. This data model should be embodied in open tools for database access and data exchange, and multi-platform in scope to work with existing hardware and software in as many individual laboratories as possible, promoting equal access. User client tool download and launch should be automatic, obviating the need to acquire new software to access the database and view datasets and metadata.

Although this technical component of user interoperability is crucial to effective development of standards and tools, interoperability must arise from and be built upon neuroscientists' views of, uses for, and models incorporating data. Technical 'solutions' are no solution at all unless they address perceived needs of the community they are designed to serve, and implement them in a manner compatible with community tools and techniques.

This paper combines a review of design principles with a description of two prototype database projects. One serves the primate somatosensory community with recordings from cortical neurones, largely from awake behaving animals, but is designed to be expandable both to other somatosensory areas, such as rat barrel fields, as well as to broader auditory, visual and motor systems physiology. It may be accessed at <http://cortex.med.cornell.edu>. The other database includes a range of data from invertebrate and other reduced preparations; reflecting a major prototype organism, it is called APLYSIA, a recursive acronym for APLYSIA Proficiently Lets You Scan Identifying Attributes, and is available at <http://mollusc.med.cornell.edu>. These prototypes benefited from parallel and earlier designs for neurophysiology data resources, including the time-series data protocol (see http://soma.npa.uiuc.edu/isnpa/tsdp/tsdp_api.html), the University of Southern California (USC) Brain Project (Arbib *et al.* 1996; Grethe *et al.* 1996) and SenseLab (Shepherd *et al.* 1998).

In addition, the paper suggests design methodologies to enhance interoperability of neuroscience data resources, towards advancing goals of the Human Brain Project initiative (Beeman *et al.* 1997; Dickson *et al.*, this issue; Huerta *et al.* 1993; Koslow & Huerta 1997; Mazziotto *et al.* 1995; Mazziotto, this issue; Shepherd *et al.* 1998). Some findings have been described previously in abstract format (Abato & Gardner 1999; Abato *et al.* 1998; Gardner & Erde 1995, 1997a,b; Gardner *et al.* 1998, 1999, 2000a,b; Knuth *et al.* 1999).

2. METHODS

(a) *Multi-platform technologies*

Hypertext mark up language (HTML), hypertext transfer protocol (HTTP) and Java drive platform-independent client-server data exchange for the cortex and mollusc neurophysiology databases. The Internet global data network and associated multi-platform multimedia data exchange protocols form the world wide web (Gardner 1995). Major web protocols include HTTP (Berners-Lee 1992), HTML (Berners-Lee & Connolly 1994) and XML (see section §2c). HTTP transmits requests from clients and information in response from servers. HTML provides platform-independent encoding of formatted text, graphics and other data. Broad data types are definable by additions to the multimedia internet mail extensions (MIME) types. Although any data type may be transmitted server-to-client using MIME-type extensions, HTML and web browsers generally provide multi-platform support only for formatted text, graphics, sounds and movies.

The development of viewers for complex neurophysiological datatypes requires an additional multiplatform tool—the Java language environment (Gosling & McGilton 1995)—which allows the creation of data query and display programs that run on any client machine, and that are automatically downloaded to clients with a requested matching datatype. We currently standardize on the Java development kit (JDK) 1.2 (Java2) for server processes and JDK 1.1.8 with Swing (Java foundation classes) for clients, reflecting the delayed implementation of Java2 virtual machines on all our target platforms. To ensure compatible Java Virtual Machine environments and largely uniform performance, we establish reference platforms for Microsoft Windows, Macintosh and several variants of UNIX. User instructions are provided for obtaining the Swing classes and (for some platforms) the plug-in architecture that enables JDK 1.1.8 (and eventual Java2) support. HTML-based code uses web browsers running client-side JavaScript 1.1 or later; and a JDK 1.1.8 or later Java Virtual Machine.

(b) *Programming standards aid generality*

Current technologies are used to implement this architecture; note that except for the use of Java and XML, alternative methods exist for each architecture element. Informix Online Server with Universal Data Option (version 9.14) is installed on <http://cortex.med.cornell.edu> and <http://mollusc.med.cornell.edu>, each a Sun SPARC system. Core database code is written in SQL92 with SQL3-like extensions as implemented by the Universal Data Option, which in turn is derived from the ILLUSTRATE object database.

For the mid-layer, we designed a QueryServer, providing interactions with the database for query, retrieval and description of data for insertion, a DataServer, providing interactions concerning stored datasets, their presentation, and their upload, as well as a public, external protocol, and a Java Application Server supplying client applications. QueryServer uses Netscape Enterprise Server 3.6, using server side JavaScript's database access API. The DataServer Java application runs JDK 1.1, the collections API from JDK 1.2, the Informix JDBC driver and standard XML parsers.

The latest versions of the client applications DataViewer and QueryTool have been coded in Java. QueryServer, DataServer and clients communicate via XML. All project-written code is designed toward the Open Source model, using the Lesser General Public License of the Free Software Foundation, supple-

mented to acknowledge funding sources and to allow free use in exchange for specific recognition that intellectual property rights for datasets belong to the experimenter-submitters, not to databases or publishers.

(c) *XML standard for structured data*

XML is employed as a data description language and as a method for specifying standards for data and data model descriptions, along with the Resource Description Format (see <http://www.w3.org/Metadata/RDF>). XML (see <http://www.w3.org/XML/>) is a standard adopted by several scientific communities, spanning chemistry to mathematics (for a review and links, see Sikorski & Peters 1998) to subfields of biology (see http://www.proteometrics.com/BIOML/bioml_toc.html) and medicine (Friedman *et al.* 1999). As a data description language, XML has spawned multiple tools, standards and proposals that both use and enhance the language. These include XML Schema, Namespaces, Document Content Description, Data, Linking Language, Pointer Language, MIX Mediator System and multiple style and Java schemes. Additionally, XML is platform- and technology-independent, requiring only the ISO Latin alphabet, so XML descriptions and datasets need not be made obsolete by advances in hardware or software technology, supporting permanence of data resources.

3. A COMMON DATA MODEL FOR NEUROPHYSIOLOGICAL DATA EXCHANGE

We began with the design of a data model as a general informatics framework for neurophysiology data that was: (1) compatible with an intuitive view of data; (2) implementable using contemporary database and data exchange methodology; (3) inclusive of heterogeneous neurophysiological numerical datasets and the complex metadata structure that distinguishes neuroscience data from sequence and structural biology datatypes; and (4) focused on data useful for exchange, collaboration, comparison and *post hoc* analyses.

