Model Formulation

Common Data Model for Neuroscience Data and Data Model Exchange

DANIEL GARDNER, PHD, KEVIN H. KNUTH, PHD, MICHAEL ABATO, STEVEN M. ERDE, MD, PHD, THOMAS WHITE, MD, ROBERT DEBELLIS, ESTHER P. GARDNER, PHD

Abstract Objective: Generalizing the data models underlying two prototype neurophysiology databases, the authors describe and propose the Common Data Model (CDM) as a framework for federating a broad spectrum of disparate neuroscience information resources.

Design: Each component of the CDM derives from one of five superclasses—data, site, method, model, and reference—or from relations defined between them. A hierarchic attribute-value scheme for metadata enables interoperability with variable tree depth to serve specific intra- or broad interdomain queries. To mediate data exchange between disparate systems, the authors propose a set of XML-derived schema for describing not only data sets but data models. These include biophysical description markup language (BDML), which mediates interoperability between data resources by providing a meta-description for the CDM.

Results: The set of superclasses potentially spans data needs of contemporary neuroscience. Data elements abstracted from neurophysiology time series and histogram data represent data sets that differ in dimension and concordance. Site elements transcend neurons to describe subcellular compartments, circuits, regions, or slices; non-neuroanatomic sites include sequences to patients. Methods and models are highly domain-dependent.

Conclusions: True federation of data resources requires explicit public description, in a meta-language, of the contents, query methods, data formats, and data models of each data resource. Any data model that can be derived from the defined superclasses is potentially conformant and interoperability can be enabled by recognition of BDML-described compatibilities. Such meta-descriptions can buffer technologic changes.

J Am Med Inform Assoc. 2001;8:17–33.

Neuroscience spans a range from biochemistry through physiology, pharmacology, and anatomy to development, behavior, learning, neurology, and psychiatry. Investigations probe nervous systems using techniques for data collection and analysis derived from fields as diverse as genomics, biophysics, computer science, and psychology. The scope and range of neuroscience data is thus ever more complex, and the number of laboratories acquiring and analyzing data digitally continues to increase.

Many current and developing neuroscience data resources are built on data types, techniques, interchange methods, and models that are local to disparate neuroscience communities. However, there is a significant and growing need among neuroscientists to exchange and compare complex and disparate experimental data. Consistent, predictable, flexible,
and descriptive data description and examination methods may aid in the evaluation and analysis of raw and derived data, the verification of results, and the testing of models.

This data diversity underscores the complex requirements for interoperability among different domains of neuroscience and emphasizes the care needed to design any unitary schema for experimental neuroscience data management. For example, neurophysiology recordings are collected from sources and sites ranging in level from the molecule to the whole organism. To make the resulting data sets accessible to investigators in any of several domains of neuroscience, data models are needed to place them in context. Such models need to incorporate a non-simply-connected hierarchic multilevel anatomy that is sufficiently intuitive to encourage use by suppliers and consumers of data.1–6

The timing appears appropriate for exploration of interoperability standards for neurophysiology. It is not too early, because there is growing experience with projects developing individual, often complementary, approaches to data storage and distribution that reflect the present fragmented state of data representation. It is not too late, because current database projects are research prototypes, adaptable rather than committed to incompatible schemas.

In this paper, we propose the Common Data Model (CDM) not as a complete scheme for interoperability but as a framework for planning for data exchange in neuroscience.

**Design for Interoperability**

**Interface Design Can Advance Interoperability**

Interoperability among developing neuroscience data resources can be served by providing universal standards for data exchange that compatibly describe the data models, formats, and contexts of different resources. To minimize the overhead for researchers, these methods should allow quick and easy identification of data type, source, and context, aiding interpretation. Without requiring changes in the underlying data models, such standards should permit not only exchanging data sets, but also querying and responding with information about methods, models, anatomy, and references.

These standards for compatibility can be provided by a high-level meta-specification that encompasses the domains over which interoperability is required. Both standards and individual focused schemas should be describable in this meta-representation and expressible via interfaces with defined syntax and semantics. Interfaces for data exchange describe data and specify data formats, interfaces linking data models recognize data model intersections, interfaces that transmit data context both mediate metadata exchange and generate, parse, and translate queries. Interfaces can be designed either to conceal the internal representations of the resource they link or to publish it, thus explicitly describing information about data and data model internals.

Interfaces, unlike data or data models, need to be uniform, so that each interface is compatible with any potential data resource that requires interchange. Uniformity obviates the need to design ad hoc interfaces between every pair of databases, a task whose effort grows as the square of the number of resources. To achieve acceptance by the community, these interfaces and related techniques should be intuitive, complete, extensible, multiprotocol and, ideally, human-readable. In addition, their design should recognize the eventual need for graceful migration to newer techniques and architectures.

**Design Goals**

Informed by investigators’ needs, the nature of neuroscience data, and case studies, we list design goals for the CDM in particular and for neuroinformatics standards and tools in general. Although no model will achieve all these goals, protocols that address them (both at a high level of domain meta-description and at the low level of data formats and Internet protocols) will have a greater chance of serving the field. The utility of neuroinformatics methods is likely to be enhanced if they are:

- **Complete.** No single data model that is broad enough to express the entire current domain of neuroscience is likely to be of sufficient specificity to adequately serve focused domains of that discipline. Because of the extent and variability of the domain of experimental neuroscience, this goal is likely to be met only through interface definitions that conform to the additional goals listed below.

- **Extensible.** The model must not be limited to the current domain of neuroscience but must be flexible enough to respond to changes in the field.

- **Compact.** The model should be as terse as possible in expressing only those data descriptors necessary to select and interchange data sets.
- **Efficient.** Implementations of the model should use the minimal set of necessary resources—time, space, bandwidth, and learning. The model should not demand excessive processor time, storage devices, network bandwidth, or maintainers’ time. This goal is aided by compactness.

