University of Texas at El Paso

From the SelectedWorks of Arshad M. Khan, Ph.D.

2003

Tools and approaches for the construction of knowledge models from the neuroscientific literature

Gully APC Burns, *University of Southern California*
Arshad M. Khan
Shahram Ghandeharizadeh, *University of Southern California*
Mark O'Neill
Yi-Shin Chen, *University of Southern California*

Available at: https://works.bepress.com/arshad_m_khan/19/

Neuroinformatics
Copyright © Humana Press Inc.
All rights of any nature whatsoever are reserved.
ISSN 1539-2791/03/01:81–110/$20.00

Original Article

Tools and Approaches for the Construction of Knowledge Models from the Neuroscientific Literature

Gully A. P. C. Burns,[*,1] *Arshad M. Khan,*[1] *Shahram Ghandeharizadeh,*[2] *Mark A. O'Neill,*[3] *and Yi-Shin Chen*[2]

[1] The K-Mechanics Research Group, 3641 Watt Way, Hedco Neuroscience Building, University of Southern California, Los Angeles, CA 90089-2520; [2] Henry Salvatori Computer Science Center, University of Southern California, Los Angeles, CA 90089-0781; [3] The Bee Systematics and Biology Unit, Oxford University Museum of Natural History, Parks Road, Oxford, UK

Abstract

Within this paper, we describe a neuroinformatics project (called "NeuroScholar," http://www.neuroscholar.org/) that enables researchers to examine, manage, manipulate, and use the information contained within the published neuroscientific literature. The project is built within a multi-level, multi-component framework constructed with the use of software engineering methods that themselves provide code-building functionality for neuroinformaticians. We describe the different software layers of the system. First, we present a hypothetical usage scenario illustrating how NeuroScholar permits users to address large-scale questions in a way that would otherwise be impossible. We do this by applying NeuroScholar to a "real-world" neuroscience question: How is stress-related information processed in the brain? We then explain how the overall design of NeuroScholar enables the system to work and illustrate different components of the user interface. We then describe the knowledge management strategy we use to store interpretations. Finally, we describe the software engineering framework we have devised (called the "View-Primitive-Data Model framework," [VPDMf]) to provide an open-source, accelerated software development environment for the project. We believe that NeuroScholar will be useful to experimental neuroscientists by helping them interact with the primary neuroscientific literature in a meaningful way, and to neuroinformaticians by providing them with useful, affordable software engineering tools.

Index Entries: Neuroinformatics; literature; knowledge models; bibliographic; database.

*Author to whom correspondence and reprint requests should be sent. E-mail: gully@usc.edu.

"Even the most active neuroscientist spends more working hours in reading, reviewing and writing scientific reports than on direct experimental effort"

Floyd Bloom, 1978,
Trends in Neuroscience 1(1):1

"There have been few changes to the traditional methods of neuroscientific information gathering, sharing and analyzing: namely reading research journals and traveling to scientific meetings"

Floyd Bloom, 1995,
Trends in Neuroscience 18(2):48-49

Introduction

A prominent and compelling justification for the development of neuroinformatics-based approaches is that our subject is extremely complex. There are currently over 50,000 working neuroscientists in the world producing enough data to fill more than 300 journals in a wide variety of subdisciplines ranging from psychology, linguistics, and animal behavior to neuroanatomy, electrophysiology and molecular biology (Chicurel, 2000).

A widely shared perspective within the field is that the most effective approach to information management is to build a large-scale collaborative network of seamlessly integrated repositories of raw data (Koslow, 2000). These repositories may be linked to the primary literature using modern web-based publishing technology to give the data structure and form. This approach has formed the basis of collaborative projects such as CERN (the European Laboratory for Particle Physics near Geneva), the Stanford Linear Accelerator (SLAC), and the Human Genome Project and has proven itself to be very powerful within other disciplines. However in neuroscience, this approach is hindered by the lack of consensus around the data's theoretical structure.

In some regards, the primary scientific literature forms the basis for all human understanding of a subject; it is the crucible where scientific discoveries are validated, tested, confirmed, or rejected. We assert that the theoretical structure embodying the subject is an emergent property of the observations, interpretations, arguments, or hypotheses contained within the literature's constitutive publications (Burns, 2001a). Furthermore, the literature's large size and scope, lack of standardization, variable quality, and largely linguistic (i.e., qualitative, nonmathematical) nature mean that these emergent theories are often computationally unwieldy.

It is for this reason that we directly focus on representing and analyzing the contents of the literature with knowledge management techniques. If we consider "data" to be unattached, unstructured values; pieces of "information" are then data with additional structure, and explanation; and "knowledge" would be defined as information that is considered *in the context* of other information (Blum, 1986). While technology has developed to accelerate the delivery of published information to the modern scholar, few, if any, tools exist to expand our understanding of it. The transition from paper to electronic publishing extends the structure of journal articles by providing not only access to raw data, but also to embedded dynamic data viewers, and even computational modeling tools for readers to really explore the data underpinning a publication. *The functionality of NeuroScholar contrasts with this by permitting users to create models of their own knowledge across large numbers of such papers (whilst providing access to other knowledge models from other users)*. We expect that the emergent properties of such a system will provide a powerful new approach for generating neuroscientific theories.

Currently, the function of the NeuroScholar system is to address the following question: "What is the complete neuroanatomical circuitry underlying a specific physiological phenomenon?" In order to address this question

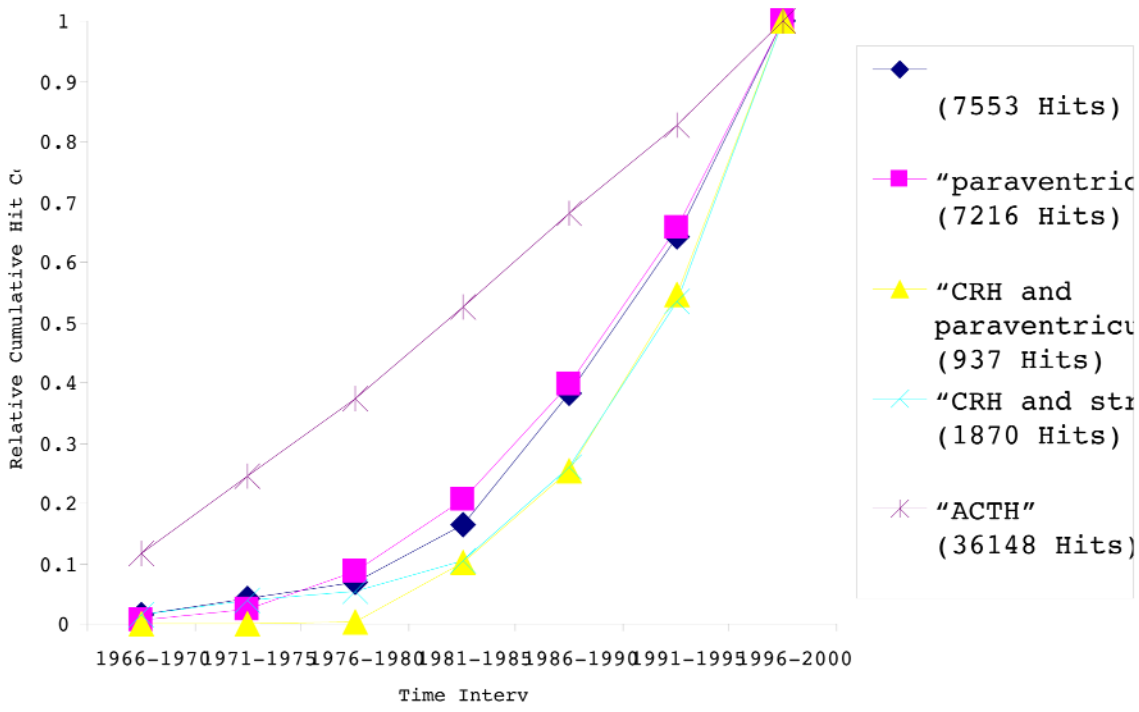

Fig. 1. Graph showing the normalized number of hits returned from searches of the PubMed literature database using specified search terms. **[CFO]**

effectively, we first distinguish the phenomenon of interest (say for example, the release of a hormone in response to stress), identify which regions of brain tissue are involved in this phenomenon, and finally, study all the neuroanatomical connections linking these regions. In principle, the NeuroScholar system may be used for any species, as long as a complementary electronic neuroanatomical atlas is available for use by the system. At present, we only support data with a neuroanatomical atlas of the rat (Swanson, 1998).

NeuroScholar's utility can be emphasized by considering the size of the task of building a useful representation of a large literature. Since our stated example focuses on the stress response and Corticotropin Releasing Hormone (CRH; *see* Table 1), we performed some broad searches on the National Library of Medicine's PubMed website (accessible from http://www.ncbi.nlm.nih.gov/) to estimate the size of the literature from the following keyword combinations: "CRH," "paraventricular," "CRH and paraventricular," "CRH and stress," and "ACTH"(*see* Table 1). As shown in Figure 1, the number of publications conforming to the search ranged from almost one thousand (for the keywords "CRH and paraventricular") to over 30,000 (for the keyword "ACTH"). Clearly, these global searches are prohibitively large for an individual scientist to manipulate. In most cases, the publication rate on the specified subject is increasing. As described in the next section, it is impossible for an unaided worker to address questions across the whole literature. This only becomes possible within the type of neuroinformatics framework we describe here.

Within this article, we describe a set of modular neuroinformatics tools that combine to form a prototypical application (called NeuroScholar) that is designed to act as a knowledge management system for the neuroscientific literature. We have previously described NeuroScholar's underlying strategy and theoretical basis (Burns, 2001a), and its system's design in detail (Burns, 2001b).

We place potential users of a knowledge management system of the literature on a continuum. At one end, *experimental specialists* focus on their own personal perspective of the literature. At the other, *neuroinformaticians* may concentrate on using individual components of our system to strengthen their own software development work. Our framework supports the entire continuum of users. In this paper, we begin by describing the high-level functionality of the overall system, and then provide examples of the specific modular tools that we are implementing. NeuroScholar provides a suite of tools for *experimental specialists* to build and use models of their own knowledge, and for *neuroinformaticians* to utilize the functionality of our system within their own. We also provide software engineering tools for *neuroinformaticians* that assist their development work.

The Utility of NeuroScholar for the Experimentalist: A Hypothetical Example Concerning the Study of the Stress Response

Large-scale questions such as "How is stress-related information processed in the brain?" are very difficult to answer. As noted earlier, the volume and complexity of the information is too large for an individual to process effectively. Here, we attempt to provide a "real-world" example of how NeuroScholar can aid the experimental specialist in addressing such issues, using the question posed above as our example.

We first define the problem by describing some questions currently being asked about it. We then introduce how NeuroScholar helps address these issues. One point to remember is that NeuroScholar is not providing solutions to the questions being asked per se, but is providing useful tools to help the experimental research community answer these questions. The promise of this work is that individual users will be able to collectively build large-scale *knowledge models* of, for example, the complete neural circuit of the system involved in the stress response. *In this way, the utility of NeuroScholar is that it helps experimentalists answer questions by helping them first ask the questions in a meaningful way.*

Defining the Question

We introduce the problem with neuroscientific definitions from within this specialized field of study. We then examine some large-scale questions currently being asked by experimental specialists within the field and how NeuroScholar aids them in this process.

What is stress?

It is important to first define what we mean by stress (which is not, in itself, universally agreed upon within the field). Many physiologists instead try to understand stress by examining the stimuli that elicit stress ("stressors"; Selye, 1974). Stressors may include those that disturb the homeostatic mechanisms of the body (e.g., dehydration, infection, hemorrhage), as well as those that threaten an individual's state in less clear-cut ways (e.g., restraint, footshock).