The common data model is centred on neurophysiological and biophysical experimental data and associated representations, including time-series and other (1D) datasets. Time-series datatypes include a variety of data records representing amplitudes of physical quantities sampled directly by analogue-to-digital (A/D) converters at a uniform rate. These include electrophysiological recordings with intracellular or extracellular microelectrodes, patch-clamp electrodes, microensors of pH, ions or molecules, force or position transducers, or activity monitors. The datasets delivered can encompass spike trains, PSPs, synaptic currents, ion flows through single channels, changes in neurotransmitter concentration or ionic composition of extracellular microenvironment or intracellular compartments. Time-series data can also include behavioural variables such as muscle force, eye position, or circadian activity, or monitors of stimulation parameters such as pressure or temperature—any physical variable sampled at a uniform rate.

One-dimensional datasets include numerical time-stamps of event markers of uniform amplitude. Examples include spike firing times, occurrences of possibly unitary events, such as miniature synaptic potentials, or the onset or end of stimulation or motor activity. Such data are usually represented as series of pulses or dot rasters. In

addition to these event lists or spike time-stamps, other 1D datasets include sets of values to be binned for histograms, as well as post-binning counts or fractions. These datasets include perievent time-histograms (PSTHs), auto or cross-correlograms, interspike interval histograms (ISIHS), channel open-time histograms, etc.

Datatypes were selected based on their utility for *post hoc* reanalysis (see section § 1d). Time-series and histogram data can clearly benefit from methods for data exchange that expand the number of investigators analysing the data or the number of methods brought to bear. Data with bivariate representation, such as current–voltage curves characterizing membrane conductances, are less likely to benefit from these exchange methods than either time-series or histogram data. However, the larger the bivariate dataset, and the more different forms of analysis that can be applied, the greater is the advantage that can be gained by data exchange. Bivariate data are additionally useful for summarizing parametric relations, and are therefore included in the common data model. Not yet included is another common class of data: sets of single values varying with experimental manipulations, often represented as bar graphs in clusters with each cluster representing a different experimental condition. These data are often difficult or unrewarding to analyse *post hoc*, both from the format and from the usually small number of data points.

To parse these datatypes and enhance data sharing among submitters and requestors, we designed a neurophysiologically-intuitive hierarchical model for experimental data and metadata. The wrapper encompassing linked data is the ‘experiment’. All data are submitted as part of an experiment and database searches similarly return experiments. Experiments contain one or more ‘views’ of data. A view is a single figure panel with metadata, including a descriptive label. Each view type corresponds to one of three major classes of data: time-series, histogram or bivariate. Video data are included in the data model but not yet implemented. Views present sets of related ‘traces’: numerical datasets—time-series sweep or x – y curve. Figure 1 shows both a schematic view of this hierarchical model for data, as well as a display of labels returned by the database from a query, showing how user tools implement the experiment, view and trace schema.

The common data model thus permits the provision of multiple views forming a descriptive experiment. For example, different views can compare stimulus-evoked and spontaneous activity, or responses of the same neurone evoked by different stimuli. Each view can include multiple traces, such as spike trains, sampled records of a time-varying stimulus, and event flags annotating specific records at different time points. The experiment→view→trace schema provides a visual syntax for neurophysiology data, complemented by CV metadata attributes (described below in § 5) that together form a lexical syntax and semantics for neurophysiology data.

The underlying schema, supporting multiple electrophysiological techniques and preparations, is based on five superclasses: data, site, method, model and reference elements (Gardner *et al.* 2001). Data elements include datasets and quasi-data, such as simulation results, also allowing coordination with model-based studies that are compatible with our basic scheme. Figure 2 shows the

Figure 1. The common data model implements an intuitive experiment→view→trace hierarchical wrapper model for datasets, including time-series, histogram and bivariate datatypes spanning many contemporary techniques.

(a) Presents this organization schematically, indicating recording technique and datatypes encompassed and the presence of metadata at each level of the hierarchy. (b) Data lists returned by a query follow the experiment→view→trace schema, as shown by an early QueryTool running under Sun Solaris. Three experiments were returned by the query, and the experiment entitled ‘Aplysia muscle K currents...’ is expanded to show titles of two contained views and several traces.

data element design, including the experiment, view and trace hierarchy, and the specific classes of views and traces accommodating different groups of recording methodologies or datatypes. Also presented in figure 2 are links from data element to the other four top-level superclasses. Site elements span systems to compartments. Method elements include protocols, transforms and modelling engines. Model elements include

Figure 2. Data element representation of datatypes and data wrappers defined for microelectrode neurophysiology, and shared by our invertebrate and cortical databases. Other domains of neuroscience can similarly define distinct datatypes and more granular or encompassing wrappers. This scheme, shown as an entity-relationship diagram (Chen 1976), presents definitions of database relational tables implementing experiment, view and trace. Each is shown with names and datatypes of characteristic data and metadata attributes. Experiment, view and trace are subclasses of the superclass data element, with inheritance shown diagrammatically using IDEF1X underlined-circle notation as specified by the National Institute for Standards and Technology (1993). More specific subclasses extend the inheritance relations; asterisks indicate tables defined without specific additional attributes. Database foreign key (_FK) relations assign traces to views and experiments. Diamonds indicate relations between data elements and site, reference, model and method elements, linking datasets with sites such as neurones. Entity-relationship and entity-entity stoichiometry is specified on each connector line: a bar indicates exactly one and the circle and crow's-foot specify none-to-many.

hypotheses and parameter sets. Reference elements encode bibliographies.

4. IMPLEMENTATION OF NEUROPHYSIOLOGY DATABASES ENCAPSULATING THE COMMON DATA MODEL

(a) Goals and overall design

Neurophysiologists searching for data and viewing results need not be concerned with implementation or internal details, as our schema encapsulates technical detail. Neurophysiologists need only know the appropriate Internet address (<http://cortex.med.cornell.edu> or <http://mollusc.med.cornell.edu>) and what data we provide. Data submitters need only accompany datasets with metadata sufficient to characterize the experiments.

Database and user tools are designed to incorporate implicitly the common data model. Client tools, reflecting goals for platform-independence described above, were additionally designed to be supplied by the databases project upon demand. Following this schema, no versioning is needed for client tools, which are always up-to-date. A goal that is not yet fully realizable, awaiting standardization on uniform Java-enabled Web browsers, is to require no action on the part of users to obtain any software component.

Recognizing the need for credit assignment, any data can be annotated with appropriate citations, including authors' names and publication data. For journals indexed by MEDLINE, this is the publication's unique UID; for others, appropriate reference formats are available.

