ELSEVIER

Neurocomputing 38–40 (2001) 1633–1641

NEUROCOMPUTING

www.elsevier.com/locate/neucom

Towards a federated neuroscientific knowledge management system using brain atlases

Gully A.P.C. Burns[a,*], Klaas Stephan[b], Bertram Ludäscher[c], Amarnath Gupta[c], Rolf Kötter[b]

[a]*Knowledge Mechanics Research Group, USC, Los Angeles, USA*
[b]*C. & O. Vogt Brain Research Institute, Düsseldorf, Germany*
[c]*San Diego Supercomputer Center, UCSD, San Diego, USA*

Abstract

The topic of federated databases has received much attention within the domain of neuroinformatics and is widely perceived as the eventual solution to the problems inherent in building truly interoperable informatics systems. We describe a feasibility study of a methodology strategy for building federated informatics systems with a specific example from informatics approaches involving the neuroscientific literature. We examine the logistical issues concerning linking two database systems (CoCoMac and NeuroScholar) via a method based on the use of transformations between neuroanatomical parcellations (Stephan et al., Phil. Trans. R. Soc. London B 335 (2000) 37–54).

Keywords: Neuroinformatics; Database federation; Connectivity; Brain parcellation

1. Introduction

Database interoperability is a pressing issue within the field of neuroinformatics. Under current funding drives such as the Human Brain Project from the National Institue of Health, the number and variety of repositories of neuroscientific information is rapidly expanding [6,11]. The associated integration problems can be classified as follows [12]: system integration (hardware/software platforms, transport protocols,

* Corresponding author. Room 428, Hedco Neurpsciences Building, University of Southern California, 3614 Watt Bay, Los Angeles CA 90007-2520 USA. Tel.: + 1-213-740-7489; fax: + 1-213-741-0561.
E-mail address: gully@usc.edu (G.A.P.C. Burns).

0925-2312/01/$ - see front matter
PII: S0925-2312(01)00520-3

etc.), structural integration (relational, object-oriented, semistructured data), and semantic integration (common vocabularies, conceptual models, ontologies).

In particular, the problem of making systems interoperable requires practical solutions of data translation between different database schemas. Even for systems that employ designs that are highly integrated, translation would still be required in order to communicate with other databases or upgrade database designs if changes are required.

Within this paper, we describe the implementation of a method where users describe how two databases overlap in their representations with a target data model made up of data that may be found in both systems. We then describe how the contents of each database may be translated to representations that have similar organizations to the target, and then how rules of mediation may be used to translate the each system's data to the common representation.

Neuroanatomical atlases often provide the common substrate for different neuroscientific studies, so we use two databases that contain very similar types of data and base the mechanism of interoperability on the neuroanatomical parcellation schemes being used. Two such systems that deal with tract-tracing data, reported in the literature are CoCoMac and NeuroScholar.

The CoCoMac project of Klaas Stephan and Rolf Kötter of the C. & O. Vogt Brain Research Institute in Dusseldorf is a valuable data resource comprising over 17,000 experimental findings from over 1000 tract-tracing experiments describing neuroanatomical connections in the macaque Monkey [13] (see also http://www.cocomac.org/). The NeuroScholar project is at the beta stage of development, and exists largely as a system design. In several ways, NeuroScholar's semantic design was based on that of CoCoMac but it emulates an object-oriented approach rather than a relational one [2] (see also http://neuroscholar.usc.edu/). For this study, we embed test data derived from the same source as a selected study in the CoCoMac system and perform the translation to the common framework accordingly.

Importantly, the CoCoMac project originated the conceptual framework of the objective relational transformation ('ORT') which is the formation that we adopt as the common neuroanatomical substrate between the two databases. The ORT paradigm is based on evaluating how data is plotted within brain regions in a given parcellation scheme (described within ORT as 'extension codes') and then how brain regions in different parcellation schemes may be spatially related to each other (described in ORT as 'relation codes'). The implementation of ORT on the CoCoMac system includes algorithmic support to trace how data mapped into one parcellation scheme, may be translated to other schemes [13].