The hypothalamic paraventricular nucleus (PVH) is considered to be the final output pathway in the stress response. The PVH is defined as a pair of densely packed wing-shaped nerve cell clusters occupying a small, dorsomedially located volume of the hypothalamus. This is a critical staging area for integrated, adaptive responses to stress (Swanson,

Table 1—A Guide to Terminology and Abbreviations Used Within This Paper

Abbreviation / Terminology	Definition
'Data'	Unstructured measurements
'Information'	Data with structure and meaning
'Knowledge'	Information in the context of other information
'Knowledge Model'	A computational representation of an individual user's perspective of information from the literature defined in context of either information from another source or from fragments in the literature
'Fragments'	Individual excerpts of data, taken from published sources (i.e., journal articles, books, databases, etc).
'Ontology'	"An explicit formal specification of how to represent the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them" (Howe 2001)
'Experimental Specialist'	End-users of the NeuroScholar system who are principally concerned with using the system to build knowledge models
'Neuroinformatician'	Developers within the field of Neuroinformatics who may want to use the underlying infrastructure of the VPDMf and the KMC within their own systems
KMC	The general 'Knowledge Management Core' infrastructure upon which the NeuroScholar system is built
VPDMf	A software engineering paradigm called the 'View Primitive Data Model framework' that permits forward and reverse engineering based on the Unified Modeling Language (the 'UML')
View	A hybrid composite data object defined within the VPDMf from one or more linked classes within a data model
View Instance	Instance data contained within the structure of a named View
View Specification	A description that names which classes from a data model from a View and how they are associated (used within the VPDMf)
View Graph Definition	A graph-based representation of the interactions between several views derived from the same data model within the VPDMf
View Graph Instances	A graph-based representation of interrelated View Instances
UML	The Universal Modeling Language, an widely-used object-oriented design methodology
XML	The eXtensible Markup Language.
ORT	The Objective Relational Transformation methodology for translating data between parcellation schemes (used in the CoCoMac system).
PVH	The paraventricular nucleus of the hypothalamus
PVHmpd	The medial parvocellular division of the paraventricular nucleus of the hypothalamus
'HPA Axis'	The pathway of activation from cells in the PVH, to cells in the anterior pituitary gland to cells in the adrenal cortex cells implicated in the stress response
Stressor	Stimuli that elicit stress
CRH	Corticotropin releasing hormone
CORT	Corticosterone
ACTH	Adrenocorticotropic hormone

1986; Sawchenko and Swanson, 1989; Swanson, 1991; Herman and Cullinan, 1997; Sawchenko et al., 2000). Activation of this region ultimately results in the release of pituitary and adrenal hormones in the bloodstream, which can then exert a variety of effects on both central and peripheral target tissues.

The PVH consists of many distinct subgroups of cells (Swanson, 1991), only one of which will be discussed here. Specifically, neurons within the medial parvocellular division of the PVH (PVHmpd) respond to stress-related inputs by synthesizing the hormone corticotropin-releasing hormone (CRH) and releasing it into the bloodstream. Once released, CRH activates the anterior pituitary gland, causing it to release the adrenocorticotropic hormone (ACTH), which in turn, travels down to activate the adrenal cortex, causing it to release glucocorticoids, such as corticosterone (CORT). CORT then acts upon multiple tissues both centrally and peripherally to mobilize the body's energy stores in response to stress.

This pathway of activation, from PVH cells to anterior pituitary cells to adrenal cortex cells, is an example of the "hypothalamo-pituitary-adrenal" axis (de Groot and Harris, 1950). The output of the HPA axis, for example the CRH-ACTH-CORT response, is a major indicator that physiological responses to stress have been triggered. It is no surprise then, that the synthesis and release of CRH, ACTH, and CORT, in the HPA axis are under exquisite control at a number of levels (e.g., Axelrod and Reisine 1984; Dallman et al., 1987; Watts and Swanson, 1989; Tanimura et al., 1998; Tanimura and Watts, 1998; Sapolsky et al., 2000). Activation of the HPA axis is generally considered to be the hallmark of the stress response, with the PVH serving as "the final common pathway for all types of stress response mediated by the central nervous system" (Swanson, 2000).

From this, we emphasize the following points:

a) It is useful to study stress in terms of its physiological causes (stressors).
b) A stressor can preferentially activate certain brain regions to produce a characteristic "pattern of activation."
c) The PVH is a critical brain region involved in the brain's response to stress and is composed of many subgroups of cells.
d) One PVH subgroup, the PVHmpd, responds to stress-related signals by synthesizing CRH. CRH can then trigger a cascade of hormone release, first involving ACTH from the pituitary, and then CORT from the adrenal cortex.
e) The activation of PVH cell groups (such as the PVHmpd), which often results in the production and release of hormones in the bloodstream, is a hallmark of the brain's response to stress.

We now address some likely questions posed by experimental specialists.

Questions of Interest for Experimental Specialists

How does the brain discriminate between stressors? As noted earlier, it is generally thought that part of the way that the brain can differentiate between different types of stressor is from the selective activation of the neural circuitry that processes the information encoding each type. These "patterns of activation" differ for each type of stressor. A conceptual challenge facing experimentalists is to precisely document the regional and connectivity relationships for the brain subjected to different types of stress.

Can stressors act similarly on the brain? A related question is whether the patterns of activation caused by one stressor are always mutually exclusive from the patterns elicited by another. It is now well known that some brain regions are activated by multiple types

of stressor, while others only respond to stressors of a specific type.

How does a single cell population such as the PVH integrate the various signals encoding information about different stressors? As previously noted, multiple stressors may act, in part, upon common sets of target areas in the brain. For these common areas, some mechanisms must exist that allow for signals encoding information from multiple stressors to somehow be integrated at the level of the single cell or cell population. How is this achieved? For example, PVH neurons become activated in response to cold stress as well as metabolic stress, such as hyperglycemia (excess blood glucose). How do these neurons act when both stressors occur at the same time (i.e., when signals conveying both types of stressor arrive at a PVH neuron essentially simultaneously)? Does the hormone release that typically forms the output response of PVHmpd cells increase in amplitude or frequency in any way?

Addressing these questions requires a seamless integration of relevant data from a variety of primary literature sources within a coherent, shared structure to which multiple users can contribute. We now describe in detail ways in which the NeuroScholar system can contribute to helping solve this problem.

The Usefulness of NeuroScholar Knowledge Models Concerning the Brain's Response to Stress

Organizing the Primary Literature

Perhaps the most logical starting point for any experimentalist attempting to make sense of data within this subject is having a means to organize the primary literature in a useful way. Traditionally, experimental specialists have used reference management programs to help create databases of the literature in which they are most interested. This method continues to be a useful, albeit simple, way to organize publications. However, few tools have been developed that allow the user to take only relevant portions of the actual published literature. Such "parts of papers," or *fragments* (*see* "The Neuroscholar System as a Whole") may include data tables, photomicrographs, electrophysiological traces, and, in the case of papers published electronically, even supplemental data (including published sets of raw data) or digital animation. It would be useful to store such fragments within a convenient environment that allows the user to make sense of them.

NeuroScholar's user interface (the components of which are described in "NeuroScholar Components") provides such an environment. As shown in Figure 2, the user interface provides a basic "workspace" within which a user may link multiple fragments of information. Fragments from this paper can then be linked to the document, including textual and graphical fragments. NeuroScholar contains tools to create fragments efficiently and store them in a database (*see* "Fragmenter") for subsequent retrieval and viewing within a user's workspace environment. Figure 3, for example, illustrates how we may delineate Figure 4 from Rho and Swanson (1989) as a graphical fragment. With this tool, the user may use the fragment as the basis for definitions and assertions within his or her knowledge model. The utility provided by the Fragmenter is to represent portions of the actual contents of the papers themselves and is clearly not commonly available within conventional bibliographic software.

Through the use of fragments, the literature required to answer the questions posed in the preceding section can be readily broken down and organized into fragments that may then be associated with other information according to the preferences of the individual user. For example, electrophysiologists may collect only fragments of data represented by stimu-

lus and recording traces (or text fragments describing such graphical data), whereas anatomists may only collect fragments depicting and/or textually describing structure-function relationships within stress-related circuitry. This latter task is made easier by other component tools of NeuroScholar, as described next.

Patterns of Physiological Intervention and Activation Related to Stress

As described in "Questions of Interest for Experimental Specialists," one question posed by experimental specialists is how the brain discriminates between stressors. Stressors produce characteristic "patterns of activation." Many studies have described such patterns, which involve regions both common and unique to multiple stressors. Moreover, while two stressors may exert their actions upon a common brain region, a single stressor can also have opposite effects on two different regions. For example, CRH synthesis in two brain regions, the PVH and the central nucleus of the amygdala (CeA), are markedly decreased and increased, respectively, in response to elevated glucocorticoids (Watts, 1996; Makino et al., 2002). Also, a single stressor can activate more than one subgroup of a brain region. For example, metabolic stress can increase the activation of enzymes in multiple divisions of the PVH (A.M. Khan, unpublished observations), each of which may be mediating different aspects of the response to this stressor.

The sample questions described above are typical of an individualized enquiry conducted by a single researcher and, as such, may be of little or no interest for other scientists outside or even within the same field. It is important to state that these questions are all interlinked, since they all involve shared concepts; namely the pattern of activity within the brain under a specific physiological condition and the structure of the neuroanatomical circuit serving as a physical substrate for that activation. Thus, the specialized knowledge models that individual researchers use to answer specific questions are based on components that will be naturally useful to other researchers.

NeuroScholar helps address these issues by providing users with tools that can help them delineate "brain volumes" according to their own choosing and represent these volumes with reference to a brain atlas. As shown in Figure 4, NeuroScholar's AtlasMapper plug-in (described in detail in "The Atlas Mapper") can allow a user to delineate an enclosed volume on a template of the brain region of interest, obtained from an electronic atlas file. Figure 4 specifically depicts a brain volume delineated by a user and superimposed upon multiple subdivisions of the PVH. This, for example, might be useful if one wishes to note a pattern of stress-induced cellular activation within neuronal populations that do not necessarily conform to the boundaries of published atlases. This method of delineation may be used to describe lesion sites, injection sites, regions of labeling, sites of activation, or any other data located physically within the brain.

It should also be mentioned that the literature reporting such data (i.e., activation patterns from different stress patterns) is growing rapidly, and this method immediately places each newly published set of results within a framework that straightforwardly permits different data to be compared. Tools such as those provided by NeuroScholar will be indispensable for users who need to sort through such large numbers of papers to begin formulating general hypotheses about the data at a systems level.

Integration of Stress-Related Information at the Level of the PVHmpd

The Fragmenter and AtlasMapper plug-in can also be used to organize information concerning the many inputs arriving at PVH neuroendocrine neurons mediating the final out-

put pathway of the stress response. The principal question asked in the preceding section in relation to these inputs was how signals conveying stress are received and integrated by the cells within the PVHmpd. This question has received a fair amount of attention in the literature (*see* reviews by Swanson, 1986, Swanson et al., 1987, Herman and Cullinan, 1997, Sawchenko et al., 2000) and remains an active area of experimental investigation.

Roughly fifty different sources of neural inputs to the PVH have been identified using axonal transport tracing methods, making the afferent control of PVH function extraordinarily complex (Swanson, 2000). PVHmpd neurons are known to receive four major sources of neural input (reviewed by Swanson, 1986): (1) catecholaminergic inputs from the brainstem that convey primarily viscerosensory information; (2) subfornical inputs, some of which contain angiotensin II as a transmitter; (3) inputs from the bed nucleus of the stria terminalis (BST), a limbic region that is believed to be a principal conduit of information arriving to the PVH from the neocortex; and (4) a variety of inputs from the hypothalamus itself. Without computational support, making sense of the data describing these regions and their inputs to PVH would be highly challenging. Within the NeuroScholar system, it becomes a matter of using the Fragmenter and Atlas Mapper tools to link the original figures and text of the input regions to delineations a standard atlas and then manage the accounts of connections with this neuroanatomical organization within the NeuroScholar user Interface.

NeuroScholar as an Aid for Designing Experiments for Stress-Related Research

In addition to providing experimental specialists with a systematic means to keep track of the primary literature, their own experimental results, and the relationship between these two sets of information, NeuroScholar also provides users with a tool to help design the experiments themselves. Experimental research plans that include iterative steps, in particular, can be readily outlined using NeuroScholar's experimental flowchart plug-in (*see* "The Experimental Flowchart" for more elaboration on this topic; *see* Fig. 5).