These goals are efficiently met with a standard three-tier client-server model, shown schematically in figure 3. It includes:

1. a unified server-side data repository implemented by an object-relational database engine for storage, relation and querying of numerical datasets and metadata;
2. a midlayer interface linking client and database while shielding users from internals, with query parsing, bi-directional data selection, translation and delivery, and system services and management of user accounts; and
3. client applications providing database access via graphical query assembly, transient data caching, display of data and metadata, user registration and data upload.

Decisions were made during the design phase of the proposed project to ensure that the hardware and software technology selected would be compatible with the very different needs of small and focused *versus* broad neuroscience communities, as well as with datasets of very different size and number. Ours is an open model in several senses, with not only extensible datatypes and methods, but also free distribution of datatype specifications, database access, and query and viewer tools.

In contrast to the user interoperability promoted by these databases and their underlying common data model, interoperability among developing neuroscience data resources requires, in addition, universal standards for data exchange that compatibly describe any resource's data model, scope, search criteria, formats and methods, aiding database–database exchange of data, data models and queries. Toward this end, we have begun to abstract and extend the common data model to provide a framework for database federation (Gardner *et al.* 2001). Complementing this schema, data exchange is aided by interfaces using biophysical description mark up language (BDML), an XML-based data description language for describing data and data models. Such interfaces do not require existing or independent data resources to redesign their data models or methods. Instead, they are designed to offer mappings through external interfaces that identify commonalities and so facilitate interoperability.

(b) Database design

Neurophysiology data include not only numerical datasets of many types but also extensive textual and other metadata. To encompass this diversity and embody our data model, the project has implemented the cortical database using an object-relational database engine, which combines features of traditional relational databases with object-oriented extensions to provide for new datatypes. Relational databases optimize data handling by robust tools and use structured query language (SQL; Astrahan & Chamberlin 1975), which permits joining disparate regions of data space, as a standard interface. Object extensions transcend the relational model not only by adding unlimited datatypes and output formats but also by aiding design via class inheritance and function polymorphism. Because the object-relational database engine allows inheritance of datatypes, specific views and traces are flexibly and extensibly defined as subtypes of more general ones (figure 2). A generic view includes identifiers and the integer number of traces displayed; **Time_series_view** inherits all of these and adds time axis scaling. A single **time_series_view** can include multiple traces from any or all subtypes of **time_trace**.

Major technical requirements of the database include the storage of complex objects, including annotated physiological data and images as well as text, and the ability to search parsable text tags within or associated with these complex objects. We have developed a unified dataset storage scheme for our databases, featuring universal binary large objects (BLOBs) for datasets, and an extensive set of relational tables for textual, numerical and logical metadata. The BLOB/table dichotomy, shown in figure 2, is transparent to users and we maintain referential integrity using both the database engine and additional internal references.

The multidimensional generalized BLOB design can store many different numerical and non-numerical data of any dimensionality. The design encompasses such diverse datatypes as time-series (continuously sampled, spike times and event lists), bivariate and histogram data. To include all data relevant to a dataset in BLOB format would require parsable BLOBs, necessitating extensions to standard database schemas and reducing performance. Although our BLOB format makes provision for limited annotation, most relevant metadata, including extensive metadata describing datasets, neurones, protocols and references, are stored instead in relational tables, using CV search terms common to submitters and requestors.

An example of datatype design (not fully implemented), including integration of data and metadata, is provided by the histogram class. Like time-series, histogram data encompasses multiple subclasses and techniques, including single channel amplitude and duration data as well as several subtypes of single-unit spike timing histograms (figure 4). Datasets of five distinct types are comprised by the schema, each a trace in the experiment→view→trace scheme: unbinned raw data points, or four pre-binned representations, including counts, probabilities or percentages, event counts and event rates. Each of the pre-binned types can also include high and low outside-range values. Datasets are represented in BLOB form in different ways depending upon type, including vectors of FLOATs containing individual data points that are to be binned and counted, vectors of FLOATs containing bin percentages or rates, or vectors of INTs containing bin counts. Trace metadata can include bin start and width, number of bins or last bin, and number of trials summed. Each histogram view can include additional metadata including horizontal axis information.

We emphasize that although the object-relational scheme eases implementation, both the specific data model and our methods for database–database interoperability are architecture-independent, with alternate implementations possible.

(c) Mid-layer

The use of a middle tier (figure 3) provides public methods for database access that are independent of our data storage schema, also allowing the migration or substitution of database architecture or technology noted above, without affecting the user interface. Two components of the middle tier, QueryServer and DataServer, implement goals of platform independence and insulation of users from database internals. DataServer–database communication uses JDBC interfaces to the database. The Java QueryServer is responsible for mediating most

Figure 3. Multi-tier data exchange architecture for neurophysiologist users and interoperability with other neuroscience data resources. (a) Shows multi-platform Java-based tools for neurophysiologist users, including QueryTool and Virtual Oscilloscope. (b) The DataServer mid-layer processes enable CV user searches and deliver client applets as well as XML-wrapped data and metadata. (c) The object-relational database receives queries from the mid-layer and returns data and metadata using the JDBC interface.

communication between project-designed client tools and the database. The QueryServer uses XML to communicate between the client tools and the mid-layer, both for query processing and the consequent DataServer data and metadata return, with the granularity of the response specified in the query.

Because the model includes supplying not only views but also datasets for analysis by the client, DataServer can also filter various input and output formats used by data acquisition and analysis programs—and converts them to and from our internal BLOBs. This scheme potentially allows non-text numerical datasets to be delivered by the DataServer as external entities in any of several formats, each properly referenced by the XML response.

All client–DataServer communication is by web standard protocols: HTML, XML, HTTP, MIME-types, Java and TCP/IP. These standards provide seamless download and display of any supported datatype to any client, as well as dynamically updated server-side client applets and screens, search parameters and Java methods.

(d) *Client tools for access, query and display*

To maximize ease of use, and therefore the likelihood that the community will use this resource, access and

viewer software tools are designed to be as intuitive as possible. Access to and search of the database, and display of returned data, utilize project-designed DataView, Virtual Oscilloscope and metadata viewer programs, designed to incorporate our data model and provide informed queries and dynamic views of data. The database and query methods are designed for investigator-directed retrieval of information by browsing or directed queries of textual metadata. The data model may permit future use of feature extraction algorithms that analyse datasets rather than metadata, but the present generation client tools do not provide this capability.