- **Simple yet scalable.** To be achievable with limited resources and maintainable with ease, the model should be simple enough to be implemented for a small project yet useful for large data repositories and exchange schemes as well. This goal typically competes with completeness and extensibility.

- **Platform-independent.** A good model minimizes assumptions about hardware and software. A data exchange standard, almost by definition, needs to be independent of any systems being used to implement it. Although this goal is directed here to those that actively maintain models and resources for data archiving, it should be applied as well to interfaces for users submitting and acquiring data.

- **Convertible.** It should be easy to convert data into other formats. The model needs to recognize the data and metadata needs of multiple existing and future formats.

- **Human-readable.** The structure of the model, and thus its comprehension, should be interpretable without machine translation or familiarity with programming meta-language conventions. Ideally, neuroscience researchers should be able to examine data in the model and both classify and work with the data without using any specialized tools.

- **Analytic.** The model should facilitate data analysis with both general-purpose and domain-specific tools.

- **Familiar.** The model should use intuitive conventions familiar to users in the neuroscience community. Finally, implementation protocols should similarly be consistent with current standards and protocols for networked communications.

### Data-driven Designs Aid Interoperability

Our models and methods for interoperability, like those for sequence and protein structure databases, are data driven; that is, they are focused primarily on data sets, data representation, and allied metadata. Data are ideally persistent and model-independent. Furthermore, data quality for several neurophysiology techniques is largely observer-independent. The reanalyzability of many data types both focuses and justifies the schema. Sets of data descriptions and definitions, relations between data, and descriptive attributes for data, sites, methods and models provide a data model that is easily implementable and broadly acceptable to neurophysiology communities. This focus on data may be extended beyond data banks to other classes of resources, including reference databases, registries, and collaboratives.

The data model incorporates knowledge representation only secondarily, as it is needed for data specification. This excludes ontological definitions as well as is-a and has-a relationships, such as:

- shaft and spine are-components-of dendrites
- conductance is-a property-of membranes
- synapse is-a relation-between two neurons

The design decision recognizes that the knowledge base of contemporary neuroscience is highly complex, dependent on interactions within and among multiple levels of anatomic organization and functional systems. This knowledge base is also fluid, with both broad concepts and usable specifics currently under active exploration. A data-driven model avoids many of the complexities of designing and parsing with inference engines, the heterogeneous semantic nets that would be needed to implement an evolving knowledge-driven schema.

### Common Data Model

#### Design Based on Two Prototype Databases

For two neuroscience database projects, we initially designed the CDM as an open and intuitive set of structures and methods for brain data and metadata that could be readily used by experimental neurophysiologists and implemented using object-relational technology.

In this design, we targeted data sets for which reanalysis was likely to be productive, including time series and histograms describing electrode recordings of neuronal activity. These data were organized in a hierarchic experiment > view > trace model, which has a visual syntax resembling that of journal figures. The model is now general enough to encompass cortical units, invertebrate neurons, and most methodologies for intracellular, patch, extracellular single-unit, and multi-electrode recording.

Because the selection, interpretation, and further analysis of such data sets depends strongly on recording conditions and methodology, as well as on functional and anatomic characteristics of the neurons from which records are made, the CDM design
included extensive descriptive metadata. These controlled vocabulary metadata descriptors of protocols and neuron properties provide lexical syntax and semantics and enable searches. Notice that we here define metadata as neurobiological descriptors characterizing neurophysiology data sets. This usage differs from other common definitions of the same term, including the database-related definition of metadata as comprising the internal structure and organization of a data resource implementation.

The intuitive data model design was also embodied in project-designed viewer and query tools aiding database access and data exchange within the two targeted areas of neurophysiology. Java and extensible markup language (XML) technologies enabled data search and dynamic display of data on several contemporary hardware and software platforms, aiding user interoperability.

Extending the Common Data Model to Aid Interoperability Between Databases

We now propose the CDM as a standard to enable interoperability between neuroscience data resources with related but disparate content. Whereas the initial design and associated user tools implicitly encapsulated the data model, the extended CDM explicitly specifies its data model and its standards for data and data model sharing. It abstracts data types, dataset formats, and metadata to span additional domains of neuroscience. This abstraction includes definitions of superclasses for data, recording site, methods, models, and references. It is designed to allow for future expansion of metadata to explicitly specify data-set structure and domain knowledge.

When fully implemented using an XML-based interface called biophysical description markup language (BDML), as well as a hierarchic attribute value implementation of controlled vocabulary (each described below), the CDM will have the capacity to mediate among disparate neuroscience database projects and similar resources with compatible rather than identical data and data models. Beyond its role in supporting community data resources, the CDM is thus intended to supply neuroinformatics with an example, not a mandate, for interoperability.

Superclasses and Elements

The core of the CDM is the identification of abstract entities from which components of compatible data models can be derived. The range of neuroscience data and metadata suggested that a flat data model characterizing data records alone would be insufficient, favoring instead a set of five superclasses—data, site, reference, model, and method elements (Figure 1). These superclasses were selected to be general enough to express data models from disparate domains of neuroscience, enabling a template-based approach to interoperability.

Data elements include data sets and wrappers. Site elements transcend neurons to describe subject organism, slice, region, circuit, subcellular (axon, soma, dendrite, and subcategories) and submembrane compartments. Method elements encode protocols and preparations, whereas model elements define data-related hypotheses and parameter sets describing mathematical or similar models. Reference elements include bibliographic record types.