As NeuroScholar is still under development, efforts are ongoing to tailor the components of the system to the needs of the experimental specialist (as well as neuroinformatician). The preceding section provided a small sampling of what can potentially be achieved using the NeuroScholar system. We welcome the input of our colleagues who come across this article and find a specific need for the NeuroScholar system that has not been explicitly described here. Indeed, the promise of this system ultimately rests in its operation by as many users as possible.

Computational Implementation of the NeuroScholar System

The development of new approaches in the field of neuroinformatics is supported by best-practices and approaches from computer science. Three key computational concepts form the foundation for our work: the separation of processes, data schemata, and data instances, modularization (within designated namespaces), and the use of sound software engineering practices.

If processes are the mechanisms that implement a specific functionality, then the data schemata and instances provide the context for that functionality. By distinguishing between these three components explicitly, we greatly expand the applicability of our tools outside of the scope of this project. If we design the processes to work under data with different data schemata, then the context may be adapted for other tasks.

The systems we have built are modular and multi-level so that a subsystem provides a

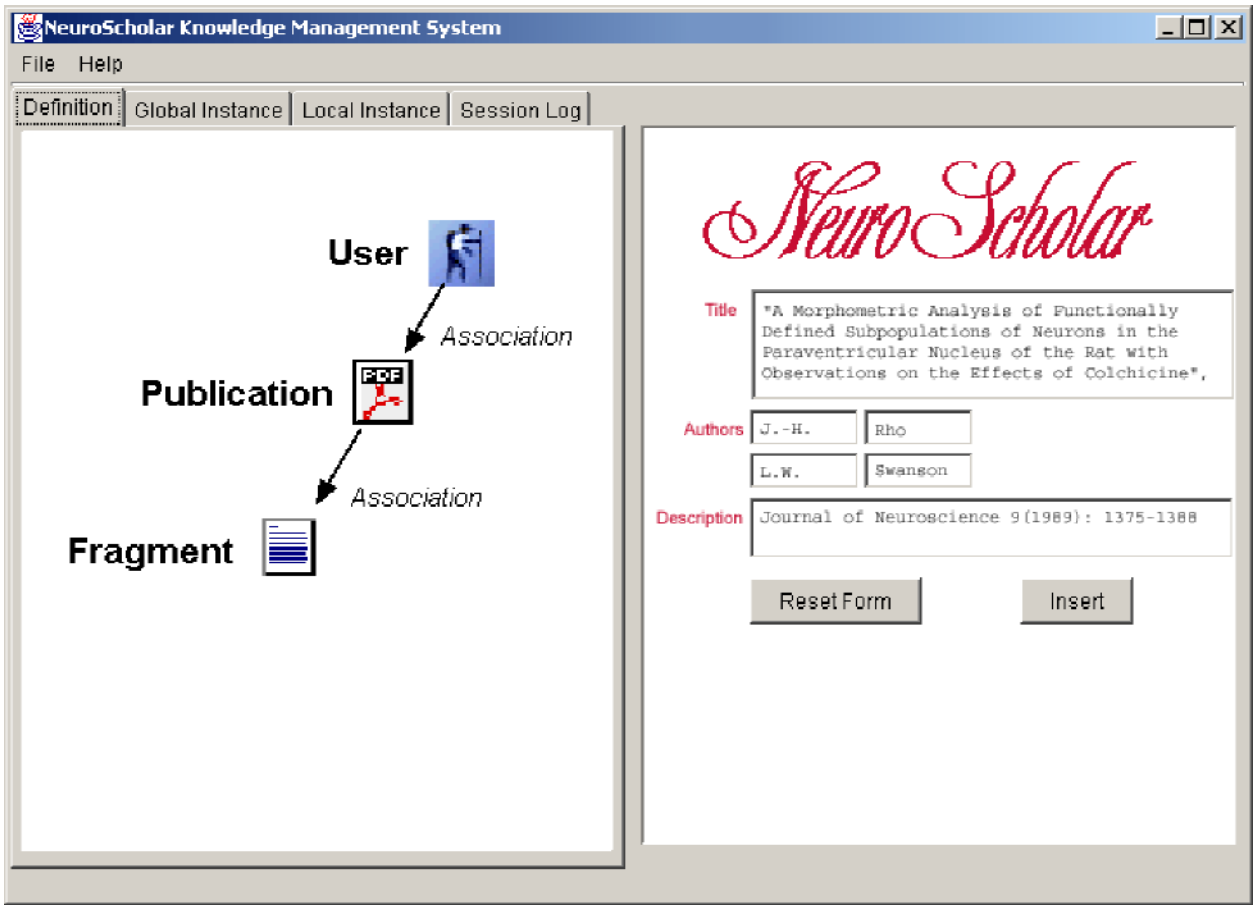

Fig. 2. A screen shot from the user interface of the NeuroScholar Knowledge Management system. **[CFO]**

specific functionality (either as a contributing web service or a plug-in component for the overall application). Web services exist where both software and data are provided as a computational service implemented as remote procedure calls. These then may act as the constitutive components of cooperative applications that may be dispersed across a network such as an intranet, the Internet, or both.

In this section, we describe some of the subsystems that combine to provide the functionality of NeuroScholar. Within this section, we describe this functionality in a top-down manner. We begin with components of NeuroScholar's user interface that permit users to manipulate specialized data types such as fragments from papers, delineated volumes of brain tissue, and representations of experimental design (*see* "The NeuroScholar System as a Whole" for the system as a whole and then "NeuroScholar Components" for each individual subsystem). The next subheading describes how we embed these data types into a framework that links scientific descriptions, interpretations, and rules to the primary literature and permits users to annotate and discuss the contents of the system (*see*the "Knowledge Management Core"). The lowest level of the system is called the View-Primitive Data Model framework (VPDMf), and provides a data-management methodology to support

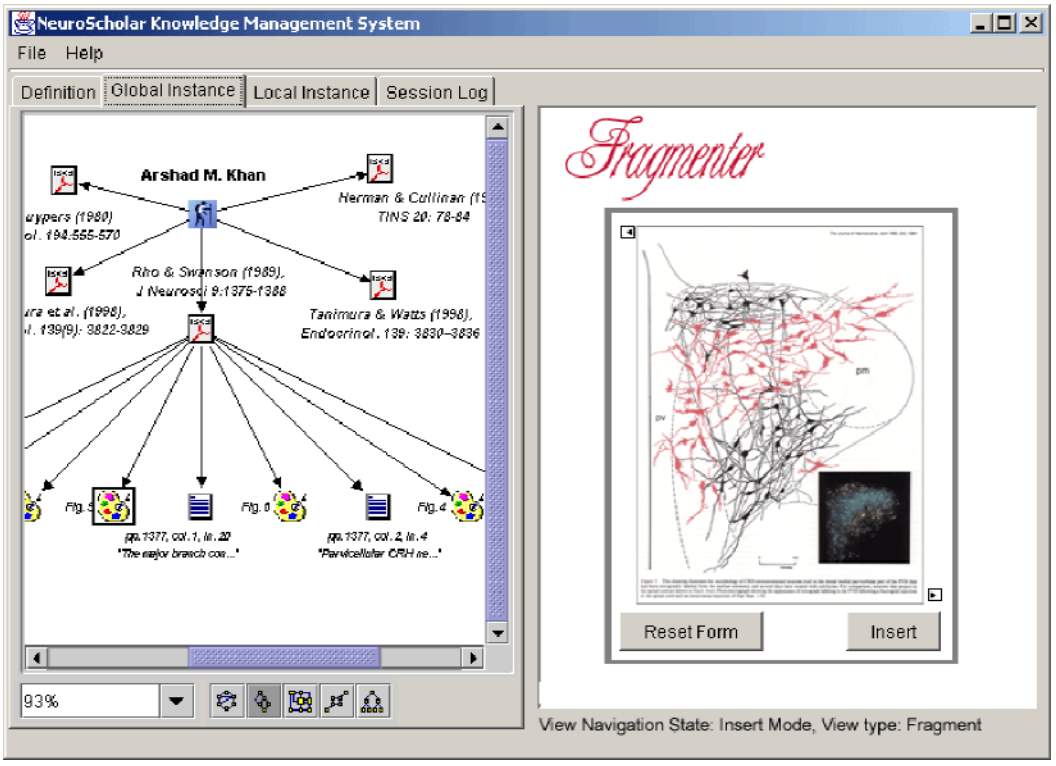

Fig. 3. Screenshot illustrating the Global View Graph Instance and the Fragmenter. **[CFO]**

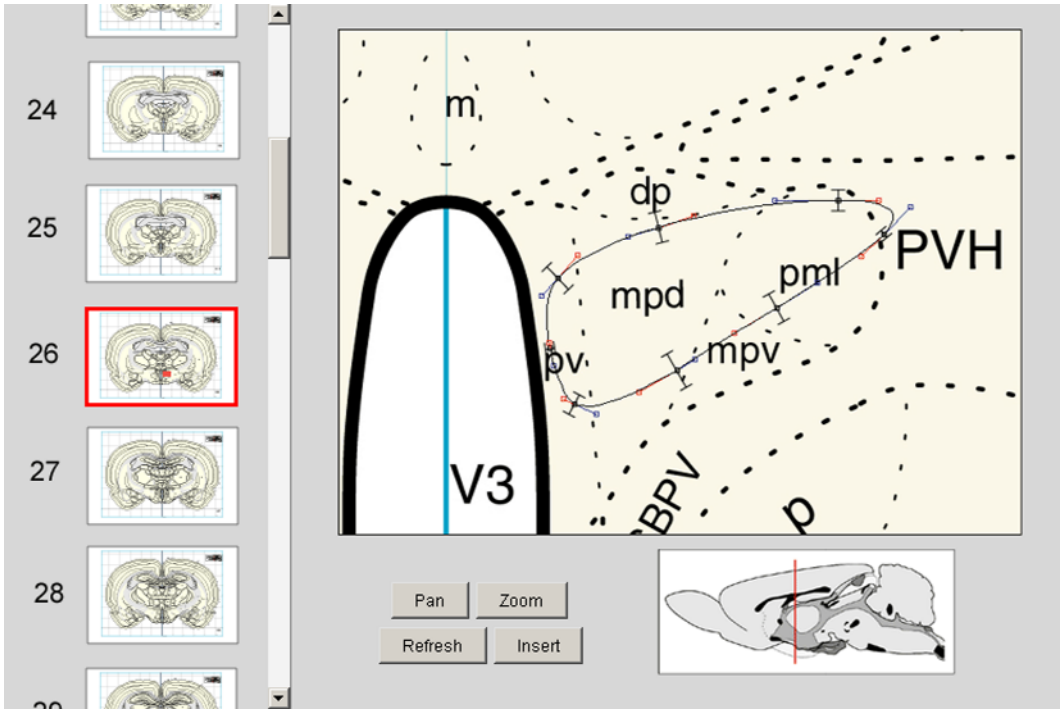

Fig. 4. Detail from the AtlasMapper plugin showing the delineation of labeled CRH cells shown in fragment screenshot in Figure 3. **[CFO]**

the other components (*see* "The View-Primitive-Data Model Framework"). It is important to note that while the different subsystems interoperate cooperatively within the framework of NeuroScholar, they may also function outside of that framework as independent tools.

All software developed within this project is open-source except where development work involves third-party commercial software (from http://www.thebrain.com and http://www.tomsawyer.com) where we do not publish our code. This is to protect our commercial partners' intellectual property. All available software is documented at http://www.neuroscholar.org/ and may be accessed by navigating from that site to http://www.sourceforge.net/.