In order to perform the first step of the translation process, we must provide a versatile interface to databases of different types. The view-primitive-table-column (VPTC) data modeling project is an object-oriented PERL-based application to provide scripting support for database design. It is designed to provide a uniform representation of data where a 'View' is a graph (i.e. a network of nodes and edges) of linked 'Primitives' derived from an underlying relational, object-relational or object-oriented framework. Here, we use this framework to provide a methodology for database federation. The VPTC model is being developed as an open-source Perl

application (http://vptc.sourceforge.net) by Gully A.P.C. Burns and is at the alpha stage of development.

This paper is a simple feasibility study demonstrating the linkage between two databases: CoCoMac and Neuroscholar. We illustrate a method to generate a limited federation between two heterogeneous neuroinformatics systems that use a common neuroanatomical framework. We consider the process of translating data from CoCoMac to our target object-oriented representation in detail.

2. Methods and results

2.1. Input data

The data used in this study is taken from a published tract-tracing study concerned with corticocortical connections of subregions of the supplemental motor region in the Macaque monkey (Areas F3 and F6 [8]). This study conforms to the general design of tract tracing experiments (see [4] for recent review of this technology). Our example data is based on Diamidino Yellow (2% in 0.2 M phosphate buffer at pH 7.2, [9]) which is a fluorescent retrograde technique. Within this study, a number of injections were made into subregions of the supplemental motor area and labeling was described in other regions. In this study, this data is collated into both the NeuroScholar and CoCoMac databases and translated to a target representation (see below). We restrict our attention to one injection within one animal (Case 1, the injection into the Arm region of F3 [8]) and a limited number of labeled regions within the output data.

Specifically, the original text of paper describes the injection site simply: "*Figure 3 illustrates the pattern of retrograde cortical connections following injections in thre rostral part of F3 (arm field)*." [8, p. 118]. Descriptions of transported labeling in other regions can be broken into sections based on the identity of the labeled region. We will consider five such sections of text from [8, pp. 118–120]. The first refers to labeling in region F1: "*In F1, the labelled cells are located in the arm representation of this area. Most of them lie on the cortical convexity anterior to the precentral sulcus*." The second refers to a more complex pattern of labeling found in are *F2*: "*In F2, two irregular stripes of labelled cells can be recognized. One stripe starts from the precentral dimple and extends rostrally to the border with F7. The other is located more laterally near the border with F4. The upper part of F2 where the leg is represented* (*Kurata, '89*) *is devoid of labelled cells*." The third section refers to labeled cells in *F4*: "*In F4, there are a large number of cells labelled dorsally in correspondence to the arm field* (*Gentilucci et al., '88*)." When considered by readers, this data is normally considered in conjunction with illustrations but for the purposes of this example, we restrict ourselves to the textual data shown above.

2.2. Target representation

Fig. 1 shows a schematic account of the target representation to which both data from CoCoMac and NeuroScholar must be translated. The figure illustrates the most

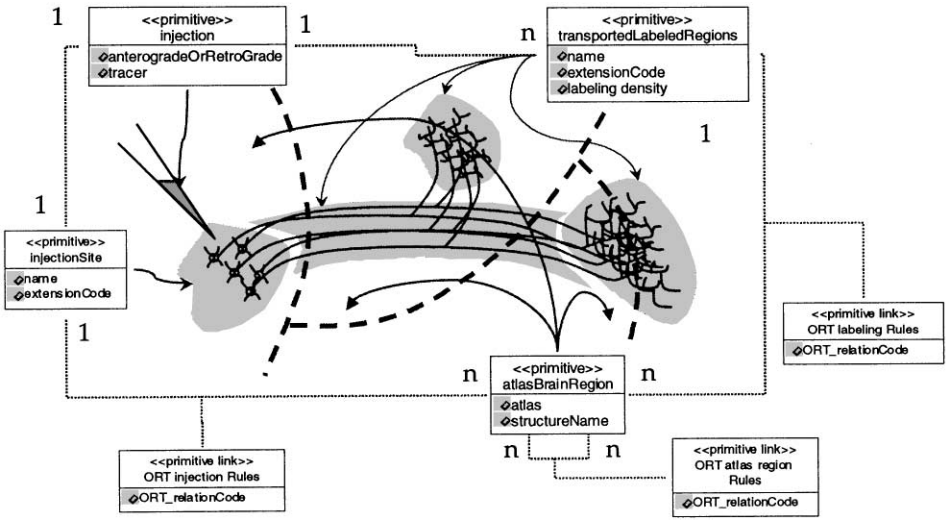

Fig. 1. Schematic diagram of target representation. Each primitive (see later) is presented schematically and semi-formally as classes in the unified modeling language.