From each superclass, specific neuroscience data types are defined in an inheritance tree. Each is a first-class entity characterized by descriptive attributes defining and characterizing either the superclass or specific classes derived from it. Attributes are also searchable, providing specificity for queries selecting among similar data. Relations link entities derived from the same or disparate superclasses.

Data Element Central to Data-driven Databases

To serve data-driven databases, the CDM allows the specification of a wide range of data types and wrappers, each deriving from the superclass data element. Unlike sequences, much neuroscience data is not self-
defining, emphasizing that data sets require a defined set of metadata attributes that describe multiple aspects of particular data sets and distinguish them from others.

In our cortical somatosensory and invertebrate identified neuron databases, the CDM is implemented using an object-relational scheme that stores numeric data sets in binary large objects (BLOBs) and allied metadata in relational tables. However, the model can define both internal implementation-dependent and external implementation-independent representations of data types.

In these databases, neurophysiology data types derived from data element include hierarchic wrappers for experiment, view, and trace, but the superclass scheme allows for extension or modification by any neuroscience community for which other classes of data elements—either more granular or more encompassing—are more natural units for data acquisition, analysis, and exchange. Data elements can thus be defined, for example, as individual data points or as any wrapper scheme, including hypothesis test, sequence, or animation.

**Site Representation in Common Data Model**

Neuroscience data are recorded from sites ranging in scope from micrometer patches of individual cell membrane through brain regions varying in size and specificity, to gross motor systems. These are abstracted as site elements, of which neuron, compartment, and region are among the definable subclasses. A sample site element definition for cortical neurons is shown in Figure 2.

Among neurophysiology communities, operational descriptions of recording sites differ, even for such fundamental entities as neurons. For mammalian electrophysiologists recording in vivo, the identity of the neuron ("single unit") providing data most often has scope limited to the set of recordings it yields, usually in an individual experiment. It can be specified only approximately, perhaps by functional or anatomic areas, by depth or by coordinates, or

---

**Figure 2** Site element description for the somatosensory database. Site element abstracts neuronal anatomy, encompassing neurons, regions, and other anatomic structures. For cortical microelectrode neurophysiology, recording sites are primarily cortical neurons, defined here as subtypes of mammalian neurons. Attribute names ending in _CV have controlled vocabulary values. In this entity-relationship diagram, subclasses of site element are shown with names and data types of characteristic data and descriptive metadata attributes. Diamonds indicate relations between sites and other elements, including controlled vocabulary for neurotransmitter and receptor classes (not shown).
through physiologic parameters such as receptive field or firing patterns.

In contrast, for many invertebrate preparations, neurons are stable identified individuals, allowing libraries to be constructed and recordings assigned confidently to a specific site. The molluscan neuron scheme, presented in Figure 3, is thus different from that for mammalian neurons, but each is derived from the same site element superclass.

For the somatosensory database, the primary site element is cortical neuron, itself a subclass of mammalian neuron; defining attributes are specified in both mammalian and cortical neuron classes (see Figure 2). For neuronal location, we incorporate any of several attributes—cytoarchitectonic or functional area, stereotaxic or other coordinates, depth and cell type. The basic scheme is extensible beyond the cortex to encompass other mammalian neurons that may be relevant to cortical physiology.

Experimentally determined functional neuronal response profiles are often the major distinguishing characteristic of cortical neurons. Descriptive identifying metadata include the neuron’s receptive field, location, and firing patterns. For receptive field, the data model specifies location, modality, and adaptation, each with its own controlled vocabulary tree. Recognizing that receptive fields are often broad, multi-modal, and stimulus-dependent, each neuron may have multiple sets of these complex attributes.

Acknowledging community use of disparate criteria, several attributes for neuronal location are defined,
including Brodmann and functional areas, depth, and cell type; all are allowed, but none required. Any of several atlas coordinates can be specified as well.

These neuronal site elements are inadequate for other methodologies in which larger- or smaller-scale anatomic structures are examined. Imaging and behavioral data sum multi-neuronal, regional, or whole-organism responses. Subneuronal site elements can include anatomic distinctions such as soma, apical dendrite, spine or shaft, or segments with attributes defined to permit comparison to compartmental models.

Subjects can also be defined as site elements, with specific subcategories for classes of experimental animals, human experimental subjects, or patients. Hierarchies of sites can be defined, as well as relations connecting and grouping sites. Such relations can be analogous to those developed for data element, to place sites of finer granularity within larger regions or nuclei or to specify circuits as site elements and connections as site-site relations.

The Institute of Medicine report that helped define neuroinformatics\textsuperscript{10} specified neuroanatomy as an organizing principle for brain information. Similarly, an NSF report that urged establishment of identified neuron databases recommended that the neuron be the focus of all such efforts.\textsuperscript{12} This reliance on neuroanatomy spurred development of the site element, as well as the site–data relation, which provides a powerful tool for linking data sets to the anatomic locations from which they were recorded. However, as important as organizing principles are techniques, hypotheses and models, and literature references, and each of these, along with site element, gives rise to first-class elements complementing the data element that is the core of the data-centric model.

**Common Data Model Abstracts Elements for Method, Model, and Reference**

Because most neuroscience data are highly dependent on experimental techniques and protocols, we abstract these as method elements, a category that also includes modeling simulation engines or schemas. For the two neurophysiology databases, protocol metadata are sets of controlled-vocabulary terms organized in triads of effector, temporal pattern, and target. Effectors are hierarchic, each arising from broad categories—chemical, electrical, mechanical, visual, auditory, thermal, behavioral, and surgical. Target trees span the whole preparation to specific organs, areas, or neurons.