The NeuroScholar System as a Whole

In this section we consider the system as a whole, rather than focus on an individual subset of the system's design or implementation. The way the NeuroScholar system works is illustrated in Figure 6. The top tier of the diagram represents the heterogeneous scattering of information that occurs in the world as "publications." At the present time, we only work with online papers that are expressed in the format known as portable document format (PDF) from Adobe (http://www.adobe.com/). We provide tools to extract textual "fragments" from a paper (*see* "Fragmenter") and save a pointer to the location of that fragment in the top level of the NeuroScholar system. Within NeuroScholar, each user has a "workspace" designated to them, where they may construct a computational representation of their knowledge. These "knowledge models" are considered the user's intellectual property and may be published within the system, or external to the system via NeuroScholar's web interface. In this section, we discuss the high level applications of interest to *experimental specialists* before entering into the low-level components of interest to *neuroinformaticians*.

The fragments serve as our image of the primary literature and as such they support the definition of knowledge models by the user. Every single entity within an individual's user space must be supported by links to fragments. It is possible to link knowledge models to entities from other users' spaces as well in order to build a consolidated view, but the entities themselves must be linked to the primary literature.

The symbols on the top tier of Figure 6 refer to non-pdf document types and databases. In future iterations of the NeuroScholar system, we will be able to treat *any* source of data on the web as a publication as long as we can navigate to intelligible fragments within it and have confidence that the fragments will persist in the same online location over time. This may include raw data from other neuro- or bioinformatics systems such as graphs, images, neuroimaging, or time-series data files. At present, we have deliberately restricted ourselves to peer-reviewed articles to ensure that the fragments being used in the system arise from reviewed sources. In order to extend this mechanism to non-reviewed sources of data, we permit users to attach their own personal "reliability score" to individual sources (based on structures from the Knowledge Management Core, *see* "Knowledge Management Core").

According to the definitions of the Unified Modeling Language ("the UML" Rational 1997), if an object is "an entity with a well-defined boundary and identity that encapsulates state (attribute and relationship values), and behavior (operations and methods)," then a class is a description of a set of objects that share the same attributes, operations, methods, relationships, and semantics.

We use a uniform approach to the different types of data being processed that uses and extends the basic object-oriented class struc-

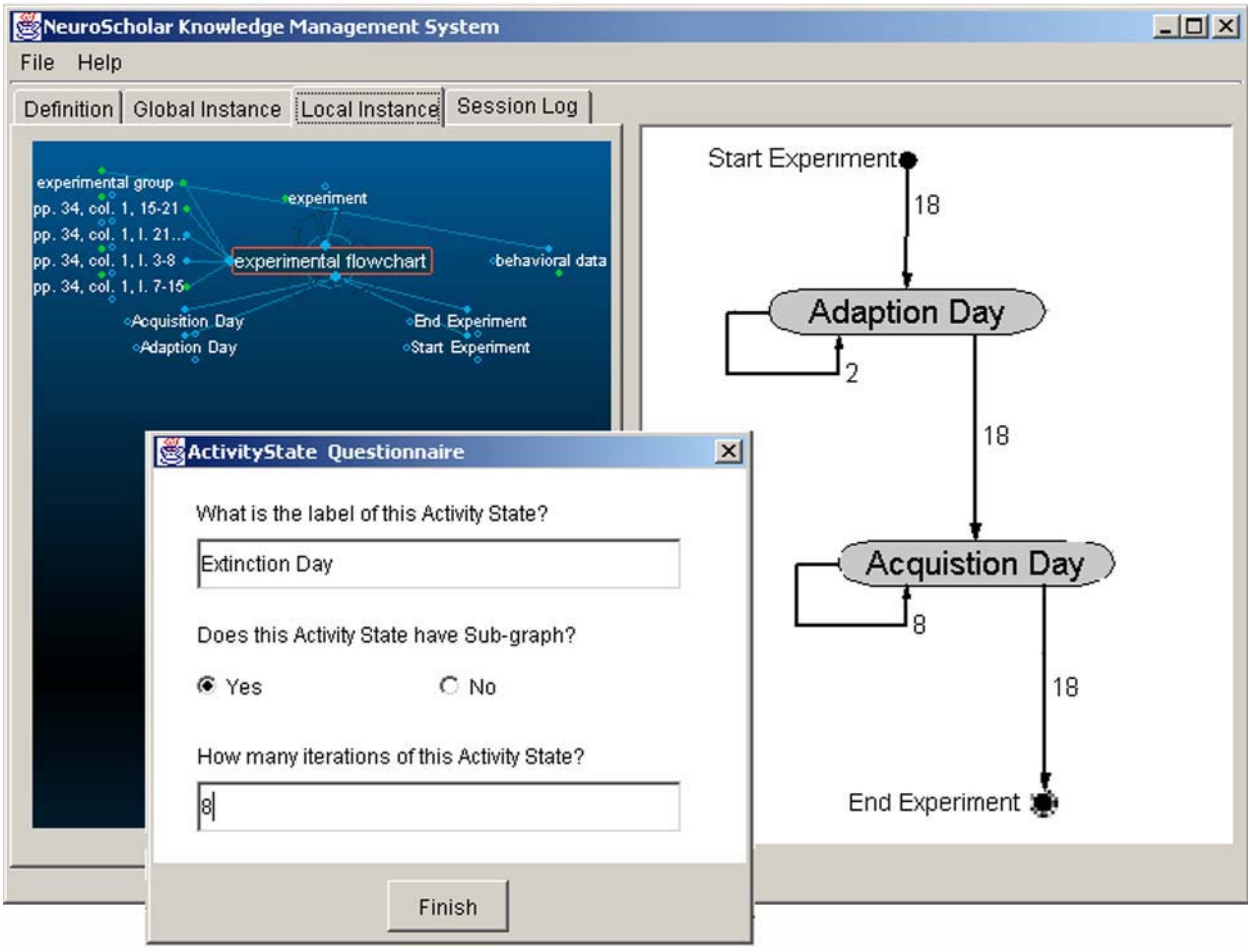

Fig. 5. The experimental flowchart control with the local View Graph Instance viewer from http://www.theBrain.com. The flowchart refers to the first ethological study of the classical conditioning in the rabbit eyeblink response (Gormezano, 1962; #606) where animals are "adapted" to their housings over two days, they acquire the conditioned response over 8 days and the response is made "extinct" over the final 8 days. **[CFO]**

ture of the UML. Each publication, fragment, entity, annotation, and rule (the different species of computational item defined within the knowledge management framework of NeuroScholar, *see* "Knowledge Management Core") is considered to simply be a "View" within the data-management framework software at the system's lowest level (*see* "The View-Primitive-Data Model Framework"). Each view is essentially a composite object (similar to the concept of "materialized views" in relational databases) made up of combinations of interlinked classes from the system's underlying data model. Each view may be represented as an encapsulated node in a so-called "View Graph" where the associations, overlapping relationships, and relative enclosure of different types of items can be represented as edges in the same graph (i.e., if View B's constituent classes were a subset of the classes of View A, then an edge from B to A would exist in the View).

A screenshot of the basic user interface of the NeuroScholar system is shown in Figure 2,

illustrating the View Graph Definition (a graph that shows the possible relationships between Views) and a simple form generated for data in a "Publication View." The definitions of views are related to each other by virtue of the fact that the set of classes in one view may be related to the set of classes in another; they may also be connected via association in the UML representation. These relationships may be calculated to form the basis of the definition of the View Graph (*see* Figure 7 for the underlying schema navigated in this process). This system utilizes View Graphs in three forms. Figure 2 shows a View Graph Definition with the possible links between the User, Publication, Graphical Fragment, and Textual Fragment Views derived from the schema shown in Figure 3. The left hand pane of Figure 3 shows the "Global View Graph Instance" which illustrates all the View Instances that are currently loaded and available to the user. Users may reformat and rearrange this graph by hand or by using layout tools. Thus, in Figure 2, we illustrate the organization of the software displaying a view. The left hand panel shows the relationships between three views defined over our simple example from Figures 2, 3, and 5. The views shown here are denoted by the User, Publication, Textual Fragment, and Graphical Fragment (each one derived from one or more linked classes from the schema shown in Figure 7).

Within the NeuroScholar system, we use commercial software packages to manipulate and navigate our graph representations. The graph-based representation shown in Figure 2 uses the commercial graph drawing software from Tom Sawyer Software (http://www.tomsawyer.com/). This software provides several functionality including automatic layout functions, subgraph representations (*see* "The Experimental Flowchart"), zoom, and graph editing.

NeuroScholar Components

Neuroscholar's primary goal is geared to cater to the *experimental specialist*. We wish to provide computational tools for scientists who otherwise would not consider using approaches from modeling or computational neuroscience. We describe some of the user interface methodologies that were built on top of the data model described previously (Burns, 2001b). These user-interface methodologies deal with some of the issues that neuroscientists are forced to consider almost every time they read a paper: these include selecting the excerpts of online papers that form fragments within the system ("Fragmenter"), tools to assist the delineation of structures on a brain atlas in a way that displays the users' uncertainty concerning the delineation ("The Atlas Server" and "The Atlas Mapper"), and graphically describing the design of an experiment ("The Experimental Flowchart"). The development of these tools was designed to make the process of interacting with the literature as effortless as possible for the user whilst empowering users to disseminate knowledge.

Fragmenter

All interpretations in this system are based on excerpts from the primary literature itself. While building a system to represent work describing the neural connectivity of the rat, we discovered that copying the text of each excerpt into the database was the rate-determining step of data entry (Burns, 1997; Burns and Young, 2000). In this application, we sought to simplify this process based on neuroinformatics techniques of annotating passages of text (Ovsiannikov and Arbib, 2001). We simply wanted to be able to select excerpts from a journal article, store them in our system, and then manipulate them as we would any other view within the system as a whole.