minimal representation of a tract-tracing experiment possible after consideration of the two relational schemes of the two databases. An injection of a specific tracer that may be classified as either 'retrograde', 'anterograde' or 'both' is made into a specific region. After enough time has passed for the tracer to be taken up by parts of neurons passing through the injection site and transported within these cells, the animal is killed and its brain processed to reveal the transported label. When reported in an article, experimentalists present this data as accounts of labeling density within specified regions of the brain. In order to be made interpretable, these labeled regions and the location of the injection site must be related to standard parcellation schemes. Fig. 1 also shows classes and association classes defined in the universal modeling language (or 'UML') that model this representation [1]. As shown in Fig. 1, classes correspond to 'primitives' and association classes correspond to 'primitive links' within the VPTC model.

2.3. Relational schemata

CoCoMac is a fully normalized relational database that already contains a large amount of data (and may therefore be considered a legacy system, where changes to the schema will necessitate translation). NeuroScholar is a relational emulation of an object-oriented framework (and is not fully normalized for that reason) and is at a beta-testing stage. For this paper, both systems are implemented in Microsoft Access 97.

Several features of CoCoMac's representation differ from that of NeuroScholar and are worthy of note: (1) 'Precision of Description' codes (or 'PD' codes), (2) the existence

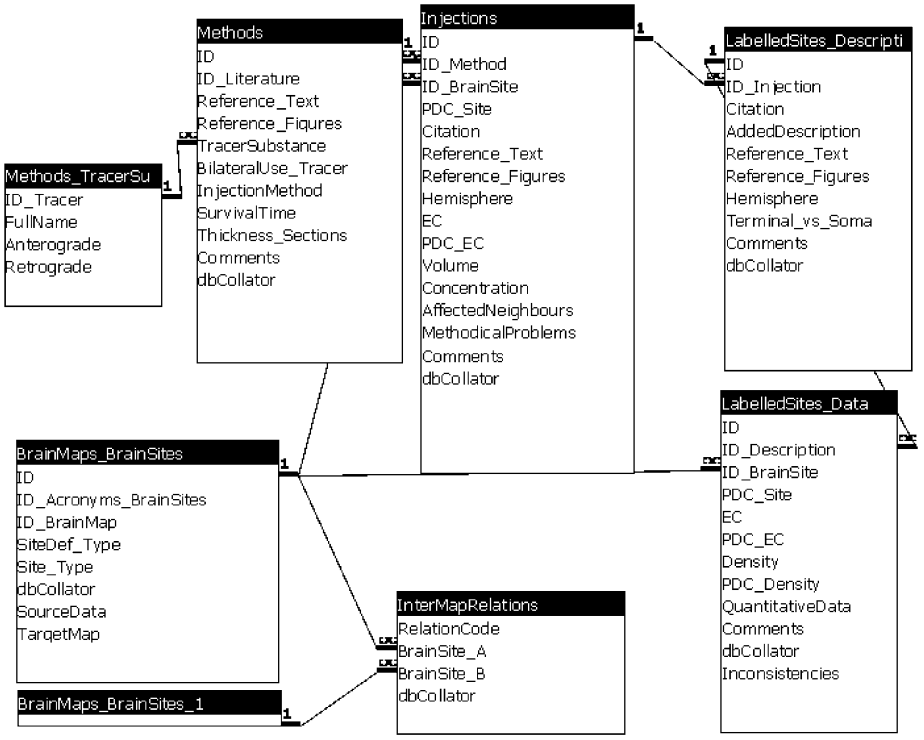

Fig. 2. The part of the database schema of CoCoMac of interest to this study.

of 'Citation' columns. These are all system-specific features that present difficulties that may necessitate translation in order to generate a mediated view. The PD codes are CoCoMac's method of keeping track of the precision of data as an indicator of their trustworthiness. This formulation is inherently difficult to match to other databases accounts since judgments of data reliability are subjective (although the designers of CoCoMac were careful to ensure that the PD codes promote an objective approach to data collation). The NeuroScholar system adopts a slightly different approach to this issue by presenting the primary data (text from the paper) in the fragment table (Fig. 2).