We have rejected solutions based on single platforms or specific methodology and adopt instead an open universal client model utilizing Java and application server technology to enhance web-based protocols. Current versions of client tools are coded entirely in Java to advance interoperability by providing compatibility with any of several computer platforms. Viewers as well as data are automatically downloaded to display experimental data selected from the search results. This application server model additionally ensures that the client QueryTool, Virtual Oscilloscope, and metadata viewer methods are current, with the latest versions of each supplied as needed automatically and seamlessly by the mid-layer. As a result, any neurophysiologist can access metadata attributes,

type	dataset	raw/binned	format	viewer	
				re-bin?	display options
1	1D data values	raw	float	yes	counts, %, p
2	counts	pre-binned	integer	no	counts, %, p
3	percents	pre-binned	integer	no	% or p
4	event counts	pre-binned	integer	no	counts or rates
5	event rates	pre-binned	float	no	event rates

uniform *trace* metadata:

Bin_start, Bin_width, Number_of_bins or Bin_end

(additional metadata for type 4: Number_of_trials_summed)

uniform *view* metadata:

H_axis, H_units

Figure 4. The histogram datatype includes both raw 1D data values, allowing dynamic binning by future versions of the Virtual Oscilloscope, as well as four classes of pre-binned sets of values, representing counts or percentages (or probabilities) for a single dataset, or event counts or rates for multi-trial runs that include nulls in the statistics. For these cases, the data model includes in addition high-field and low-field values. As the data model is largely independent of viewer issues, each datatype potentially permits multiple viewer options. Metadata are found at view and trace levels.

form a query from them, access the full database, and view the dataset using the latest versions of client tools updated as datatypes evolve or parameters change.

Figure 5 presents the Java QueryTool, which constructs queries for metadata attributes. It allows selection from server-maintained controlled-vocabulary terms using a menu-driven point-and-click interface that can traverse any level of the hierarchical tree. Results of such searches are graphically presented according to the experiment→view→trace model (figure 1).

Figures 6 and 7 show the multi-platform Java-based Virtual Oscilloscope, displaying time-series and histogram data on a variety of contemporary hardware and software platforms. Experiments returned by the Data-Server to the viewer are collections of views, each displaying a rectangular figure pane containing one or more coordinated dataset traces, following the data model.

The display and delivery methods are designed to present an interactive, dynamic view of neurophysiological data rather than a static picture, giving the user several options for interactive control of scaling, styles and metadata display by the user. To do this, we download actual datasets to the viewer and construct dynamically created views to display the dataset. We thus transcend existing methods for multi-platform delivery of static images, video or animations. As figure 6 shows, one advantage of this model is that users can alter display sweep and other parameters of the time-series or histogram data displayed, transcending the information available from a static journal figure displaying the same dataset. Tools are being tested that give users the option to save datasets to their local file systems as well, aiding meta-analyses and reanalysis. This dynamic display also avoids mere duplication of pre-existing figures, avoiding copyright infringement of related static figures in the literature.

Figure 7 shows that the generality of the Virtual Oscilloscope encompasses processing and displaying datasets found in invertebrate and reduced preparations, including sampled patch-clamp and voltage-clamp-time-series and related histogram data as well as the time-stamped records for spike trains.

5. CONTROLLED-VOCABULARY METADATA

(a) *Design goals for neurophysiology metadata*

A database of data records alone is ineffective for a multi-user community unless each dataset is accompanied by appropriate metadata that identify it and distinguish it from others. Interpretation of physiological records depends more heavily than does genomic data upon these complementary metadata, which should include how, from where and by whom, the data were recorded, and also conform to community usage.

As noted above, it is difficult to search time-series and other datasets themselves, given the subtle distinctions among spike train dynamics or after-potential shapes. Searching parsable, descriptive text, metadata requires that text values have the breadth to span the domain while providing adequate selectivity. In a discipline as diverse as neuroscience, multiple optional categories of metadata provide flexibility, as does the ability to extend metadata without obsolescence. Design should permit efficient implementation to speed queries and methods should transcend specific architectures or technology. For future federation, schemes should interoperate and be generalizable to other domains of neuroscience and to non-neuronal biophysical data.

(b) *Controlled-vocabulary schema*

To satisfy these design goals, we implement neuroscience metadata using a CV semantic and syntactic environment. To ensure uniformity when entering data

Figure 5. QueryTool constructs searches by selecting CV descriptive terms. The Java QueryTool provides intuitive database access via searches of metadata describing datasets, neurones, protocols and references. The screen shown presents cortical neurone receptive field and motor behaviour descriptors. Values of most attributes are organized in hierarchies of CV terms; terms can be selected at any level of the hierarchy. Details of our hierarchical attribute-value implementation enabling selectable specificity searches are presented in Gardner *et al.* (2001). Performance is identical for any user with a Java (JDK 1.1.8 or later) browser, here shown on Macintosh using the Macintosh Runtime for Java Virtual Machine.

descriptions and searching databases, the same vocabulary and environment are shared by submitters and requestors. We have tried to identify a set of near-orthogonal hierarchical descriptors that span each of our domains while providing specificity for searches among similar data. We have chosen to model a reduced set of metadata terms that represent classes of datatypes, techni-

ques and protocols, and neurones and other recording sites. Because both attributes and the CV values they can have are stored in the database and maintained on the server, search criteria are always up-to-date, and available via forms and menus to generate user requests. This design is moreover easily generalizable to other preparations and fields of neuroscience.

Figure 6. Histogram and time-series Virtual Oscilloscope is platform-independent, interactive and dynamic. Somatosensory cortical peri-stimulus time-histograms and multi-trace neurone activity shown via spike replicas; data are from the laboratory of E. P. Gardner. Experimental data are graphically parsed as view panes containing one or more traces. Data and presentation are identical in each of the four panels, with sweep adjusted individually for each view to emphasize user control of the dynamic display. Each panel shows an earlier version 2 DataViewer on (a) Macintosh, (b) SGI, (c) Sun Solaris and (d) Windows NT. Both data and Java viewer applets are downloaded automatically via our mid-layer application server on all platforms.

We have selected small sets of metadata with simple vocabularies rather than utilizing established complex medical or anatomical schemes such as the unified medical language system (UMLS; Lindberg *et al.* 1993). CV values were inspired by the Society for Neuroscience keyword lists constructed by Bernice Grafstein (personal communication) and were developed with the assistance of prototype user communities. To aid ease of use by investigators rather than expert curators or searchers, these values are designed to be intuitive and as self-defining as possible (Cimino 1998). Although CV favours rather than establishes uniform descriptions, this simplicity aids uniformity and reduces the ambiguity and opportunities for error inherent in complex schemes. Recognizing the rapid advances and fluid terminology

that characterize contemporary neuroscience, such a vocabulary needs to be expandable with minimal obsolescence. To achieve this, we have designed a hierarchical attribute-value schema that permits adding terms of increased specificity while retaining compatibility with existing entries (Gardner *et al.* 2001).