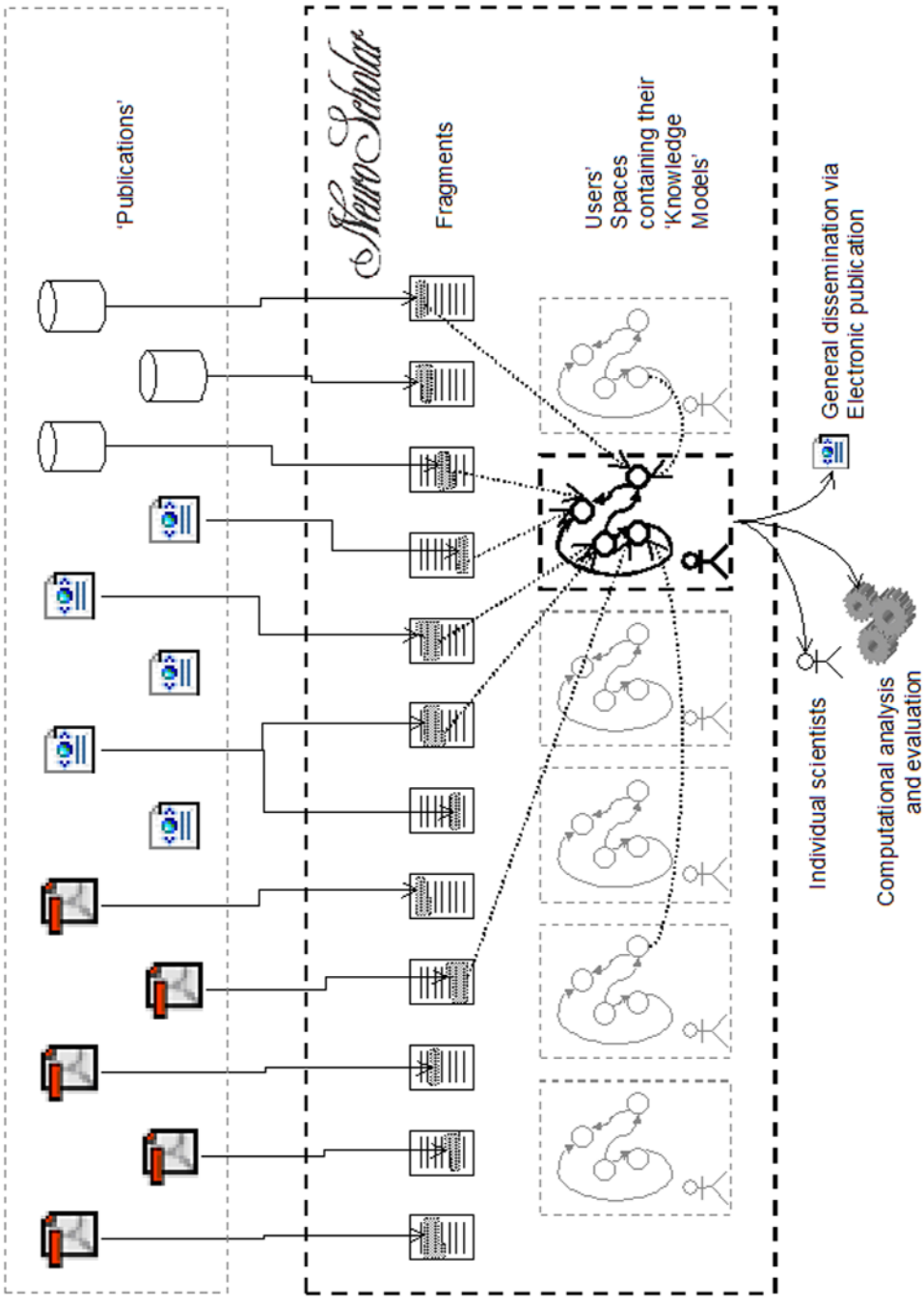

Fig. 6. The high level structure of the NeuroScholar system. **[CFO]**

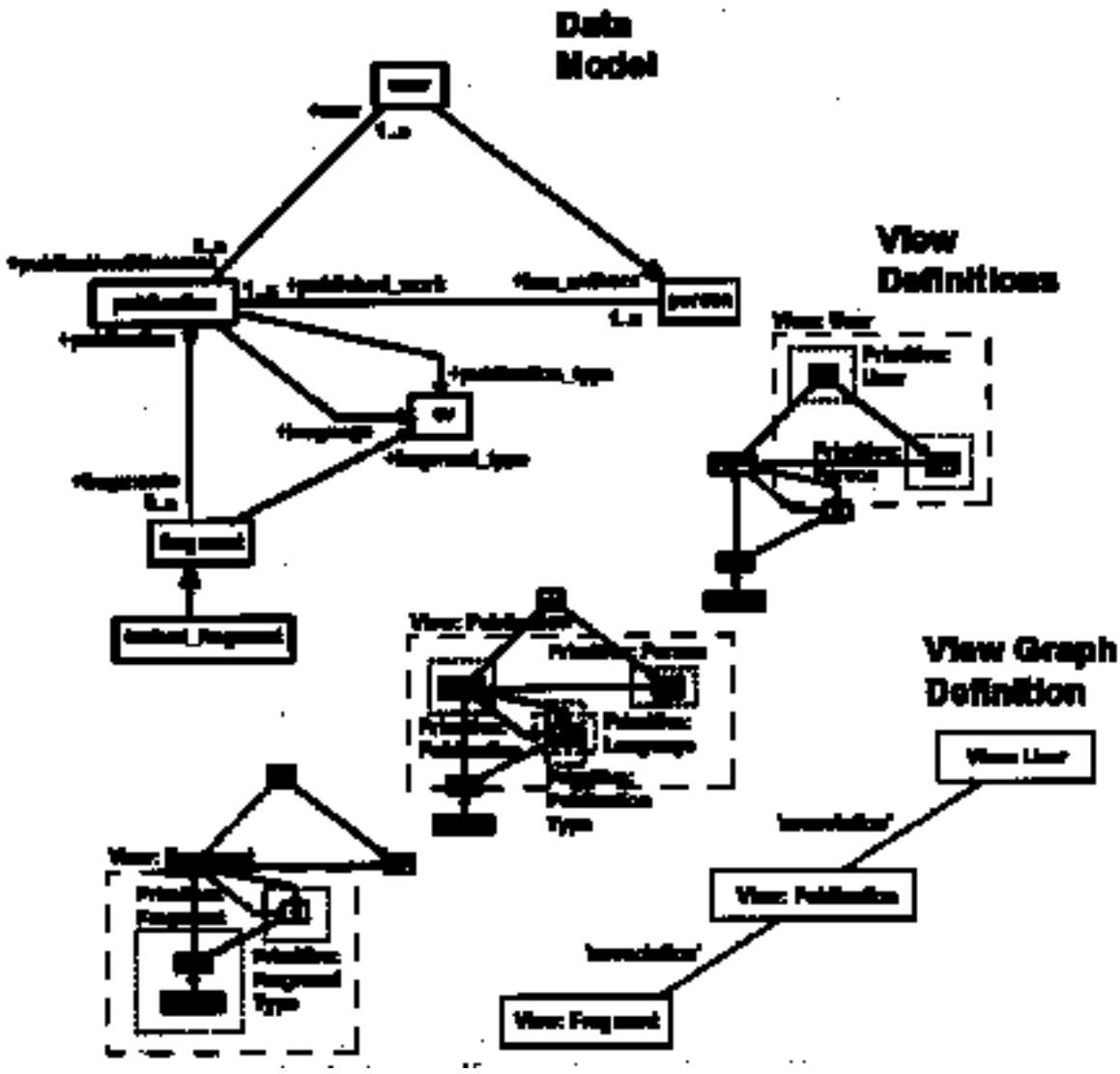

Fig. 7. The relationship of the data model, Views, and View Graph for a simple example. **[CFO]**

The interface is referred to as the "Fragmenter" and is illustrated below.

The Fragmenter serves as a form for textual and graphical fragments for online documents in pdf format. There are some copyright issues that we have to circumnavigate; in most cases we would not be permitted to store a reproduction of the text and figures of the document in our document without permission. We address this issue by only storing the location of the delineation around the relevant text or figure on the page, the page number, and a unique index citation to the article that points the user directly to the article (such as the PMID identifier on the PubMed system). This means that users must have *external access* to the publication from which they wish to extract fragments (as files on their local

machine, for example). If permission is given by the publisher, users may be able to access online articles directly. The right hand pane in Figure 3 shows the Fragmenter in use, illustrating a graphical fragment View Instance (Rho and Swanson, 1989).

It is hoped that the use of the Fragmenter will have a significant impact on the way that scientists perform their research. The act of referring to the original fragment from the original publication necessitates a rigorous approach to the literature to minimize misinterpretation and miscommunication. Not only will users be able to track what precise fragments support or refute their interpretations, authors will be able to track exactly how people are interpreting their work.

The Atlas Server

Web-based atlases are a widely used, effective web service (*see* http://www.mapquest.com for the current leader in this field). Other websites routinely use maps generated by these web services as plug-ins within their own application. In this way, we propose a web service based on neuroanatomical atlases for neuroscientists. In this system, parameters are delivered to the application via a Universal Resource Locator (URL) that specifies the atlas level number, the desired zoom factor, and the coordinates of the desired view to receive a scaled, cropped, and annotated bitmap image of the desired region of the atlas. The atlas server is used directly by the Atlas Mapper project (*see* "The Atlas Mapper") to provide the images on which the Atlas Mapper's delineations appear. It is also possible to search for named landmarks (i.e., areas and nuclei) so that the label of the structure in question appears in the center of view. Our current version is based on the Swanson atlas of the rat brain (Swanson, 1998) but could use any electronic atlas that is expressed as a set of Adobe Illustrator or PDF files so that the system could be used for any species with a sufficiently detailed electronic atlas. This project, like any other that involves copyrighted information, will require approval from the publishing house that owns the atlas.

The Atlas Mapper

The unification and standardization of neuroanatomical nomenclature in order to alleviate much of the past and present confusion surrounding the naming of brain regions has been an ongoing effort for over 100 years (Wilder, 1896; Bowden and Martin, 1995; Bowden and Dubach, 2003, this issue). Enforcing a set of standardized definitions lacks flexibility, however, and may not allow researchers to describe neuroanatomical delineations in the detail they would like (Swanson, 1998). Rather than devise new treatments for the neuroanatomical nomenclature (i.e., the names researchers use to denote brain structures), our approach (called the "Atlas Mapper") allows users to delineate volumes of brain tissue explicitly on a standard atlas using mapping techniques from standard drawing applications. They may then save those delineations in the knowledge management system.

Figure 4 shows a typical screenshot (corresponding to a region from the graphical fragment shown in Figure 4) of the Atlas Mapper control in "Insert" mode.

These delineations are based on the perceptions from users derived from images in papers. As such, they are unlikely to be very accurate for several reasons (size of printed image, plane of section of the image, etc.), and so our system permits users to estimate the error in their delineation through the use of Fuzzy Bezier Splines (FBS). The graphical controls of FBS are shown in Figure 5. Essentially, the FBS handles define the extent of a "border zone" where inclusion in the structure may be defined probabilistically so that the central

anchor point has a 0.5 probability of being inside the structure and each of the Fuzzy Spline Handles lie one standard error either inside or outside the structure. FBS curves are drawn on more than one section to provide a stack that delineates a volume. Within the NeuroScholar system, many such volumes comprise a map that may describe how different regions of the brain possess different characteristics (such as histological labeling patterns generated from neuroanatomical experiments).

The Experimental Flowchart

The NeuroCore project seeks to provide a "base ontology" for neuroscientific data by defining a data model with base classes that may be extended for different laboratories and data sets (Grethe et al., 2001). One section of NeuroCore that was carefully emphasized was a generic table to capture the experimental method for individual papers. We have elaborated this idea by representing the workflow of the experimental method in a paper as a modified UML activity diagram. This is shown in Figure 5 where the user is adding an activity to the workflow (in this case, the "extinction day" procedure).

The experimental flowchart plug-in allows users to build descriptions of the organizational flow in an experiment in a method described in detail elsewhere (Burns, 2001b). Figure 6 illustrates the organization of a seminal eyeblink classical conditioning study published in 1962 (Gormezano et al., 1962, also discussed in Burns, 2001b) where each "activity state node" in the flowchart on the right-hand side corresponds to an experimental day (each node contains subnodes that denote procedures performed that day). Each edge that passes between nodes has a weight corresponding to the number of animals involved in that transition (18 in Figure 5). When an edge connects a node to itself, its weight shows how many times that step is repeated. This tool permits experimental procedures to be depicted graphically in a straightforward manner. Within the finalized system, we intend to link activity states that are concerned with specific measurements to the actual data themselves. The structure of this connection has been described previously (*see* Figure 3 in Burns, 2001b) and will appear as linked Views within the View Graph.

The final variety of View Graphs is illustrated in Figure 5. The example of a "Local View Graph Instance" only shows the current View Instance on display in the center surrounded by its immediate neighbors. We use commercial software specifically designed for knowledge navigation (http://www.thebrain.com) to move between Views so that only the local knowledge is displayed. By using graph-based approaches that may be generated automatically from data models and view specifications we hope that users may straightforwardly organize and navigate through large knowledge models.

Knowledge Management Core (KMC)

The intentions and the design of the knowledge management core have been described in detail previously (Burns, 2001a; Burns, 2001b) and will be briefly reiterated here. There are five key capabilities that the KMC delivers.

1) **Users may interact directly with the contents of the primary literature.** Rather than supporting an interpretation of data by a reference to the paper, users may support their ideas by linking to the relevant paper, page, and passage of interest as fragments.
2) **Users may build models to represent papers' data, methodology, and conclusions.** Users may build models that accurately capture the knowledge content of papers; we provide specialized tools to accomplish this based on malleable data models. Justification for the definition of these models is derived from links to supporting fragments.
3) **Users may build computational models of**

their own knowledge using human reasoning. Users may model their own knowledge based on their representation of papers' contents.

4) **Users may argue, refute, support, and question the knowledge models of other users in the system.** Users may allow their knowledge models to be accessed by other users on the system and the KMC provides toolsets that permit that conversation to be constructed surrounding the knowledge models in the system. Users may also query the data based on their preferences and opinions, tracing work that they have said they believe to be reliable.
5) **Knowledge models may be aggregated and analyzed to address the specific question under study.** We provide tools to summarize the contents of knowledge models into a larger model. We also provide data analysis techniques to map the organization of these data summaries in order to provide a complete description of the subject under study.