The CDM makes provision for model or simulation parameter sets as subclasses of model element. This distinction between parameter sets and modeling engines reflects the separation between conceptual and methodological phases of model building. Model elements can be defined to be compatible with a wide range of models, from shallow or phenomenological to deep or biophysical. In addition, databases that include hypotheses, concepts, diagnoses, or fits to experimental data can define each of these as subclasses of model element. Maintaining such parameters as a distinct superclass, rather than incorporating them in data elements, recognizes that data are distinct from models; maintaining the distinction removes limits on post hoc analysis. Indeed, an important rationale for data exchange is to stimulate these analyses, independent of prior fits or hypotheses.

However, distinction should not imply separation, since the schema allows relations linking data, method, and model elements. This can include any of several representations. One is a semantic tree relating data sets, not hierarchically (as our existing data element wrappers do) but in a protocol-dependent way—\textit{wild-type/mutant, control/experimental, derived-from, or pre-during-post}. To provide explicit connections between hypotheses and data sets testing, confirming or disproving them, semantic trees are needed for hypotheses within model elements and for relations between model element and data element, including \textit{defines, depends-upon, tests, and predicts}.

The reference element superclass is offered as a standard description of publications and bibliographic information for the neuroinformatics community. One of its unique features is a \textit{contribution} field to specify the nature of the author–article relationship, enabling future utilization by journals to define the areas for which each author of a multi-author work is responsible.

**Metadata Extensions**

The cortical and invertebrate databases and project-specific tools encapsulate an implicit model of data structures and methods. For this reason, identifying and selecting data sets required that only values of a narrow set of metadata be specified. This set is limited in scope to attributes that vary with experimental design and results and, in level, to descriptions useful for selection and interpretation of data sets by informed investigators. Such explicit controlled-vocabulary metadata attributes of data, site, and method elements serve to specify preparation,
methodology, protocol, experimental conditions, and experimenter, and thereby enable selection and interpretation of data sets by users familiar with the domain.\textsuperscript{11}

To aid database federation and interoperability and offer standards for automated search, query, and exchange of biophysical data, we plan to expand the data model beyond the high-specificity experimental level metadata, implemented by the two neurophysiology databases, to include two additional levels of metadata. One is low-level data-set structure—the granularity, dimensionality, format, and precision of each data set. The other is high-level—metaknowledge representation of the syntax and semantics of the particular knowledge domain, enabling automated recognition of database contents.

Metadata scope is similarly expandable to include local and analytic types as well as the currently implemented global metadata. Local metadata provide audit information often found only in laboratory notebooks, such as date and time of experiment or animal designator. Analytic metadata, analogous to

---

### Recording technique:

<table>
<thead>
<tr>
<th>extracellular</th>
<th>intracellular</th>
</tr>
</thead>
<tbody>
<tr>
<td>single electrode</td>
<td>voltage-clamp</td>
</tr>
<tr>
<td>multielectrode array</td>
<td>cell-attached</td>
</tr>
<tr>
<td></td>
<td>o-o</td>
</tr>
<tr>
<td></td>
<td>i-o</td>
</tr>
<tr>
<td>whole-cell</td>
<td>nerve cuff</td>
</tr>
<tr>
<td>macroelectrode</td>
<td>field/surface</td>
</tr>
<tr>
<td></td>
<td>scalp</td>
</tr>
<tr>
<td></td>
<td>cortical</td>
</tr>
<tr>
<td></td>
<td>neuro-intrinsic</td>
</tr>
<tr>
<td></td>
<td>neuro-via-dye</td>
</tr>
<tr>
<td></td>
<td>light intensity</td>
</tr>
<tr>
<td></td>
<td>MEG</td>
</tr>
<tr>
<td></td>
<td>radial</td>
</tr>
<tr>
<td></td>
<td>planar</td>
</tr>
<tr>
<td></td>
<td>mechanical</td>
</tr>
<tr>
<td></td>
<td>pressure</td>
</tr>
<tr>
<td></td>
<td>force</td>
</tr>
<tr>
<td></td>
<td>displacement</td>
</tr>
<tr>
<td></td>
<td>translation</td>
</tr>
<tr>
<td></td>
<td>rotation</td>
</tr>
<tr>
<td></td>
<td>radial</td>
</tr>
<tr>
<td>sound intensity</td>
<td>thermal</td>
</tr>
</tbody>
</table>

### Motor behavior action:

| hand/arm               | flex/extend                         |
|                        | pronate/supinate                    |
|                        | abduct/adduct                       |
| manipulate             | reach/point/press                   |
|                        | crack                               |
|                        | grasp                               |
|                        | palpate                             |
|                        | translate/rotate/push-pull          |
|                        | reject                              |
| leg/limbs/foot         | flex/extend                         |
|                        | pronate/supinate                    |
|                        | abduct/adduct                       |
| locomote               | rotate                              |
|                        | step                                |
|                        | swim                                |
| eyes/extraocular muscles | fixate                           |
|                        | saccade                             |
|                        | smooth pursuit                      |
|                        | vergence                            |
|                        | nystagmus                           |

### Cortical neuron functional area:

| somatosensory          | SI                                  |
|                        | SI                                  |
|                        | insular/retroinsular                |
|                        | postauricular/gilfd                |
|                        | posterior/parietal                 |
| motor                  | M1/F1                               |
|                        | SMA/presMA/F3/F6                   |
|                        | cingulate                           |
|                        | MEF                                 |
|                        | PMd/F2/17                           |
|                        | PMv/I4/F5                          |
|                        | FEF                                 |
|                        | preF/preM/orbitoF                  |
| visual/multi           | V1                                  |
|                        | V2                                  |
|                        | V3                                  |
|                        | V3a                                 |
|                        | V4                                  |
|                        | VP                                  |
|                        | MT                                  |
|                        | temporal                            |
|                        | parietal                           |
|                        | MST                                 |
|                        | LIP                                 |
|                        | VIP                                 |
|                        | AIP                                 |
| auditory               | A1                                  |
|                        | R/L/CM/lateral                     |
| speech/language        | Broca's                            |
|                        | Wernicke's                         |
| olfactory              | piriform                           |
|                        | periamygdaloid                     |
| gustatory              | hippocampal                        |
|                        | CA1                                 |
|                        | CA3                                 |
|                        | dentate                            |
|                        | subiculum                          |
| entorhinal cortex      | amygdala/periamygd cortex           |