The other difference between the two databases was the treatment of the original text from the paper itself. Within CoCoMac, these data appear as 'Citation' columns within each table that refers to the text, and in NeuroScholar these data appear as separate entries rows in an entirely separate table [2]. We decided to omit this data from the target representation in order to keep it as simple as possible.

2.4. The VPTC representation of CoCoMac's data

We illustrate the construction of a mediated view for a specific example taken from CoCoMac's representation of data from [8]. From the text described above (see

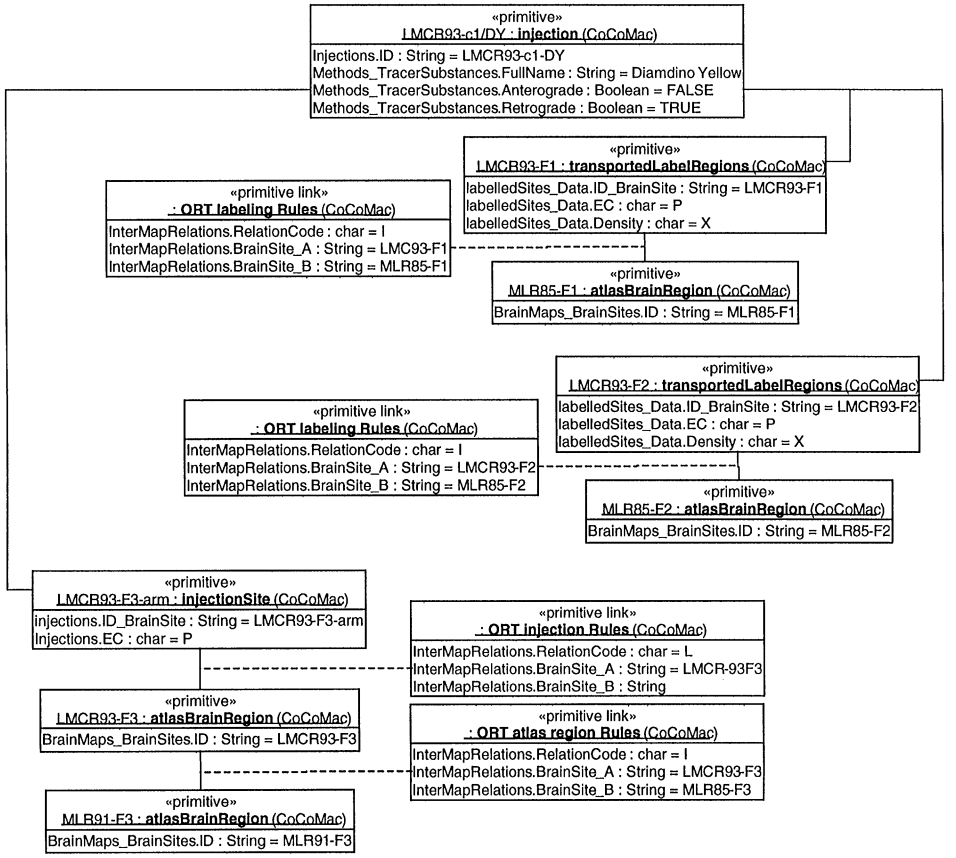

Fig. 3. The VPTC representation of the data extracted from the CoCoMac database expressed as an object diagram in the universal modeling language.

'Input Data'), the collators of CoCoMac entered data into several tables. We superimpose additional structure over that design based on the VPTC model and illustrated in Fig. 3.