Essentially, the KMC is an implementation of our underlying data management system (*see* "The View-Primitive-Data Model Framework") concerned with the defining high-level entities and rules that are extended further by the definition of the NeuroScholar system's data model and toolsets within the domain of neuroscience (*see* Burns, 2001b). The KMC might, in principle, be extended into other domains as well. This section is concerned with some of the technology we use at this level to evaluate users' opinions and to interpret the rule sets generated within the system.

There are four constructs we use to record human opinions within the system: *Comments* (where users may annotate the system's contents however they would like), *Justifications* (where users are required to justify the definition of an item in the system by linking it to another piece of data and to explain the link), *Viewpoints* (where users score their confidence levels in the attached piece of information), and finally a *Judgment* (where users may select between two items that have been shown to contradict one another). These constructs are included to provide users with a clear way of capturing and reconstructing their own reasoning rather than attempting to use computational inference to automate it.

The usage of the NeuroScholar system will depend on the opinions and preferences of its users. To accommodate this, we have incorporated support for "soft queries," which employs a fuzzy-logic based aggregation technique, to permit users to retrieve customized information from the system based on their preferences. Our proposed soft query technique can further use genetic algorithms to extract users' confidence values for different users in the system based on their usage behaviors. This methodology has been applied in the field of e-commerce applications (Chen and Shahabi, 2002).

The soft query method aggregates high-level entity data and the corresponding human perceptions (such as confidence values to other users, authors, journal, experimental methods) in order to provide customized results that are appropriate to users' preferences (Chen and Shahabi, 2001). This method also allows users to consult and adopt other users' opinions. In this way, we assert that "if user **x** believes user **y**, the concepts that satisfy the query criteria based on **y**'s judgments can be retrieved for **x**." For example, assume user **x** does not provide any confidence values concerning which methodology he prefers. If user **x** believes user **y**, who has specified confidence weights for different methodologies, the soft query method would also take user **y**'s opinions into consideration during aggregation processes for user **x**. Within the system, a user could assign confidence weights for many of the different computational entities within the NeuroScholar system (individual users, authors, specific journals, experimental methods, etc.), so that when querying the system,

each object in the result set will be prioritized according to the weighted aggregation data.

The Portable UNIX Programming System (PUPS) uses homeostatic computational processes to run robust, adjustable computations (O'Neill and Hilgetag, 2001). PUPS has been used to build optimized maps showing ordering or clustering of complex data sets in neuroinformatics (Hilgetag et al., 1996a; Hilgetag et al., 1996b). We will use PUPS in conjunction with users' opinions to map the current contents of each user's knowledge modeling space dynamically, so that each time a user updates the model, the map will accommodate changes. We are initially focused on evaluating each user's account of the neural connectivity between the structures of interest to them.

The PUPS system is coded in ANSI-C and was developed under the LINUX operating system using the Free Software Foundation GNU compiler tools. PUPS has been ported to a number of POSIX.1b compliant operating systems including OSF1, Solaris, SunOS 4.1 and BSD4.4. PUPS is supported as a open-source project on Sourceforge (http://www.sourceforge.net/projects/pups.

The View-Primitive-Data Model framework (VPDMf)

The VPDMf forms the base of the technology described in this article and as such forms the foundation of almost all the software described here. This section is geared to the interests of dedicated neuroinformaticians: specialists whose primary interest lies in the construction and evolution of computational systems themselves. Here, we describe the functionality and design of the VPDMf as a set of computer aided software engineering (CASE) tools that neuroinformaticians may have direct access to at zero cost. CASE tools permit the standardization of tool sets, straightforward communication of software design, and the acceleration of code development with forward/reverse engineering methods. Although common in industry, broad adoption of CASE software in academia is slow (due mainly to the high cost of these products such as Rational Rose). The VPDMf does reproduce some of the automated methods of commercial tools, but it also provides an entirely novel approach to representing a system by superimposing a formalized framework over the populated data model to encapsulate content into "Views." Within this section, we discuss a simplified example based on a small section of the design of the KMC (see "Knowledge Management Core (KMC)" Burns, 2001a) to illustrate how views are built within the VPDMf and these definitions may be used to build navigable models of the data model that support the Graph-based approaches shown in Figures 2, 3, and 5.

Figure 7 shows a class diagram in the UML that illustrates the static characteristics of several classes defined in the KMC. For example, instances of the "Publication" class are citations to the literature. The "n-to-n" association between the Publication and Person classes signifies that each cited paper must have "one or more" authors, and each person may be an author of "one or more" publications. Each publication refers to the controlled vocabulary (CV) class in two attributes, the "publication_type," and the "language." The principles of object-oriented design and the use of the UML are well documented, and will not be described here (*see* Rumbaugh et al., 1999).

The VPDMf provides an abstraction method that may be superimposed over a data model to encapsulate related classes into "Views" made up of "Primitives" spanning a small portion of the data model. As shown in the central section of Figure 7, every instance of the Publication class would be linked to one or more authors as instances of the Person class, a CV object denoting the language of the paper and another CV object denoting the type of the publication. It is straightforward to

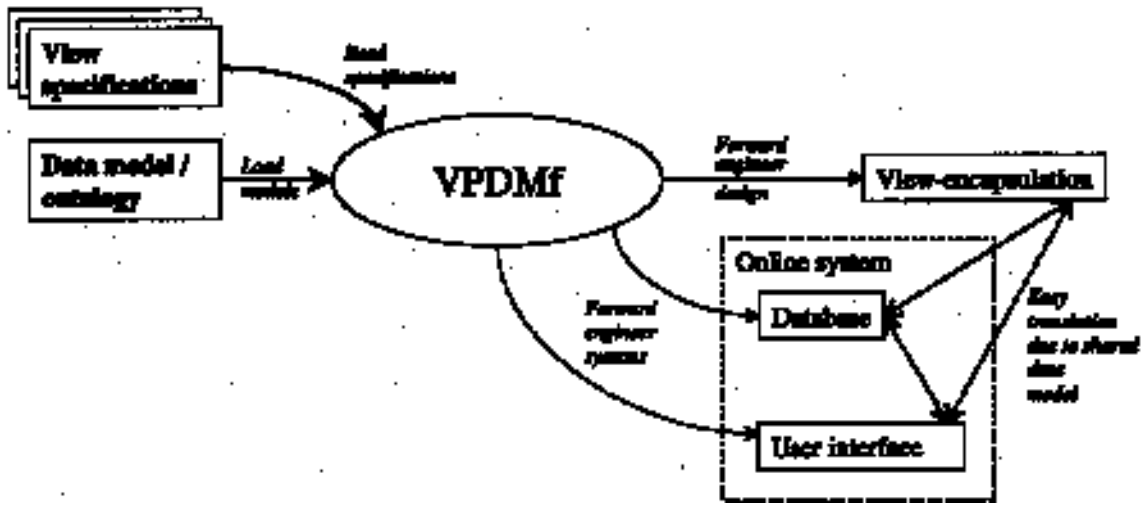

Fig. 8. Program flow underlying forward engineering in the VPDMf. (Burns, G. A. P. C., pp. 46, 7/24/2002). **[CFO]**

specify the structure of the view by (a) describing the class-composition of each primitive, (b) naming which primitive is considered "primary" (i.e., this primitive always has a cardinality of one and forms the core of the view definition), and finally (c) describing the associations that link the primitives together. All other data that must be taken into account when dealing with the representation (such as the type of each attribute, the cardinality of each association, etc.) is explicitly described in the data model. These three pieces of information are referred to in this article as a "View Specification."

Naturally, it is possible to define a large number of views by superimposing different tiled, enclosed or overlapping definitions onto the data model. This permits us to construct a formal methodology for navigating from view to view by traversing these relationships. For example, if we defined a "Textual Fragment" view based on a single primitive made up of the Fragment and Textual_Fragment classes, we can calculate that there is one route in the data model linking the two views: the 1-to-n aggregation association represented shown with a diamond at one end (the apparent overlap would not be considered since each Primitive has non-overlapping conditionality placed on their attribute values). By capturing these relationships, we permit a graph-based representation for navigating between Views called (somewhat unimaginatively) a "View-Graph." This forms the basis of the graph-based data navigational systems in Figures 2, 3, and 5.

The data model diagram in Figure 7 could easily serve as the conceptual design of a dataset in several different applications: a database scheme, object-oriented classes in a user interface or a schema description of web-based resources. In Figure 8, we illustrate the use of software engineering techniques to automate the generation of a working system (made up of these three specific applications). The input data for this process consists of the specification of a data model combined with an appropriate set of View specifications (as defined above). Importantly, the process of communicating between the three components is simplified enormously since they all derive from the same design.

The schema of the physical system based upon the conceptual design may undergo lan-

guage- or system-specific design changes (for example, adapting object-oriented models to a relational design require that n-to-n relationships are represented by the insertion of an intermediate link class and the addition of attributes for primary and foreign keys, Ullman and Widom, 1997). Within the VPDMf, this transformation process is entirely automated.

The "view encapsulation" component is an XML-based wrapper around the database to provide a standardized view-based web interface to the contents of the system. This may provide a way of mediating knowledge between systems with dissimilar data models (Burns et al., 2001). This may also provide the basis for publishing the system as a web service since web services use XML formatters to communicate (*see* work in the W3C consortium concerning SOAP messages at http://www.w3.org/TR/SOAP/).

The most powerful aspect of this framework is that it incorporates a model of the UML itself (similar to the reflection capabilities of Java, *see* Campione et al., 2002). This feature allows the VPDMf to be superimposed over any software that may itself be represented by the UML. The applicability of this is very widespread since the UML is designed to be generally applicable within the software engineering industry (Rumbaugh et al., 1999). Data models may be described in the UML or the XML Schema language (Duckett et al., 2001).

Discussion

NeuroScholar strives to provide an online environment for the comprehensive evaluation and interpretation of the neuroscientific literature in order to answer specific, stated, high-level questions. This will permit neuroscientists to expand their theoretical view of the subject by removing ambiguity concerning the large-scale information sets that describe characteristics for the wider system. The main obstacles to designing and building this system may be distilled into three main issues: (1) Neuroscience is heterogeneous and nonstandardized (with regard to techniques, data, interpretations or the nomenclature); (2) the information in the literature is subjective and contextual; (3) most experimental neuroscientists are not inclined to adopt new computational methods if the methods are technically problematic or unreliable. Here, we briefly discuss our strategies for overcoming these obstacles and how they relate to existing work.

Consider that the process of "Knowledge engineering" involves three interlinked subdisciplines: logic, ontology, and computation (Sowa, 2000) where the term "ontology" refers to "an explicit formal specification of how to represent the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them" (Howe, 2001). We assert that the lack of standardization across neuroscience's various domains can really only be addressed by defining explicit, unambiguous ontologies for each of the domains in turn and then relate the components of different ontologies to one another to permit interaction and translation. If users feel that it is scientifically inappropriate to adopt any given set of standards within a specific context, then they should be encouraged to define their own as long as they describe how to map from their descriptions to the standard set.

For example, the ambiguity within the neuroanatomical nomenclature was mentioned briefly in a previous section (*see* "The Atlas Mapper"). The objective relational transformation (ORT) project defines a methodology to use set-theory to track the relationships between different brain structures and then translate the data embedded in those regions between different parcellation schemes (Stephan et al., 2000). This method provides a practical mechanism of standardization without relying on individual researchers to agree

to adopt a standard reference scheme. Within this approach, standards will emerge over time as the most widely used solutions and if new discoveries force a change of approach or terminology the system can naturally evolve. The KMC supports this functionality within its use of rules and set theory (*see* "Knowledge Management Core" and Burns, 2001b).

An important characteristic of the neuroscientific literature is that it is too large for one individual to definitively understand everything about a given subject when considered for the whole brain. Every neuroscientist's understanding of the literature can be considered subjective. We specifically target this concern within NeuroScholar by providing a unique "workspace" for users so that they may examine the literature, build their own knowledge models, and then use the knowledge models in analyses or publish them so that other users may adopt (or refute) them.