---

**Figure 4** Samples of controlled-vocabulary values for each of four attributes. One (recording technique) is shared by cortical and invertebrate databases, one describes molluscan neuron receptive fields, and two are used as neuron descriptors and search terms in the cortical database. Hierarchies expand to the right, with broader terms shown to the left in each column. The hierarchic attribute-value search algorithm shown in Figure 6 enables selectable specificity.
material frequently located in the methods section of publications, include values required for further analysis of data, such as noise filter settings and sampling rates as well as dimensions and units.

Hierarchic Attribute-Value Implementation of Controlled Vocabularies

Controlled vocabularies enable unified definitions and targeted search for attribute values—descriptive key words describing such items as recording types and techniques, neuron anatomy and receptive field, and experimental protocols. To serve experimental neurophysiology, we have devised a set of controlled vocabulary terms based in part on the Society for Neuroscience key word list, to partially define a data definition language and lexical grammar for neuroscience. Specific design decisions and selection criteria are detailed elsewhere, and comparisons with standard clinical schemes are presented below.

In addition, the controlled vocabularies are hierarchic. For example, Figure 4 illustrates several hierarchies defining attributes that are useful in somatosensory and molluscan neurophysiology. In our current implementation, controlled vocabulary values are served dynamically as XML and presented via Java hierarchic menus so that users always see current terms. Each controlled vocabulary term can be associated with a short definition, to aid selection by users. Although this scheme is general enough to allow for variable definitions, depending on language or on the source of the enquiry, we have specified a unitary definition for each term.

Hierarchic Attribute-Value Implementation is Efficient

Both attributes and their controlled vocabulary values are database objects, represented in a simple hierarchic attribute-value (HAV) relational table schema (Figure 5). The HAV representation may be viewed as an extension of the well-known entity-attribute-value (EAV) scheme with the addition of a second table to implement the hierarchy. The EAV scheme has proved useful for neuroscience data as well as, in its traditional role, for the electronic patient record. Whereas EAV aids efficient storage of sparse data, HAV permits efficient specification of context and constructs simple semantic trees. Although the HAV specification is independent of any particular implementation and, indeed, can be represented in XML or Java, in relational or object-relational databases a pair of tables is sufficient to represent controlled vocabulary terms, additional clarifying glossaries enhancing entries via short definitions, and the attributes specified by each term, as well as the multilevel hierarchy.

For simplicity and ease of use, we implement the vocabulary values for each attribute as a tree rather than as the more general directed graph. Trees are relatively shallow, with a depth rarely exceeding four from the explicit root. Depth is also restricted by excluding implicit tree roots shared by every value within a focused domain of neuroscience. Such shallow representations reduce the need for tree traversal during recursive query parsing as well as manual tra-
A query term finds an attribute of value value if and only if:

\[ \text{term} \equiv \text{value} \mid (\text{parent}_\text{term}) \text{value} \]

Enabling fine intra-domain queries at specified selectivity ... hand finds:

\[ \ldots .\text{cutaneous.glabrous_skin.hand} \]
\[ \ldots .\text{cutaneous.glabrous_skin.hand.volar–palm} \]

but not \[ \ldots .\text{cutaneous.glabrous_skin} \]

Broad interdomain queries are implemented as well ... cutaneous finds:

\[ \ldots .\text{cutaneous.glabrous_skin} \]
\[ \ldots .\text{cutaneous.glabrous_skin.hand} \]
\[ \ldots .\text{cutaneous.hairy_skin} \]

Figure 6 Hierarchic attribute-value (HAV) outer-match search algorithm enables selection of either operational or high specificity. The algorithm finds entries described by the search term or any of its more specific child terms. This approach favors the broad returns consistent with operational specificity yet allows high specificity searches when desired. The HAV schema is also compatible with exact-match search algorithms.

Hierarchic Attribute-Value Design Empowers Searches with Selectable Precision

For controlled vocabulary to be comprehensive yet specific, attributes must be able to take values of varying degrees of precision. Given a tree representation, queries and descriptors can each be specified at any of several levels, rendering exact-match searches an imperfect scheme. Moreover, highly specific values, serving a focused domain, may be inappropriate for queries from a database concentrating on a different domain of neuroscience. Search methods must accommodate this variable specificity.

To provide the capability for both broad and focused queries, and to minimize misses, the HAV schema enables both submitter or requester to specify any desired degree of granularity or precision. An intuitive outer-match search algorithm returns hits if and only if a specific search term is found in the tree extending from the root to the value of the descriptive attribute (Figure 6). Search terms thus find data sets described either by the identical term or by a more specific term from the same branch of the tree of values for that attribute. The HAV schema is also compatible with higher-selectivity exact-match searches.

Hierarchic Attribute-Value Design Aids Interoperability

The HAV schema is designed to permit the same extensive controlled-vocabulary metadata set to serve both cross-disciplinary and intradomain queries. As noted above, within a domain many controlled vocabulary hierarchies can be subsets, arising from implicit roots that span more of neuroscience. For example, the receptive field root as implemented for our cortical database is somatosensory, but the true hierarchy root is system, spanning motor, cognitive, and emotional as well as sensory functions.