Our CV scheme also includes a lexical grammar, forming a data definition language for neurophysiology. Metadata describe not only datasets but also neurones and other recording sites, as well as methods such as protocols and preparations. To serve the resulting wide range of vocabulary terms, the scheme includes specific sets of attributes in the definition of each datatype and specifies a particular vocabulary set for each attribute. Moreover, most sets of vocabulary terms are arranged in

Figure 7. The data model and Virtual Oscilloscope encompass multiple datatypes. A version 2 Virtual Oscilloscope shows intracellular and patch-clamp data, including (a) single Na channel trace and (b) amplitude histogram from the laboratory of O. S. Andersen as well as (c) Aplysia muscle delayed rectifier K currents, courtesy of K. R. Weiss; the displayed data are those shown in the query return of figure 1.

hierarchies, allowing variable specificity. With this design, syntax is extended but not restricted, allowing simplified semantics.

Unlike free text metadata, individual CV terms lack a natural-language textual context. Our schema is designed to add context in several ways, thus increasing the utility and precision of the CV. By structuring the vocabulary with attribute=value pairs, the attribute serves to specify additional context for CV values. If such attributes are associated with specific database elements, and if the database elements are additionally hierarchical, context is enhanced still further. Finally, hierarchical vocabulary terms gain further structural context from the resulting vocabulary tree (Gardner *et al.* 2001).

Following this schema, we have specified multiple metadata attributes in the definitions of many database entities. CV values include parsable descriptive text tags chosen with the assistance of two prototype communities of consultant collaborators. Reflecting both the breadth of neuroscience and the diversity of community approaches to description, many attributes are optional.

Such metadata attributes are displayed to characterize datasets returned by the database. Like the Virtual Oscilloscope, the metadata viewer (figure 8) implements the experiment→view→trace hierarchy for display of dataset-related metadata and presents as well additional metadata specifying recording site, protocols and other experiment-related metadata elements. Recording methodologies (e.g. whole cell patch-clamp) and specific neurophysiological data descriptions (e.g. isolated PSCs) are linked to each data view or trace, as applicable. Currently, XML-based metadata are converted to HTML by a parser for easy viewing and printing using a web browser; as browsers become more XML-aware, the DataServer will offer metadata formatting in either the native XML or HTML.

Both attributes and their CV values are part of the common data model, stored in each database, and served dynamically as XML or Java trees. As a result, search criteria are always up-to-date, and available via forms and menus to generate user requests. Investigator-entered experiment, view and trace labels are stored as free text to avoid limitations on these fields. Currently, axis units are also free text.

Protocol and preparation use an experiment-wide CV that is again common to submitter and requestor. As shown in figure 9, protocols and their constituents are subclasses of the *superclass method_element*, implemented as sets of CV pairs or triples. For stimuli, these triples describe the stimulus energy, the temporal pattern of the stimulus and the target—neural, sensory surface or other—to which the stimulus is directed.

Trace-, view- or experiment-linked CV specifies the semantics of neuronal description. For somatosensory neurones, experimentally-determined functional neuronal response profiles are often the major distinguishing characteristic. Receptive fields are specified using CV dual hierarchies encoding location and modality. Recognizing that receptive fields are often broad, multimodal and stimulus-dependent, each neurone may have more than one set of these dual attributes. For neuronal location, we incorporate any of several optional attributes: cytoarchitectonic or functional area, stereotaxic or other

Figure 8. Metadata viewer shows attributes describing experimental data, neurones and references. The same CV metadata used to search the database also characterize returned datasets and related neurones and protocols. Experimental metadata are presented using the same intuitive experiment→view→trace model used for data organization and Virtual Oscilloscope display. The MEDLINE UID number in the References metadata area is a live link, bringing up the PubMed database.

coordinates, depth and cell class. The representation of anatomy is based on electrophysiological and functional criteria, and does not substitute for more detailed Human Brain Project structural parcellation and mapping schemes (Brinkley *et al.* 1997; Dickson *et al.*, this issue; Mazziotta *et al.*, this issue; Mazziotta *et al.* 1995, 1997; Mazziotta, this issue). Multiple parcellation schemes, including Brodmann areas and functional descriptors, are included in the data model, each with its own evolving set of CV values, but no one scheme is mandatory for either entries or queries. The scheme thus follows electrophysiological practice in encouraging both descriptors but requiring neither. Areal assignments are often subjective, reflecting the difficulty of parcellation and mapping between schemes (Stephan *et al.* 2000). The basic scheme is extensible to additional classes of mammalian neurones with different descriptive attributes.

The invertebrate database similarly specifies a very different class of neurones, many of which are identified individuals with a distinct set of specific anatomical and functional metadata descriptors.

(c) Operational specificity

Optional attributes are consistent with our search model, which implements operational specificity sufficient for categorization, or selection of sets of related datasets, rather than absolute specificity excluding all but a unique member of the set. The goal is to implement ease of use for simple queries while allowing more complex queries when required. With a large number of attributes, selectivity may emerge from the use of multiple broad descriptors, a strategy resembling natural-language selection from free text abstracts or summaries.

In addition to its influence on the selection of simplified vocabulary, this operational specificity model also influenced design of the QueryTool. Rather than implementing searches for all metadata attributes, the Query Tool searches only a subset of attributes that appear to have the greatest capacity to disambiguate queries. The schema permits highly specific as well as broad searches. Figure 10 presents a query, represented as XML, produced by the QueryTool using a series of high-selectivity search terms.

Figure 9. Method element defines protocol descriptors. Protocols group nudges, each a two- or three-term CV descriptor specifying an experimental manipulation. For stimuli, this is an effector + pattern + target triple, describing a stimulus type, its temporal pattern and its spatial or neuronal localization. Schemes for behavioural, pharmacological and neurochemical descriptors are being refined to facilitate coverage of broad areas of experimental neuroscience. Entity-relationship schema as in figure 2.

6. DISCUSSION

A common data model underlies two neurophysiology database projects providing storage and exchange of time-series, histogram and other datasets, and searchable CV metadata describing preparations, protocols and neurones.