It is the authors' experience that when presented with this idea, some researchers feel that neuroscientists *as a community* would be reluctant to express their ideas in this way, preferring still to publish new theories within the narrative of text-based articles and reviews. We assert that a trend already exists within publishing that will naturally evolve into a system similar to NeuroScholar. For example, the Signal Transduction Knowledge Environment is an online journal (http://stke.sciencemag.org/) with the following stated purpose: "[to] maximize the efficiency with which the reader gathers, assimilates, and understands information about cell regulatory processes." The development of the NeuroScholar system enables neuroscientists to address questions that hitherto, have been impossible to formulate effectively.

Central to this concept is the usability of the system. The success of the project is completely dependent on whether the system is simple enough for noncomputational neuroscientists to understand, beneficial to users in their work, and reliable within its functionality on a day-to-day basis.

The origins of information science itself derived from early attempts to systematize published knowledge. The pioneering work of Paul Otlet was the driving force behind the International Institute of Bibliography from 1895 to 1935. He constructed the "Classification Décimale Universelle" (in English, the "UDC") as an extension of the Dewey Decimal system. This methodology acted as "an immense map of the domains of knowledge" (Otlet, 1918; Rayward, 1998) and was used for the institute's efforts to catalog large volumes of bibliographic citations (the "Répertoire Bibliographique Universel" or "RBU" contained 16 million records), graphical records (containing as many as 250,000 entries) and also "full-text documents" (containing as many as one million items in 10,000 subject files). The UDC was based on a complex numbering scheme that was designed to provide multiple routes of access to an individual document, and could be considered as an early, fully formed practical database management system.

From these beginnings, the technology surrounding library-based databases has grown massively. The National Library of Medicine houses several web-accessible databases that are used by the scientific community worldwide. The usage statistics for PubMed exceeded 329 million individual searches in 2001, which corresponds to a rate of roughly 10 per second (National Library of Medicine, internal documentation). PubMed contains over 11 million records, which, incredibly, is less than the maximum content of the RBU at its greatest extent in 1930 (Boyd-Rayward, 1998). Full text for an increasing number of journals is available online, some as an archival resource at no cost (http://www.jstor.org). Therefore, indexes of published scholarly information not only constituted the earliest form of centralized databases, but also form one of the essential tools of scholarly work.

The Brain Browser is an atlas-enabled cross-domain computational encyclopedia that was originally distributed as a Macintosh Hypercard application, and was probably the first serious attempt to generate a neuroinformatics database that was based on published (or "public") neuroscientific data (Bloom et al., 1990). This was a commercial product, designed to act as a form of computational textbook that could be annotated and expanded by the user according to individual requirements rather than a direct interface to the primary literature.

One of the necessary precursors of early studies of neural connectivity was the development of databases capable of storing the data (Nicolelis et al., 1990). A large-scale collation study of the hierarchical organization of primary visual cortex was performed without computational support by Felleman and Van Essen in 1991, which then prompted a substantial research effort into the analysis and representation of connectivity data for the macaque monkey (Young, 1992; Young, 1993; Young et al., 1995), the cat (Scannell et al., 1995; Scannell et al., 1999), and the rat (Burns, 1997; Burns and Young, 2000). Thus by the mid-to-late nineties, several independent neural connectivity applications existed (with little or no interoperability between them).

These were notably followed by the development of a literature-based database for the study of Macaque Cortical Connectivity (called "CoCoMac"; Stephan et al., 2001). This system utilizes a well-defined methodology for translating connectivity data between different neuroanatomical nomenclatures using set-theoretical rules to track inferences that permit data to be translated from one neuroanatomical schema to another (Stephan et al., 2000). CoCoMac is an example of a well-populated, fully functional relational database system based on information from the literature that permits data exchange with other systems via an XML-enabled web interface. Other projects include the NeuroHomology database, designed to evaluate the validity of homologies between brain structures in different species (Bota and Arbib, 2001).

An important feature of the earlier neural connectivity database work in the rat (Burns, 1997; Burns and Young, 2001) was that each record in the database included an abbreviated copy of the original text that described the data. This concept was identified as pivotal within development work into "summary databases" (Arbib, 2001). Within NeuroScholar, we define "fragments" to denote the "raw data" that forms the substrate onto which the interpretations of NeuroScholar may be overlaid. The Annotator project in the University of Southern California's Brain Project was formative in the conceptualization of the Fragmenter component of NeuroScholar (Ovsiannikov and Arbib, 2001).

This perspective, of *superimposing* an interpretative framework onto fragments extracted from any data source (as long as it is web-accessible) is unique to the NeuroScholar system, and may provide a powerful capability for expansion of the system's capabilities in the future. Notably, this perspective dovetails with other neuroinformatics developers who are designing their systems so that their content might serve as publications themselves (Gardner et al., 2001). Thus, NeuroScholar might be useful as an interpretive methodology working within the conventional literature as well as other emerging technologies from within neuroinformatics.

When faced with a real-world application, systems designers parameterize and describe their view of the world as a "data model" (or "ontology" according to the definition above), and importantly, their world-view is *directly* influenced by the application that they are building. If the designers are forward thinkers, they will attempt to maximize the utility of their system by designing it for more than one purpose, or they will make it con-

form to standard methods that permit sharing and reuse of the knowledge in the system (Musen, 1992). It is important to note that the software development process is iterative, based on a cycle of design, implementation, and testing strategies. From a development viewpoint, it is highly desirable to iterate through this cycle as rapidly as possible. This means that effective software engineering strategies for adapting and redesigning data models are potentially very important tools (Muller, 1999).

Different groups in neuroinformatics approach the problem of positioning their data models in different ways, with some workers emphasizing markup languages in order to facilitate communication between systems (Goddard et al., 2001), while others have accomplished interoperability by defining a common data model for different systems, (Gardner et al., 2001; Grethe et al., 2001). Other developers have simply described the data models of their systems diagrammatically in order to describe their approach as explicitly as possible (Stephan et al., 2001; Burns, 2001a). The KIND architecture (Gupta et al., 2000), adopts the ontology of the Unified Medical Language System of the National Library of Medicine (UMLS) within a FLOGIC-based system. This approach may also form the basis for mediation between databases (Burns et al., 2001).

The Protégé 2000 project at Stanford is concerned with building practical modeling tools for knowledge sharing and reuse (Noy et al., 2000). Users may download the Protégé application to their local machine and build an ontology for a specific domain. The project is open-source and is supported by an international community of developers. The VPDMf project and Protégé both use data modeling at their core but have some key differences. Protégé is a frame-based knowledge management system with emphasis placed on the development of ontologies in widely dispersed domains. The VPDMf is principally a software engineering paradigm based on the UML to allow accelerated software development. Protégé supports interfaces to other knowledge representations such as Ontolingua (Gruber, 1993), the Knowledge Interchange Format (KIF; Genesereth, 1991), the open knowledge base connectivity specification (OKBC; http://www.ksl.stanford.edu/software/OKBC/), the resource description framework (RDF) as part of the development work within the semantic web (http://www.w3.org/2001/sw/), and the XML Schema language (which is itself supported by the VPDMf).

The field of bioinformatics has many highly engineered tools for various well-defined tasks such as displaying molecular structure, performing data analysis, etc. (e.g., see the Protein Databank's website for a cross section of tools: http://www.pdb.org/). An in-depth analysis between the bio- and neuroinformatics tools is beyond the scope of this article but many of the research problems in bioinformatics work within a problem space defined by a quantitative, mathematically-tractable ontology (as defined above). As we have discussed in this paper and elsewhere (Burns, 2001b), the problem space of neuroinformatics is defined by a qualitative, nonstandardized set of ontologies. We are therefore compelled to address our technical problems in a different way (by building subjective, multi-user systems; by addressing one specified high level question, etc.) whereas bioinformatics solutions can be large-scale without having to implement these measures.

NeuroScholar is a natural extension of the technology described in many of these other projects (such as work in the analysis of neural connectivity). Many of the tool sets we are developing are similar to other developers' work (e.g., the Atlas Mapper was derived from the NeuARt project from the USC Brain Project, Dashti et al., 1997), however the com-

bination of all of these techniques into our theoretical framework is unique (Burns, 2001a).

At present, the development of ontologies and data models is not a continuous process. If the data model needs to be redesigned, changes are typically made offline and then subsidiary systems would need to be updated and the data ported to the new system. Cellular Database Management Systems (CDBMS) present a very new technology for robust, dynamic data-handling without the need for predefined data models (Gelfand, 2002 a,b).

Cellular Database Management Systems are a new technology. The process of building a database starts by having to fully define its data model, which is essentially how the database "sees the world." In general, this viewpoint is static and relatively simple (consisting of, in the absolute largest cases, several thousand typed object classes in Data Warehouse solutions). Within this database system, any and all facts present in the world may only ever be expressed within this data model and all users of the system are restricted to the same absolute vocabulary. If the vocabulary becomes obsolete, or the information in the world does not exactly conform to the rigidly defined rules of the system, at best, the system loses accuracy, and at worst, it breaks. Within the CDBMS model, it is possible to insert and retrieve ambiguous data at any level. The system works with a large network of interconnected "cells" to store data (rather than tables, or objects). Each cell is a vector of three integers and a data value. The connections between cells may be formed and unformed dynamically to reconfigure the system's data model intelligently while the system is in use. If the stated objectives of the system's inventors are realized, this will be an extremely important development, permitting data models to be changed and updated on the fly, with no loss of functionality.

The fuzzy relationships that are often found between individual data items within neuroinformatics datasets, especially those derived from knowledge management systems, are usually analyzed with numerically intensive techniques (for example, stochastic optimization, fuzzy template matching and analysis of large graphs). Thus, even with parallel or clustered computing facilities, applications may have to run for a time period of days to weeks, given the relatively low speeds of current hardware. This has a number of implications. Applications should be restartable in the event that the host system software and hardware fail. The use of dynamic load balancing in network-cluster computer environments may permit the optimal use of available resources and could facilitate the reliability of the system, since processes may migrate away from malfunctioning nodes.

The Portable UNIX Programming System (*see* "Knowledge Management Core") is designed to perform homeostatic computations. Namely, the computational processes performing the work of the calculation manage their environment through a number of mechanisms. These include automated process migration via MOSIX (Barak et al., 1993); support for recoverable processes via the Tennessee checkpointing protocol (Plank et al., 1995); homeostatic protection of data items; peer-to-peer and user-to-peer dynamic interaction with running processes; support for dynamic goal reassignment and steering parameter update in running applications; and support for manipulation and storage of complex datasets via a practical multihost/mulitprocess implementation of a persistent object store.

For the NeuroScholar project, this is of particular interest, since it is expected that the theoretical landscape of the subject of neuroscience will be dramatically reshaped by developments in neuroinformatics. As new discoveries arise, our successes will be deter-

mined by the agility with which we can shift our theoretical perspective to meet new data.

Acknowledgments

This work was supported under an RO1 from the National Library of Medicine (number LM07061). We thank the collective members of the Brainwalk discussion group in the Neurobiology department at USC for their valuable input and suggestions. Thanks to Alan Watts for his conceptual input concerning application of NeuroScholar to the CRH system. Thanks to Shyam Kapadia, Kevin Chan, Wei-Cheng Chen, Atousa Golapayegani, Shanshan Song, Yan Zhou and Ning Zhang for their contributions to the development of the application. Thanks to Cyrus Shahabi for his valuable feedback and support during the writing of this manuscript.

References

Arbib, M. (2001) NeuroInformatics: the issues, in: *Computing the Brain, 1* (Arbib, M. and J. Grethe, J., eds.) Academic Press Inc., San Diego, CA, pp. 3–28.

Axelrod, J. and Reisine, T. D. (1984) Stress hormones: their interaction and regulation. *Science* 224:452–459.

Bloom, F. E., Young, W. G., and Kim, Y. M. (1990) Brain Browser, Hypercard application for the Macintosh, Academic Press Inc., San Diego, CA.

Blum, B. (1986) Clinical Information Systems, Springer, New York.