Controlled vocabulary serving a CDM benefits from this common top-level hierarchy that both spans multiple domains and also enables domain-dependent specificity. Within a domain, it may be necessary to traverse the tree several levels from the root to provide sufficient specificity. However, for interdomain specificity, top-level attributes are likely to be more practical; otherwise, every resource would have to incorporate every other resource’s complete data model and attribute-value set, requiring development of a near-complete ontology for neuroscience. Variable specificity allows inter-domain queries to be finely detailed selections of specific data sets. For broader cross-disciplinary intradomain queries among conforming data resources, more shallow traversing of the same controlled vocabulary trees is sufficient.

The controlled vocabulary scheme we have developed is thus more reductionist and less comprehensive than many other contemporary methods. This simplicity is enabled by the use of HAV, the decision to implement operational rather than selective specificity, and two added dimensions of context. One is by the attribute class itself; that is, the same term may be used as a descriptor for more than one attribute, but the term appears and is selected only in the context of the specific attribute. The other dimension is the branch of the tree giving rise to the term—another advantage of using a tree rather than a directed graph to link terms. Thus, hand can be a term in both a sensory and a motor hierarchy.
As recognized by the Institute of Medicine neuroinformatics report,\textsuperscript{10} data models should be designed to permit changes required by evolution of the discipline. With controlled vocabulary, the implementation must accommodate changes to attributes as well. We recognize the difficulty of managing controlled vocabulary so that inevitable advances in neuroscience do not make either prior entries or the user interface obsolete.

New techniques and concomitant new classes of observations can be implemented with new values for attributes and even new attributes (representing extensions and modifications to the data model) so long as two conditions are fulfilled: 1) neither existing terms nor prior entries characterized by those terms are made obsolete, and 2) new terms and attributes are in place before or at the same time as the first submission of data that are properly described by them. The first requirement will be met so long as the data model is at all times comprehensively current; failures imply that a current data model is in fact inadequate to describe existing data. Extensions to the data model that are truly new thus extend a hierarchy rather than replace existing entries. Such modifications are facilitated by initial design decisions that provide sufficient scope of top-level values (preferably with approximately equal utility) for each attribute to adequately span the domain. If so, the implementation scheme allows increasing tree depth with no loss of data selection. Conversely, inadequate spanning by top-level values prevents expansion of the depth of an existing hierarchy, which may result in orphan data.

**Extensible Markup Language for the Common Data Model**

The CDM can be used to define similarities between data resources. It thus potentially satisfies a goal for interoperability; namely, in being applicable to any neuroscience data resource yet sufficiently specific to be of use to each resource that adopts it. However, interoperability also requires mechanisms, including interfaces, to make the definitions available and thus link disparate data resources. Such mechanisms inform both investigators and a federation of interoperable databases what data are available remotely, where they are located, and what criteria can be used to select, acquire, examine, and evaluate them.

To develop interfaces that allow commonalities in the CDM to mediate database federation and automate data exchange, we have proposed and begun to implement a schema that allows any CDM-derived data model to be described in a compatible, human-readable and machine-readable format, using the emerging standard XML.\textsuperscript{16} BDML, an XML-based data and data model description method, is designed to automate specification of data model commonalities and may thus enable the exchange of queries and data sets that interoperability requires. It extends XML with neurophysiologic vocabulary and formats optimized for data exchange, allowing any model to be defined and described using a standard language and set of tools. Figure 7 presents an example of the use of BDML to define components of the CDM.

BDML serves as an open, single standard for coordinating data models and specifying formats that does not require other database resources to alter their existing data models. To enable interoperability, each resource needs only to describe its existing data model in BDML; enable methods for accepting BDML queries and converting to local query methods, using XML parsers; and deliver data and metadata wrapped in (or referenced by) BDML.

Evolution of data models and formats, rather than making prior interfaces obsolete, requires only re-specification of the new format in the ongoing BDML.
meta-representation. BDML can be used to formulate the master specification of a database’s data model, from which SQL code, entity-relationship diagrams, and Java classes can each be derived. Following community review, BDML will be proposed as a draft standard for the World Wide Web consortium and other appropriate standards bodies.

XML was selected as the basis of BDML because of its versatility; it can be used to describe data models as well as data sets and can provide data checking for both. There is also a body of experience with and general acceptance of not just XML but SGML (and HTML), on which it is based. Even before XML, Beeman et al. postulated that SGML and data type definitions could be used as a neuroscience database description language. XML is, in addition, independent of platform architecture and implementation, and it is human-readable and easily understood. Furthermore, informal discussions with neuroinformatics developers indicate that XML and the related resource description framework (RDF) are more likely than conventional unified modeling language or object-dependent design schemas to be adopted by the Human Brain Project and other neuroscience resources. Because the XML specification excludes non-unicode data from XML documents, BDML requires that non-text-encoded numeric data sets be transmitted using embedded XML pointers that specify Universal Resource Indicator (URI) and type.

There are many active and proposed extensions to and utilizations of XML for data and metadata definition in the sciences and elsewhere. For example, evolving extensions include XML Schema and XML Data. There are ongoing efforts to fuse XML, Universal Modeling Language, and Meta Object Facility into a global scheme for data definition called XML, or XML Metadata Interchange. XML parsers, viewers, editors, and libraries are under active development, easing the design of schemas. XML name spaces provide a mechanism relating BDML-defined tags to the BDML data type definition, allowing attribute definitions to be neurophysiologically descriptive and intuitive without concern for duplication of terms used in other schemes.

Initially, we are readying data model descriptions for each of our two database projects, in the form of a public XML data type definition. These definitions will make available the CDM data type definition and our data dictionary of internal types, thus expanding our present implicit type-specific structure with optional explicit human-readable text implementations of all metadata levels. This will provide the neuroinformatics community with a description of our data and query attributes as well as a methodology to generate and submit queries and return data, structured and informed with metadata. Such descriptions would also serve as examples of a proposed standard meta-framework, useful for testing the ability to articulate specific data definitions for other resources.