The data model organizes such datasets into intuitive groups. It downloads Java tools that are broadly compatible with most hardware and software, enabling user interoperability: access by investigators familiar with the domain of neuroscience are aware of the range of data provided by the databases. The data model enables, and the database and tools implement, search and display of datasets to aid collaboration in near real-time among experimentalists and theoreticians.

Much impetus to neuroscience information exchange was provided by two documents summarizing recommendations of a variety of investigators and panels: an Institute of Medicine (1991) report on brain mapping, and a pair of National Science Foundation (NSF) workshops recommending the establishment of identified neurone databases (Comer & Jacobs 1994). Our data model and database designs originally derived from invertebrate community requests and were partially incorporated into the NSF report. Subsequently, technological developments provided enhancements to the data model, and the generality of the evolving schema suggested its utility to mammalian neurophysiology as well.

Since our scheme thus did not derive directly from Institute of Medicine recommendations, and since our focus is on electrophysiology at neuronal levels, rather than on brain mapping, it is of interest to compare our database scope and implementation with the earlier recommendations. Our reliance on the ease-of-use of sequence databases is presaged by the report, which also noted some of the differences between sequence and neuroscience data complexity. We have designed for the less canonical nature of neuronal signals and consequent enhanced requirement for metadata, which were not specified by the Institute of Medicine. The report noted that databases could have more kinds of data than journals, including 'data that could be included in a digital format' (p. 93), which can be interpreted as including the availability of actual datasets. It also noted the ability of digital information resources to provide more complete data and masses of informal data, perhaps underestimating the volume of data that it is now possible to generate within a single laboratory in a very short time. The Institute of Medicine (1991) report noted the need for multi-platform support and for standard formats and protocols, which we support.

Distinctions were made by some panels between a formal consensus database and informal collaborative databases, but this distinction need not require different datatypes. For example, the invertebrate database accommodates equally well recordings from unidentified

Figure 10. QueryTool transmits CV search attribute values via XML wrappers. The extensible markup language (XML) is used for interprocess communication from client to mid-layer. Here, queries generated using the Java QueryTool to search CV metadata trees (figure 5) are transmitted via XML wrappers. Values of terms and hierarchies of terms are shown underlined.

neurones and canonical or exemplar datasets characteristic of identified cells. Just as for polymorphisms and 'canonical' genomic sequences, similar datatypes can serve both, and simple flags can distinguish consensus from representative data.

Our model for use of specific datatypes is similar to that of FAST (Clucas & Watson 1994), in which actual datasets are transmitted to clients and scientific data viewers spawned. A similar but not identical view from the pre-Java era is that of Chhabra *et al.* (1994), who proposed a more general object-oriented model in which the display methods themselves, or pointers to them, are encapsulated into data objects.

(a) *Data flow reflects the data-driven publication model*

In selecting, examining, and analysing both local data and the literature, neurobiologists use internalized models that reflect experience and motivation. Database utility is enhanced if the underlying data model and user tools reflect these models. Figure 11 presents the data flow aiding neurophysiologist submitters and viewers of data for our databases. These also resemble the sequence database model, modified to conform to neurophysiology data.

Assembling figures for publications normally requires selection and organization of datasets and providing text descriptions that characterize the datasets. Figure 11a shows the equivalent database processes. Initiating a submission session transfers DataEntry applets and CV metadata lists to the author's or submitter's local computer. A mini-Virtual Oscilloscope displays each dataset trace as it is entered. Submitters organize their datasets as for figures, annotate by selecting descriptive metadata from hierarchical CV lists, and upload data and metadata to a compliant neurodatabase via XML, Java and a Data-Server process that adds date and author tags.

An ideal neurodatabase would be as usable as a paper journal. Figure 11b shows how our tools are designed to promote similar usability to our databases. Here, a reader initiates a database search, receiving Java applets including the QueryTool and current CV metadata lists. The QueryTool enables neurodatabase search by multiple CV descriptive attributes. Readers familiar with a field and its practitioners are likely to search by submitter's name, as shown in figure 11b; a search using experimental metadata is shown in figure 5. In response, the server delivers the QueryResults applet displaying a set of experiments from the database matching the search criteria. Readers are offered not only a choice of datasets but also actions including viewing data using the Virtual Oscilloscope, examining metadata via the Metadata Viewer, or saving datasets in XML or other formats for further analysis.

(b) *A database, not a lab notebook*

Although this scheme could be used, for example, to replace the carbon-paper duplication of a laboratory notebook by a central data archive server, the data model is designed for publication-quality data rather than as an electronic notebook. Moreover, the broad design of datatypes and the specifications of user tools reflect the disparate requirements and complementary techniques of contemporary neurophysiology, rather than the focused needs of individual laboratories. Such 'local metadata' as trial or series number or acquisition date are not part of the data model, although an acquisition token can be used for establishing links to audit trails held locally and individually by each submitter.

(c) *Analyses enabled by collections of datasets*

Any scheme for collecting and archiving data from multiple sources aids meta-analyses by supplying numbers of observations larger than a single laboratory could

Figure 11. Data flow from a user's perspective, implementing the dynamic publication model for data-driven databases. (a) Data flow for data entry, the process enabling neurobiologist submitters to upload datasets and specify descriptive metadata. (b) Data flow for readers to search the database and acquire selected data and metadata. All illustrations except for those of the remote data server and neurodatabase show screens displayed on submitters' or readers' local computers by database user tools.

reasonably collect on its own. However, the common data model and database implementation can potentially expand not just the magnitude of available data but the number and type of feasible analyses, for both mammalian and invertebrate communities.

For cortical electrophysiology, archiving data from multiple laboratories in a common format allows comparisons of datasets recorded in different cortical areas, or using differing recording techniques, or different stimulus paradigms. As for sequence databases, many of these analyses are enabled because investigators other than the submitter have access to the actual datasets.

Storage and presentation of datasets along with metadata describing neuronal recording sites has the capacity to expand electroanatomy: building maps by physiological responses, including analysis of unit properties by cortical location. Since the metadata provide functional, anatomical and protocol descriptors, common features of spike trains recorded under differing values of these parameters can be correlated with either circuits or regions they encode.

With enough datasets, and functional or subsequent histological determination of cell type, it is possible that characteristic features of records could be used to identify specific cell classes electrophysiologically. At least some cell classes have sufficient specifying qualities to make discrimination possible—thin-spike stellate versus thick-spike pyramidal cells, for example—and it may be possible to extend this schema by collecting and providing additional data. Such properties as characteristics of tonic rest discharge might demonstrate correlation with particular cell classes or cortical sites regardless of other features such as modality.