Within Fig. 3, the view with the identifier 'LMC93-c1/DY' is made up of four primitives (combined with the primitive links denoting ORT rules), and each primitive is made up of several tables. Importantly, the cardinality of each primitive (i.e., the number of items in the data array of a given primitive) is consistent across all constituent tables, thus each primitive may be conceptualized as a self-contained data entity. The VPTC model samples the relevant tables to present only those data items that are relevant to the data translation process. So under the VPTC model, a 'View' is a graph whose nodes are 'Primitives' and edges are 'Primitive Links'.

The relational schema of the NeuroScholar system represents the data with a completely different combination of tables to form the same primitives as for the CoCoMac system. The resultant view has the same graph-like structure as that

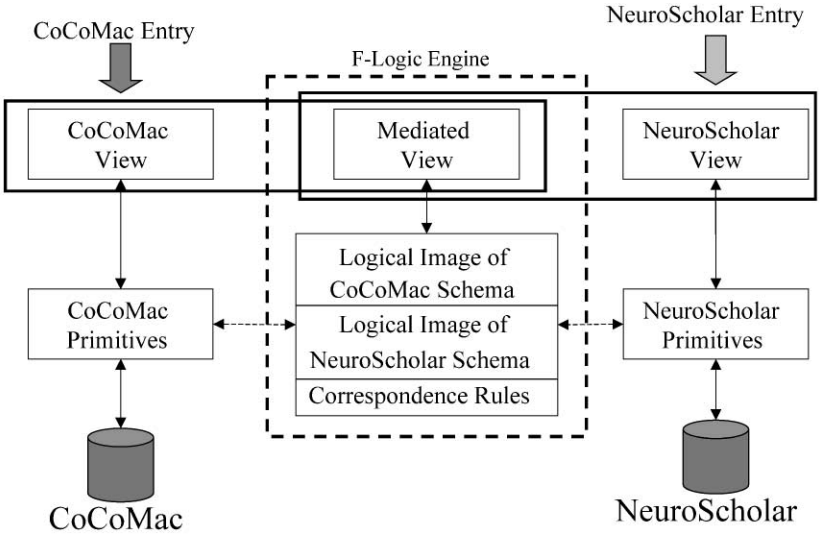

Fig. 4. The mediation strategy of this approach.

derived from the CoCoMac database; differences between the two VPTC representations arise from the precise configuration and format of individual columns within the primitives.

2.5. Mediation

The VPTC model is designed to organize columns and tables into an object-oriented–like framework for translation and evaluation. The translation itself is based on intelligent rule-based evaluation of each database to translate data into the common framework. This is illustrated in Fig. 4.

A simple example of this sort of translation occurs with data shown Fig. 3. The Methods_TracerSubstance table of CoCoMac uses two columns called 'Anterograde' and 'Retrograde' with data of type BOOLEAN to describe properties of a given neuroanatomical tracer, whereas the target representation uses a single column with a single character. Both representations are equivalent and the mediation engine translates between them straightforwardly. These correspondence rules are encoded in the deductive object-oriented database language, F-Logic [10]. This approach is well-suited to this task having been used for schema transformations [3] and information integration using semistructured data [7], as well as knowledge representation and reasoning with ontologies [5].

3. Discussion

This paper is concerned with the presentation of a method of translation to mediate between two relational databases. We have attempted to design the procedure to be

expandable to other database types, since the VPTC model includes concepts of inheritance in order to include object-relational schema (and includes input/output functions for both Oracle and Informix databases). Our strategy for data translation between systems is based on the same logic used by neuroscientists to translate data between studies and is a well-documented, published computational algorithm called the Objective Relational Transformation [13]. All components of this approach use open source software that may be obtained for free.