**Case Studies Explore the Generality of the CDM**

We intend the CDM to eventually span all neuroscience, not just the microelectrode time-series records and descriptions of neuronal recording sites currently implemented in our two initial database projects. We explore the generality and scope of the CDM by evaluating the utility of the five top-level superclasses to define, and BDML to mediate, the specifics of data models very different from the microelectrode-derived original design. We selected, as fields with markedly different techniques and data types, functional neuroimaging of auditory responses and neuropsychiatric drug response data.

**The Common Data Model Structure Is Compatible with Neuroimaging Data**

The increasing use of functional imaging techniques—including multimodal imaging of individual subjects—has emphasized the importance of comparison of data obtained by multiple recording modalities as well as the need for uniform protocols for data and metadata description. To explore the applicability of the CDM, we selected a set of multimodal auditory data as representative of these imaging-based data sets. If the data types and data models that are natural to single-mode or multimodal experimental neuroimaging can be derived from the five CDM superclasses and specified using BDML, exchange and comparison of these data with other neuroscience data resources would be simplified.

The CDM data element designed for microelectrode recording includes data types and wrappers that are directly applicable or extensible to imaging modalities. Multichannel EEG/MEG data can be represented as views containing sets of traces, each a time series. Only values of controlled vocabulary metadata, not attributes, need enhancement. Distinct modalities can be accommodated by the use of multiple views, for instance by the use of dual views to link related EEG and fMRI data.

Figure 8 shows sample subtypes of trace and view designed to accommodate two classes of non-micro-
electrode neuroscience image data sets. As noted above, the experiment > view > trace wrapper set serving the somatosensory and invertebrate databases allows extension or modification by specifying units for data of different granularity. Data element subtypes can be defined to include not only new data types but also the most appropriate wrapper classes for images and image sets commonly produced by several sensor technologies. A two- or three-dimensional image such as those obtained by MRI, fMRI, CAT, or PET, can be represented as a unitary 2D_image_trace or 3D_image_trace; fMRI image slices can be represented as sets in which each trace is a 2-D array of scalar values. Indexing metadata can be defined to order traces within such a view array. A sequence wrapper accommodates related data sets at multiple time points, thus defining animations.

Site element can give rise to additional subclasses to serve neuroimaging, including sensor, subject, and landmark. Surface EEG/MEG and other techniques require that sites be abstracted to accommodate locations of recording sensors where these are distant from neural sites. Region-of-interest representation is often data-defined yet mapped to sites. The CDM accommodates this by means of the site-data relation.

In some cases, elements of the model accommodate imaging data with changes to the controlled vocabulary for metadata. In other cases, subclasses of exist-
ing model components are derived, adding or modifying characteristic attributes. Whether such additions to the schema are focused or extensive, it appears that each can be derived from one of the five superclasses and, potentially, described in XML.

Conceptually, the CDM thus appears adequate to describe neuroimaging data, including those derived by electrophysiologic techniques such as EEG and MEG and imaging techniques such as MRI, fMRI, PET, and optical imaging. However, MRI, fMRI, and MRS data are complex, often resulting from ad hoc pulse sequences, significance criteria, transformations, and normalization methods, and further tests are needed to see whether a manageable set of metadata attributes and values, as well as relations between the needed site, data, method, and model elements, have practical utility for data description and exchange.\(^2,10,22\)

**Patient Data**

To explore the applicability of the CDM as an interface to neuroscience data markedly distinct from either microelectrode or functional imaging neurophysiology, we asked whether the data types and models appropriate to psychopharmacologic multi-patient studies could be derived from the top-level CDM classes.\(^23\) As already noted, both waveform and image studies and time series of such studies can be derived from the data element superclass, which can also accommodate clinical observations and outcomes data. The experiment > view > trace hierarchy is ill-suited to patient data and will require the design of specialized data element wrappers to encompass clinical data and metadata.

The top-level separation in the CDM between methods, designed for planned events such as protocols and administration methods, and data, encompassing results obtained from either experimental system or patients, appears to map well to clinical data, too. Conceptually, indications and diagnoses can be derived from top-level model elements, and procedure codes, guidelines, treatment regimens, and pharmacologic agents from method elements.

We do not minimize the difficulty of implementing such conceptually derivable linkages. The broad set of existing standards, with varying degrees of overlap and penetration but each with an individual data model and vocabulary, present a particular challenge.\(^14,24,25\) In addition, many existing standardized and ad hoc schemes use patient- or admission-centric records that combine patient, protocol, and result descriptors. Mapping from such clinical schemes to a data-centric wrapper will thus require the development of specialized interfaces. However, the scope of our top-level elements and these preliminary tests of mapping ability suggest that the CDM could be used to link clinical as well as experimental data.

In this context, we emphasize that CDM is designed for the exchange of selected data with sufficient metadata to place the data in an experimental context, rather than for data mining from such sources as the electronic patient record. Consequently, its focus differs from that of clinical systems designed to encode a particular knowledge base in broad scope and fine detail. Our choice of vocabulary values and schema reflects this role and explains in part why we have not been able to build on the extensive development of medical informatic vocabularies and clinical data exchange methodologies.\(^14,15,24–28\)

**Discussion**

**Interoperability Standards for Scientific Data Exchange May Make the Network an Extension of the Laboratory**

The Internet has already transformed information exchange in many areas, but not all fields of biology have benefited equally. If permissive rather than restrictive interoperability standards are developed, neuroinformatics can enhance neuroscience in the same way that genetic and genomic informatics has enhanced genetics.