Common formats for archiving and presentation of neuronal signals can also facilitate analyses of the information content of spike train and allied records. Comparison of spontaneous and evoked firing patterns may identify particular stimulus-encoding features of spike trains. Grouping of records obtained with varying attentional states, or with and without anaesthesia, or using different stimuli, will allow later analysis of informational content and extraction of common *versus* specific features.

With a large amount of neurobiological data available in common format, modellers or other computational neurobiologists will be able to evolve and test theories of information coding and processing. The melding of real data with models may also produce an increased understanding of the differences between informational aspects of neural coding and general properties of spike trains that derive from neuronal biophysics but do not encode specific messages.

Availability of datasets may aid the conduct of research independent of these cross-technique synergistic explorations of neurobiological processes. Disseminating, exchanging and sharing of neurophysiology data may also reduce unwitting duplication of work done elsewhere, freeing investigators to build upon rather than reinvent methods and to supplement rather than repeat findings. The resulting more effective utilization of data should advance more targeted use of mammalian species; this may in turn promote more selective animal use.

In many invertebrate preparations, neuronal recording sites are unique individuals, often identified in part on the

basis of characteristic shapes of an action potential, an I-V curve, or branching pattern. A networked database of characteristic recordings will permit investigators recording from a cell of interest to scan the database for relevant electrophysiological records. A cell may thus be revealed to be a previously identified neurone, or to possess a similar signature to a class of cells in another ganglion. Perhaps equally important, recordings from identifiable cells may be provided and related to data from additional techniques, proving useful in distinguishing an identifiable neurone from other, superficially similar, cells, thus leading to collaborative neuronal identification.

(d) *Data exchange methods may clarify intellectual property issues of scientific publishing*

The evolution of mechanisms for data exchange within neuroscience communities parallels an evolution of media used for scientific publishing. Online journals may benefit from some of the mechanisms we are developing for active figures, and we encourage such use on a non-exclusive basis.

Given this dual evolution, it is important that distinctions are drawn between underlying data and the form of representation published and copyrighted. Again, protein or nucleic acid sequences may provide a useful precedent for the use of neuroscience datasets. The report of a finding including the text and figures that support it should properly be copyrighted by the authors, assigned to or shared with the journal publisher. However, the neurophysiology dataset, like the sequence, should be identified with the researcher or author (with appropriate citation of a reviewed and copyrighted report, if applicable) and free for the researcher to use or to distribute. These datasets should be viewed not as substitutes for the literature. Instead, data and literature should coexist, each informing and adding value to the other.

We thank Esther P. Gardner for many thoughtful contributions to the design of the project, and for supplying data and metadata from her laboratory for development and testing of database and tools. We also thank consultant-collaborators for the somatosensory cortical database project, including Kenneth O. Johnson, Robert H. LaMotte and Harold Burton—as well as members of the invertebrate community too numerous to list—for their insights, comments and datasets. We are grateful for the contributions of Thomas White to interface design and integration. We acknowledge the programming assistance and support of Martin Andersen, Rolf Hanson, John Thaemlitz, Marcello Blinder and Matthew Tepel in earlier versions of databases and tools. This work was supported by the Human Brain Project and the National Institute of Mental Health via MH57153, the National Institute on Neurological Disorders and Stroke via NS36043, and the National Science Foundation via BIR/DBI-9506171.

REFERENCES

- Abato, M. & Gardner, D. 1999 Common data model—a data type definition for biophysical data exchange. *Biophys. J.* **76**, A197.
- Abato, M., Tepel, M., Erde, S. M. & Gardner, D. 1998 A unified neurophysiology database storage schema for time series, histogram, and bivariate datasets. *Soc. Neurosci. Abstr.* **24**, 134.