References

[1] G. Booch, J. Rumbaugh, I. Jacobson, The Unified Modeling Language User Guide, Addison-Wesley, Reading, MA, 1999.
[2] G.A.P.C. Burns, Knowledge Mechanics and the NeuroScholar Project, a new approach to neuroscientific theory, in: M. A. Arbib, J. Grethe (Eds.), Computing the Brain: A Guide to Neuroinformatics, Academic Press, San Diego, 2001.
[3] Y. Chang, L. Raschid, B. Dorr, Transforming queries from a relational schema to an equivalent object schema: a prototype based on F-Logic, International Symposium on Methodologies in Information Systems (ISMIS) Springer, Berlin, 1994.
[4] C. Kobbert, R. Apps, I. Bechmann, J. Lanciego, J. Mey, S. Thanos, Current concepts in neuroanatomical tracing, Progr. Neurobiol. 62 (2000) 327–351.
[5] A. Gupta, B. Ludäscher, M. E. Martone, Knowledge-based integration of neuroscience data sources, Proceedings of the 12th Intl. Conference on Scientific and Statistical Database Management (SSDBM), Berlin, Germany, IEEE Computer Society Press, Silver Spring, MD, July 2000.
[6] M. Huerta, S. Koslow, A. Leshner, The human brain project: an international resource. Trends Neurosci 16 (1993) 436–438.
[7] B. Ludäscher, R. Himmeröder, G. Lausen, W. May, C. Schlepphorst, Managing semistructured data with FLORID: a deductive object-oriented perspective, Inform. Systems 23 (1998) 589–613.
[8] G. Luppino, M. Matelli, R. Camarda, G. Rizzolatti, Corticocortical connections of area F3 (SMA-proper) and area F6 (pre-SMA) in the macaque monkey, J. Comp. Neurol. 338 (1993) 114–140.
[9] K. Keizer, H.G.J.M. Kuypers, A.M. Huisman, O. Dann, Diamidino yellow dihydrochloride (DY-2HCl); a new fluorescent retrograde neuronal tracer which migrates only very slowly out of the cell, Exp. Brain Res. 51 (1983) 179–191.
[10] M. Kifer, G. Lausen, J. Wu, Logical foundations of object-oriented and frame-based languages, J. ACM 42 (4) (1995) 741–843.
[11] S.H. Koslow, Should the neuroscience community make a paradigm shift to sharing primary data? Nat. Neurosci. 3 (2000) 863–865.
[12] A. Sheth, Changing focus on interoperability in information systems: from system, syntax, structure to semantics, in: M. Goodchild, M. Egenhofer, R. Fegeas, C. Kottman (Eds.), Interoperating Geographic Information Systems, Kluwer Publishers, Dordrecht, 1998.
[13] K.E. Stephan, K. Zilles, R. Kötter, Coordinate-independent mapping of structural and functional data by objective relational transformation (ORT), Phil. Trans. R. Soc. London B 335 (2000) 37–54.

Gully A. P. C. Burns is a research assistant professor at the University of Southern California working as the chief developer of the NeuroScholar system in the laboratory of Larry Swanson. His research is concerned with understanding the large-scale organization of the brain by analyzing patterns of connections between brain structures. This involves theoretical research into databases and data-mining in order to be able to quantify, organize and then analyze data describing the neuronal circuitry in a mathematically tractable way.

Klaas Enno Stephan studies medicine at the Heinrich-Heine University Düsseldorf and computer science at the University of Hagen. Working at the C. & O. Vogt Brain Research Institute in Düsseldorf with Rolf Kötter and Karl Zilles as well as with the Neural Systems Group of Malcolm Young at Newcastle, his research focuses on computational approaches to analysis of structural and functional connectivity in the brains of macaque and man.

Ludäscher is an assistant research scientist at the San Diego Supercomputer Center at the University of California, San Diego. His main research areas include integration of heterogeneous information, especially knowledge-based mediation in scientific databases, and database theory. He studied Computer Science in Germany and received his M.Sc. and Ph.D. from the Universities of Karlsruhe and Freiburg, respectively.

Amarnath Gupta is an Assistant Research Scientist at the University of California San Diego. He is a member of the Data Intensive Computing Environments group at SDSC. His research interests are in heterogeneous information integration, scientific and multimedia data modeling, and spatiotemporal data management. Amarnath received his Bachelor of Technology from the Indian Institute of Technology, Kharagpur, a Master of Science in Biomedical Engineering from University of Texas, Arlington, and his Ph.D. (Engineering) degree in Computer Science from Jadavpur University, India.

Rolf Kötter studied medicine and computer science in Germany, Britain and France. He leads the Computational Systems Neuroscience group at the Center of Anatomy and Brain Research, Düsseldorf University, to establish structure-function relationships in the brain by experimental and computational approaches. Supported by DFG LIS 4-554 95 (2) Düsseldorf.