Projected examples can be drawn from neurophysiology. Laboratory data acquisition and analysis of time series were for many years the province of the oscilloscope or chart recorder and paper records derived from them. Data analysis, either directly from signal conditioners in real time or via the recorded intermediary of analog tape, was performed on data acquired locally. Currently, the oscilloscope has been supplanted by the desktop workstation, which often combines acquisition and analysis and allows data storage as well. But the workstation also has a network port, enabling analysis of remote as well as local data and comparison of local data with network data. To enable this extension of the data model, standards are needed to allow network data to be examined and analyzed using the same tools and methods that are used for local data, and other standards are needed to specify and classify network data and metadata that make public enabling values for the remote data “notebook.”
Multiple Dimensions of Interoperability

Figure 9 schematically represents interoperability of databases as a multidimensional space in which each axis defines enabling or restricting aspects of data exchange. Along each axis, interoperability is enhanced by movement away from the origin.

User interoperability measures the ease with which users interact with a database. The initial data model, database, and tools were designed to advance user interoperability, facilitating access to an intuitive data model embodied in open tools that aid data exchange in two domains of neurophysiology. This design included a technological component promoting multiplatform availability via standard Internet protocols, ideally requiring no additional software. User interoperability also includes open access, transcending hardware and software compatibility to encompass usability that derives from conforming to the standards, practices, and common core of knowledge of a focused domain of science.

Beyond user interoperability, the technical, data, domain, and temporal dimensions illuminate database-to-database interoperability and emphasize the need for open standards for data and data model exchange. The technical dimension expresses hardware and software level standards ranging from restricted or proprietary requirements to those enabling open interchange between resources for exchanging data and data models. Some aspects of these are relatively easy to implement, requiring only designs that are not restricted to a single hardware platform, operating system, or software package and embracing open architecture instead. More difficult are compatible standards for data format description and exchange of data and data models. These should include common but preferably not proprietary methods, such as the CDM, that describe queries and data formats and identify and resolve commonalities in data models.

Two more axes are needed for domain and data interoperability, and these are not orthogonal. To achieve interoperability between resources specializing in providing different domains of a discipline, the underlying models need to be linkable. This requires that metadata scope not be restricted to the immediate focus of the domain. A goal of data interoperability is to acquire and present data from different sources, including remote laboratories and distinct techniques. This permits comparisons that transcend differences in source or technique, by representing commonalities, such as a hypothesis, location of origin (cell type or region), or experimental conditions. Data types that are not directly interconvertible impose as well a requirement for interconvertible metadata to specify the data context. It is worth emphasizing that interoperability does not require identical data or data models but does require relatable data and compatible data models.

Migration

Finally, there is also a temporal dimension to interoperability. One direction of the temporal axis specifies migration—technology advances and even architecture changes. As database architectures change, the data model must be expressible and the data representable in the evolving architecture. Standards that are, ideally, persistent should be separate from their implementation, which will evolve with computer technology and experimental methodology. Interoperability must recognize not only the diversity of contemporary computer platforms, software, and standards but also this inevitable migration. Temporal interoperability may be maximized by basing solu-
tions on methodologies that are thought likely to change more slowly, rather than quickly. As described above, to implement data and data model interfaces, we selected XML for its independence from any technology, corporate sponsor, platform, or software implementation, which may be indicative of future stability.

Internal representations and organization of data sets may differ locally and evolve with changing technology and local needs, but external representations should be standardized. Such standardization might be designed to specify, for example, how data and metadata are represented in a common format, but a more general approach would be to specify a meta-representation for describing the format of the data set, so that migration of implementation would require only re-specification of the new format in the continuing meta-representation. Databases themselves may thus serve as buffers against migrations in technology that would otherwise make operating system or application-dependent formats or tags obsolete.

Interoperability should make provision for legacy data—at least from the era of digital data acquisition—as well as for current and future data sets.

The Common Data Model Advances the Identified Dimensions of Interoperability

The CDM described in this report promotes technical, domain, and data interoperability via extensible standards to describe data and data models for neuroscience information resources. Clearly, standards exist for specific instruments and software for data acquisition or analysis, and there is no technological need for format conversion where these are identical. Even in these cases, however, specific recording parameters, methodologies, and metadata may differ, as may the subjects, the preparations used, and the hypotheses tested.

A scheme as comprehensive as the CDM would not be required to empower one-to-one collaborations between pairs of laboratories willing to invest the resources to manually rationalize data models and define conversion methods. Via the data descriptor meta-language BDML, the CDM is designed instead to introduce methods to enable data and data model exchange and to make explicit the commonalities of disparate data utilization schemes. This involves both new methods (such as controlled vocabulary and broadly applicable internal data formats) and recognition of inherent commonalities (data models, techniques, and metadata).

The CDM and its XML representation are open, general, extensible, and largely shielded from technological obsolescence. The CDM makes explicit structure-function relationships, facilitating cross-modality and cross-species comparisons.

Since this work is a component of the Human Brain Project, the goals of generality, expandability and, most important, persistence guided our selection of a text-based meta-representation for a major component of our model for interoperability. The CDM, its XML representation, and allied methods are designed to mediate not only data exchange between different laboratories within one domain of neuroscience but also between different domains. Preliminary tests suggest at least some utility of the five superclasses for such additional domains of neuroscience. Interoperability should also be interdisciplinary, and additional tests are needed to determine how well the CDM encompasses links to and federation with sequence, structural, and other data.

The authors thank the many consultants, collaborators, and correspondents within and outside the Human Brain Project who have evaluated previous versions of the CDM and have helped us evaluate its applicability to a range of domains and projects.

References

10. Pechura CF, Martin JB (eds) for Institute of Medicine (U.S.) Committee on a National Neural Circuitry Database. Mapping the Brain and Its Functions: Integrating Enabling