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990 Basic local assignment search tool. *J. Mol. Biol.* **215**, 404–410.
- Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
- Arbib, M. A., Grethe, J. S., Wehrer, G. L., Mureika, J. R., Tracy, J., Xie, X., Thompson, R. F. & Berger, T. W. 1996 An on-line neurophysiological and behavioral database for the neuroscientist. *Soc. Neurosci. Abstr.* **22**, 916.
- Astrahan, M. M. & Chamberlin, D. D. 1975 Implementation of a structured English query language. *Commun. Assoc. Computing Mach.* **18**, 580–588.
- Baldi, P. & Bruniak, S. 1998 *Bioinformatics*. Cambridge, USA: Bradford.
- Baxevas, A. D. 2000 The molecular biology database collection: an online compilation of relevant database resources. *Nucleic Acids Res.* **28**, 1–7.
- Beeman, D. E., Bower, J. M., De Schutter, E., Efthimiadis, E. N., Goddard, N. & Leigh, J. 1997 The GENESIS simulator-based neuronal database. In *Neuroinformatics: an overview of the human brain project* (ed. S. H. Koslow & M. F. Huerta), pp. 57–81. Mahwah, USA: Lawrence Erlbaum Assoc.
- Benson, D. A., Boguski, M. S., Lipman, D. J., Ostell, J., Oulette, B. F. F., Rapp, B. A. & Wheeler, D. L. 1999 *Nucleic Acids Res.* **27**, 12–17.
- Brinkley, J. F., Myers, L. M., Prothero, J. S., Heil, G. H., Tsudura, J. S., Maravilla, K. R., Ojeman, G. A. & Rosse, C. 1997 A structural information framework for brain mapping. In *Neuroinformatics: an overview of the human brain project* (ed. S. H. Koslow & M. F. Huerta), pp. 309–334 and plate 16. Mahwah, USA: Lawrence Erlbaum Assoc.
- Cattell, R. G. G. 1994 *Object data management: object-oriented and extended relational database systems*. Reading, USA: Addison-Wesley.
- Chen, P. P. 1976 The entity-relationship model—toward a unified view of data. *Assoc. Computing Mach. Trans. Database Systems* **1**, 9–36.
- Chhabra, B., Hardy, D. R., Hundhausen, A., Merkel, D., Noble, J. D. & Schwartz, M. F. 1994 Integrating complex data access methods into the Mosaic/WWW environment. Online document: <http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/DDay/schwartz/schwartz.html>.
- Cimino, J. J. 1998 Desiderata for controlled medical vocabularies in the twenty-first century. *Methods Inf. Med.* **37**, 394–403.
- Clucas, J. & Watson, V. 1994 Interactive visualization of computational fluid dynamics using Mosaic. Online document: <http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/DDay/clucas/clucas.html>.
- Codd, E. F. 1970 A relational model of data for large shared data banks. *Commun. Assoc. Computing Mach.* **13**, 377–387.
- Comer, C. & Jacobs, G. (eds) 1994 Establishing identified neuron databases: a report on workshops held at the National Science Foundation, Arlington, VA, USA.
- Felleman, D. J. & Van Essen, D. C. 1991 Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1**, 1–47.
- Friedman, C., Hripcsak, G., Shagina, L. & Liu, H. 1999 Representing information in patient reports using natural language processing and the extensible markup language. *J. Am. Med. Inform. Assoc.* **6**, 76–87.
- Gardner, D. 1995 Physiology on the world-wide web. *Physiologist* **38**, 187–194.
- Gardner, D. & Erde, S. M. 1995 Criteria for neuronal identification yield universal neurophysiological data types. *Soc. Neurosci. Abstr.* **21**, 146.
- Gardner, D. & Erde, S. M. 1997a Database schemas for electrophysiological datasets may complement genomic and protein sequences. *Biophys. J.* **72**, A118.
- Gardner, D. & Erde, S. M. 1997b Neurodatabases implement evolving neurophysiology data structures using an object-relational model. *Soc. Neurosci. Abstr.* **23**, 1048.
- Gardner, D., Abato, M. & Erde, S. M. 1998 Open neurodatabase access by Java-mediated controlled-vocabulary neuron terms and experimental protocols. *Soc. Neurosci. Abstr.* **24**, 134.
- Gardner, D., Abato, M., DeBellis, R., Erde, S. M., Knuth, K. H. & White, T. 1999 Common data model 2000: open methods for neuroscience data description and interchange. *Soc. Neurosci. Abstr.* **25**, 1910.
- Gardner, D., Knuth, K. H., DeBellis, R. & Abato, M. 2000a Hierarchical attribute-value schema enables extensible controlled vocabularies systematizing neuroscience databases. *Biophys. J.* **78**, 218A.
- Gardner, D., Abato, M., Knuth, K. H., DeBellis, R. & Erde, S. M. 2000b A dynamic publication model for data-driven neurodatabases implements interoperability for neuroscience information resources. *Soc. Neurosci. Abstr.* **26**, 2001.
- Gardner, D., Knuth, K. H., Abato, M., Erde, S. M., White, T., DeBellis, R. & Gardner, E. P. 2001 Common data model for neuroscience data and data model interchange. *J. Am. Med. Inform. Assoc.* **8**, 17.
- Gosling, J. & McGilton, H. 1996 The Java language environment—A white paper. Mountain View, CA: Sun Microsystems Technical Report. Online document: <http://java.sun.com/docs/white/langenv/>
- Grethe, J. S., Wehrer, G. L., Thompson, R. F., Berger, T. W. & Arbib, M. A. 1996 An extendible object-relational database schema for neurophysiological and behavioral data. *Soc. Neurosci. Abstr.* **22**, 916.
- Hamming, R. W. 1962 *Numerical methods for scientists and engineers*. New York: McGraw-Hill.
- Hines, M. 1989 A program for simulation of nerve equations with branching geometries. *Int. Biomed. Comput.* **24**, 55–68.
- Huerta, M. F., Koslow, S. H. & Leshner, A. I. 1993 The Human Brain Project: an international resource. *Trends Neurosci.* **16**, 436–438.
- Institute of Medicine (US) Committee on a National Neural Circuitry Database. 1991 *Mapping the brain and its functions: integrating enabling technologies into neuroscience research* (ed. C. M. Pechura & J. B. Martin) Washington, DC, USA: National Academy Press.
- Jones, E. G. 1986 Connectivity of the primate sensory-motor cortex. In *Cerebral cortex*, vol. 5. (ed. E. G. Jones & A. Peters), pp. 113–183. New York: Plenum Press.
- Knuth, K. H., Abato, M., White, T. & Gardner, D. 1999 Generalizability of common data model 2000 to representation and exchange of data for advanced neuroimaging techniques. *Soc. Neurosci. Abstr.* **25**, 1910.
- Koslow, S. H. & Huerta, M. F. 1997 *Neuroinformatics: an overview of the human brain project*. Mahwah, USA: Lawrence Erlbaum Assoc.
- Lindberg, D., Humphreys, B. & McCray, A. 1993 The unified medical language system. *Methods Inf. Med.* **32**, 281–291.
- Mazziotta, J. C., Toga, A. W., Evans, A., Fox, P. & Lancaster, J. 1995 A probabilistic atlas of the human brain: theory and rationale for its development. *Neuroimage* **2**, 89–101.
- Mazziotta, J. C., Toga, A. W., Evans, A., Fox, P. & Lancaster, J. 1997 Atlases of the human brain. In *Neuroinformatics: an overview of the human brain project*. (ed. S. H. Koslow & M. F. Huerta), pp. 255–308 and plates 14–15. Mahwah, USA: Lawrence Erlbaum Assoc.
- National Institute for Standards and Technology (US) 1993 *Integration definition for information modeling (IDEFIX)*. Federal

- Information Processing Standards Publication 184. Springfield, USA: National Technical Information Service.
- Rowe, L. A. & Stonebraker, M. R. 1987 The POSTGRES data model. In *Proc. XIII Intl. Conf. Very Large Databases*. San Mateo, USA: Morgan Kaufman.
- Schuler, G. D., Epstein, J. A., Ohkawa, H. & Kans, J. A. 1996 Entrez: molecular biology database and retrieval system. *Methods Enzymol.* **206**, 141–161.
- Shepherd, G. M., Mirsky, J. S., Healy, M. D., Singer, M. S., Skoufos, E., Hines, M. S., Nadkarni, P. M. & Miller, P. L. 1998 The Human Brain Project: neuroinformatics tools for integrating, searching and modeling multidisciplinary neuroscience data. *Trends Neurosci.* **21**, 460.
- Shu, Y., Liaw, J.-S., Berger, T. W. & Shahabi, C. 1999 Data mining for neuroscience databases. *Soc. Neurosci. Abstr.* **25**, 2259.
- Sikorski, R. & Peters, R. 1998. XML is hatching. *Science* **281**, 1164.
- Stephan, K. E., Hilgetag, C.-C., Burns, G. A. P. C., O'Neill, M. A., Young, M. P. & Kötter, R. 2000 Computational analysis of functional connectivity between areas of primate cerebral cortex. *Phil. Trans. R. Soc. Lond. B* **355**, 111–126.
- Wilson, M. A., Bhalla, U. S., Uhley, J. D. & Bower, J. M. 1989 GENESIS: a system for simulating neural networks. In *Advances in neural information processing systems I*, (ed. D. Touretzky). San Mateo, USA: Morgan Kaufman Publ.