Bota, M. and Arbib, M. (2001) The NeuroHomology Database, in: *Computing the Brain, 1,* (Arbib, M. and Grethe, J., eds.) Academic Press Inc., San Diego, CA, pp. 337–354.

Bowden, D. M. and Martin, R. F. (1995) Neuro Names Brain Hierarchy. *Neuroimage* 2:63–83.

Boyd-Rayward, W. (1998) The Origins of Information Science and the International Institute of Bibliography/Internatinoal Federation for Information and Documentation (FID), in: *Historical Studies in Information Science*, (Hahn, T. and Buckland, M. eds.) Information Today, Inc. Medford, NJ.

Burns, G. (2001a) Knowledge Mechanics and the Neuroscholar project: a new approach to neuroscientific theory, in: *Computing the Brain, 1* (Arbib, M. and Grethe, J. eds.) Academic Press, Inc., San Diego, CA, pp. 319–336.

Burns, G. A. (2001b). Knowledge management of the neuroscientific literature: the data model and underlying strategy of the NeuroScholar system. *Philos Trans R Soc Lond B Biol Sci* 356:1187–1208.

Burns, G. A. and M. P. Young (2000). Analysis of the connectional organization of neural systems associated with the hippocampus in rats. *Philos Trans R Soc Lond B Biol Sci* 355:55–70.

Burns, G. A. P. C. (1997) Neural connectivity of the rat: theory, methods, and applications, Laboratory of Physiology, University of Oxford, Oxford, UK.

Burns, G. A. P. C., Stephan, K. E., Ludäscher, B., Gupta, A., and Kotter, R. (2001) Towards a federated neuroscientific knowledge system using brain atlases. *Neurocomputing* 38–40:1633–1641.

Campione, M., Walrath, K., and Huml, A. (2002) The Java Tutorial, Third Edition.

Chen, Y.-S. and Shahabi, C. (2001) Automatically improving the accuracy of user profiles with genetic algorithm. IASTED International Conference on Artificial Intelligence and Soft Computing, Cancun, Mexico.

Chen, Y.-S. and Shahabi, C. (2002). Improving user profiles for e-commerce by genetic algorithms, in: *E-Commerce and Intelligent Methods in studies in Fuzziness and Soft Computing*, (Segovia, J., Szczepaniak, P. S., and Niedzwiedzinski, M., eds.), Springer-Verlag. Heidelberg, Germany.

Chicurel, M. (2000) Databasing the brain. *Nature* 406:822–825.

Dallman, M. F., Akana, S. F., Cascio, C. S., Darlington, D. N., Jacobson, L., and Levin, N. (1987) Regulation of ACTH secretion: variations on a theme B. *Rec Prog Horm Res* 43:113–173.

Dashti, A. E., Ghandeharizadeh, S., Stone, J., Swanson, L. W., and Thompson, R. H. (1997) Database challenges and solutions in neuroscientific applications. *Neuroimage* 5:97–115.

de Groot, J. and Harris, G. W. (1950) Hypothalamic control of the anterior pituitary gland and blood lymphocytes. *J Physiol (London)* 111:335–346.

Duckett, J., Griffen, O., Mohr, S., Norton, F., Stokes-Rees, I., Williams, K., Cagle, K., Ozu, N., and Tennison, J. (2001) Professional XML Schemas, Wrox Press, Inc, Birmingham, UK.

Gardner, D., Abato, M., Knuth, K. H., DeBellis, R., and Erde, S. M. (2001) Dynamic publication model for neurophysiology databases. *Philos*

Trans R Soc Lond B Biol Sci 356:1229–1247.
Gelfand, B. (2002a) Data cells and data cell generations. US Patent Office. USA.
Gelfand, B. (2002b) Data cells, and a system and method for accessing data in a data cell. US Patent Office. USA.
Genesereth, M. R. (1991) Knowledge Interchange Format. Proceedings of the Second International Conference on the Principles of Knowledge Representation and Reasoning, Morgan Kaufmann, Cambridge, MA.
Goddard, N. H., Hucka, M., Howell, F., Cornelis, H., Shankar, K., and Beeman, D. (2001) Towards NeuroML: model description methods for collaborative modelling in neuroscience. *Philos Trans R Soc Lond B Biol Sci* 356:1209–1228.
Gormezano, I., Schneiderman, N., Deaux, E., and Fuentes, I. (1962) Nictitating membrane: classical conditioning and extinction in the Rat. *Science* 138:33–34.
Grethe, J., Mureika, J., and Merchant, E. (2001) Design concepts for NeuroCore and neuroscience databases, in: *Computing the Brain, 1*, (Arbib, M. and Grethe, J., eds.) Academic Press, Inc., San Diego, CA, pp. 135–150.
Gruber, T. R. (1993) A translation approach to portable ontology specifications. *Knowledge Acquisition* 5:199–220.
Gupta, A., Lüdascher, B., and Martone, M. (2000) Knowledge-based Integration of Neuroscience Data Sources. Proceedings of 12th Int. Conf. Scientific and Statistical Database Management Systems (SSDBM'00), Berlin, IEEE.
Herman, J. P. and Cullinan, W. E. (1997) Neurocircuitry of stress: central control of the hypothalamo-pituitary-adrenocortical axis. *Trends Neurosci* 20:78–84.
Hilgetag, C. C., O'Neill, M. A., and Young, M. P. (1996a) Indeterminate organization of the visual system. *Science* 271:776–777.
Hilgetag, C. C., O'Neill, M. A., and Young, M. P. (1996b) Optimization analysis of complex neuroanatomical data, in: *Computational Neuroscience*, Plenum Press, Boston, MA.
Howe, D. (2001) The Free On-Line Dictionary of Computing (http://www.foldoc.org/).
Koslow, S. H. (2000) Should the neuroscience community make a paradigm shift to sharing primary data? *Nat Neurosci* 3:863–865.
Muller, R. J. (1999) Database design for smarties, using UML for data modeling, Morgan Freeman, San Francisco, CA.
Musen, M. A. (1992) Dimensions of knowledge sharing and reuse. *Comput Biomed Res* 25:435–467.
Nicolelis, M. A., Tinone, G., Sameshima, K., Timo-Iaria, C., Yu, C. H., and Van de Bilt, M. T. (1990) Connection, a microcomputer program for storing and analyzing structural properties of neural circuits. *Comput Biomed Res* 23:64–81.
Noy, N. F., Fergerson, R. W., and Musen, M. A. (2000) The knowledge model of Protege-2000: Combining interoperability and flexibility. Proceedings of 2nd International Conference on Knowledge Engineering and Knowledge Management, Juan-les-Pins, France.
Otlet, P. (1918) Transformations in the bibliographic apparatus of the sciences, in The international organization and dissemination of knowledge: Selected essays of Paul Otlet (1990), W. B. Rayward ed., Elsevier. Amsterdam.
Ovsiannikov, I. and Arbib, M. (2001) Annotator: Annotation Technology for the WWW, in: *Computing the Brain, 1*, (Arbib, M. and Grethe, J., eds.) Academic Press Inc., San Diego, CA, pp. 255–264.
Plank, J. S., Beck, M. and Kingsley, G. (1995) Libckpt: transparent checkpointing under UNIX. Proc. USENIX Winter Tech. Conf., New Orleans, LA.
Rational (1997) UML Semantics Version 1.1. Santa Clara, CA, Rational Software Corp.
Rayward, W. B. (1998) The origins of information science and the International Institute of Bibliography/International Federation for Information and Documentation (FID), in: *Historical Studies in Information Science*, (Hahn, T. and Buckland, M., eds.) Information Today, Inc. Medford, NJ.
Rho, J. H. and Swanson, L. W. (1989) A morphometric analysis of functionally defined subpopulations of neurons in the paraventricular nucleus of the rat with observations on the effects of colchicine. *J Neurosci* 9:1375–1388.
Rumbaugh, J., I. Jacobson and G. Booch (1999) The Unified Modeling Language Reference Manual, Addison-Wesley, Reading MA.
Sapolsky, R. M., Romero, L. M., and Munck, A. U. (2000) How do glucocorticoids influence stress responses? Integrating permissive, suppressive, stimulatory and preparative actions. *Endocrine Rev* 21:55–89.
Sawchenko, P. E., Li, H. Y., and Ericsson, A. (2000) Circuits and mechanisms governing hypothalamic responses to stress: a tale of two paradigms. *Prog Brain Res* 122:61–78.

Sawchenko, P. E. and Swanson, L. W. (1989) Organization of CRF immunoreactive cells and fibers in the rat brain: immunohistochemical studies. In: *Corticotropin-Releasing Factor: Basic and Clinical Studies of a Neuropeptide*, (Souza, E. D. and Nemeroff, C., eds.) CRC Press. Boca Raton, FL, pp. 29–51.

Scannell, J. W., Blakemore, C., and Young, M. P. (1995) Analysis of connectivity in the cat cerebral cortex. *Journal of Neuroscience* 15:1463–1483.

Scannell, J. W., Burns, G. A., Hilgetag, C. C., O'Neill, M. A., and Young, M. P. (1999) The connectional organization of the cortico-thalamic system of the cat. *Cereb Cortex* 9:277–299.

Sowa, J. (2000) Knowledge Representation. Logical, Philosophical and Computational Foundations, Brooks/Cole, Pacofoc Grove, CA.

Stephan, K. E., Kamper, L., Bozkurt, A., Burns, G. A., Young, M. P., and Kotter, R. (2001) Advanced database methodology for the Collation of Connectivity data on the Macaque brain (CoCoMac). *Philos Trans R Soc Lond B Biol Sci* 356:1159–1186.

Stephan, K. E., Zilles, K., and Kotter, R. (2000) Coordinate-independent mapping of structural and functional data by objective relational transformation (ORT). *Philos Trans R Soc Lond B Biol Sci* 355:37–54.

Swanson, L. (1986) Organization of mammalian neuroendocrine system. In: *Handbook of Physiology, The Nervous System, IV*, (Bloom, F., ed.) Waverly Press, Baltimore, MD, pp. 317–363.

Swanson, L. (2000) Paraventricular nucleus. In: *Encyclopedia of Stress, Vol. 3*, Academic Press Inc., San Diego, CA, pp. 130–133.

Swanson, L., Sawchenko, P., Lind, R., and Rho, J.-H. (1987) The CRH motoneuron: differential peptide regulation in neurons with possible synaptic, paracrine and endocrine outputs. *Ann NY Acad Sci* 512:12–23.

Swanson, L. W. (1991) Biochemical switching in hypothalamic circuits mediating responses to stress. *Prog Brain Res* 87:181–200.

Swanson, L. W. (1998) Brain Maps: Structure of the Rat Brain, Elsevier Science Publishers B. V., Amsterdam, The Netherlands.

Tanimura, S., Sanchez-Watts, G., and Watts, A. G. (1998) Peptide gene activation, secretion, and steroid feedback during stimulation of rat neuroendocrine corticotropin-releasing hormone neurons. *Endocrinol* 139:3822–3829.

Tanimura, S. and Watts, A. (1998) Corticosterone can facilitate as well as inhibit corticotropin-releasing hormone gene expression in the rat hypothalamic paraventricular nucleus. *Endocrinol* 139:3830–3836.

Ullman, J. and Widom, J. (1997) A first course in database systems, Prentice Hall, Inc., Upper Saddle River, NJ.

Watts, A. and Swanson, L. (1989) Diurnal variations in the content of preprocorticotropin-releasing hormone messenger ribonucleic acids in the hypothalamic paraventricular nucleus of rats of both sexes as measured by in situ hybridization. *Endocrinol* 125:1734–1738.

Wilder, B. G. (1896) Neural terms, international and national. *J Comp Neurol* 6:216–352.

Young, M. P. (1992) Objective analysis of the topological organization of the primate cortical visual system. *Nature* 358:152–155.

Young, M. P. (1993) The organization of neural systems in the primate cerebral cortex. *Pro Royal Soc London B Biol Sci* 252:13–18.

Young, M. P., Scannell, J. W., and Burns, G. A. P. C.(1995) The Analysis of Cortical Connectivity, Springer/RG Landes.