ISSN 1539–2791

Volume 1 • Number 2 • 2003

Neuroinformatics

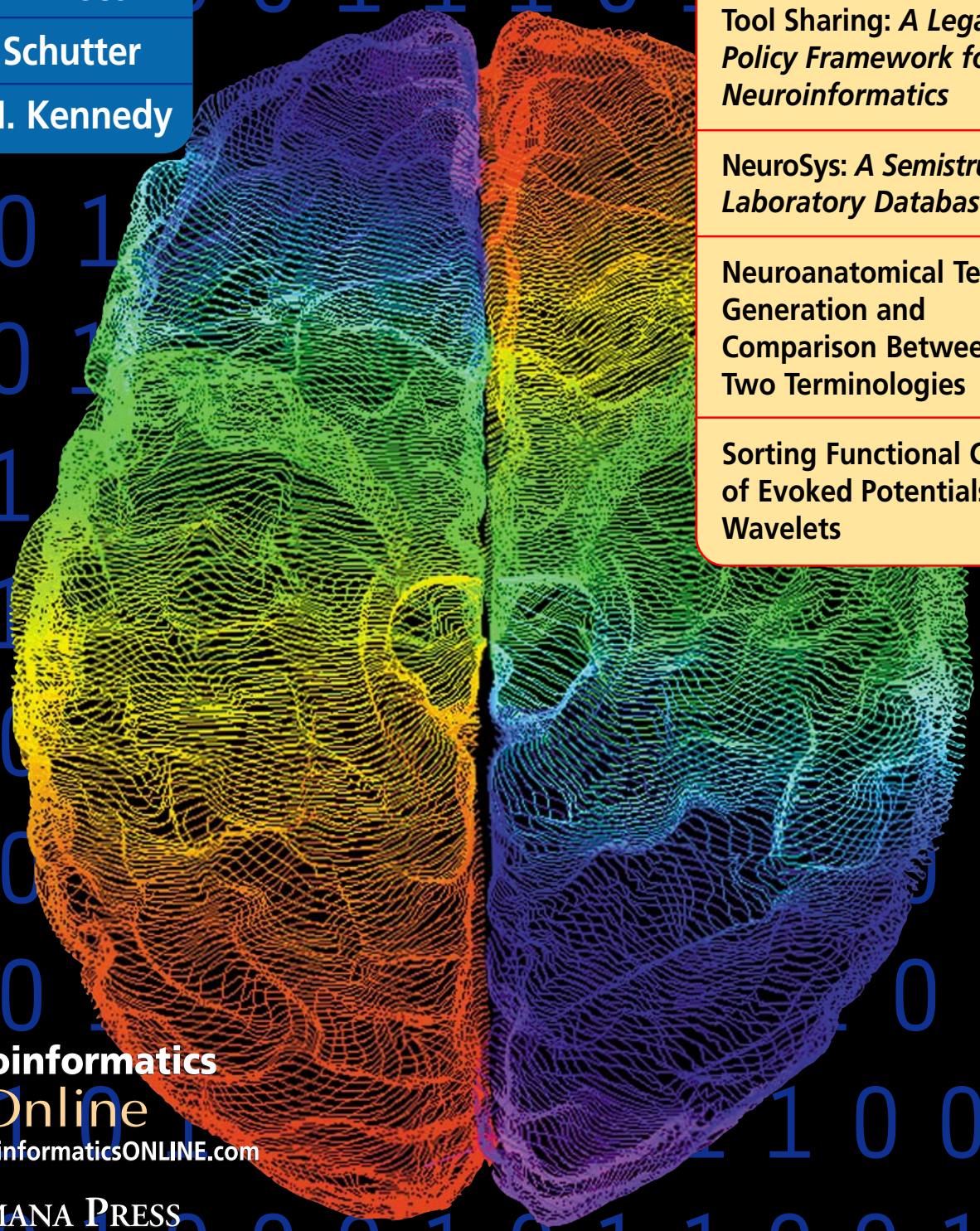
Editors

Giorgio A. Ascoli

Erik De Schutter

David N. Kennedy

IN THIS ISSUE



Humana Press

Neuroinformatics Journal

ISSN 1539-2791/03/01:149–166/$20.00

Public Policy Forum

Neuroscience Data and Tool Sharing

A Legal and Policy Framework for Neuroinformatics

The OECD Working Group on Neuroinformatics: Peter Eckersley,[a] Gary F. Egan,[*] Shun-ichi Amari, Francesco Beltrame, Rob Bennett, Jan G. Bjaalie, Turgay Dalkara, Erik De Schutter, Carmen Gonzalez, Sten Grillner, Andreas Herz, K. Peter Hoffmann, Iiro P. Jaaskelainen, Stephen H. Koslow, Soo-Young Lee, Line Matthiessen, Perry L. Miller, Fernando Mira da Silva, Mirko Novak, Viji Ravindranath, Raphael Ritz, Ulla Ruotsalainen, Shankar Subramaniam, Arthur W. Toga, Shiro Usui, Jaap van Pelt, Paul Verschure, David Willshaw, Andrzej Wrobel, and Tang Yiyuan[b]

[a] Department of Computer Science & Software Engineering, and Intellectual Property Research Institute of Australia, The University of Melbourne. E-mail: pde@cs.mu.oz.au. Eckersley's work on this project is supported by IPRIA funding from IP Australia.

[b] For the institutional affiliations of the other co-authors and members of the OECD working group on neuroinformatics, please see the end of this paper.

Abstract

The requirements for neuroinformatics to make a significant impact on neuroscience are not simply technical—the hardware, software, and protocols for collaborative research—they also include the legal and policy frameworks within which projects operate. This is not least because the creation of large collaborative scientific databases amplifies the complicated interactions between proprietary, for-profit R&D and public "open science." In this paper, we draw on experiences from the field of genomics to examine some of the likely consequences of these interactions in neuroscience.

Facilitating the widespread sharing of data and tools for neuroscientific research will accelerate the development of neuroinformatics. We propose approaches to overcome the cultural and legal barriers that have slowed these developments to date. We also draw on legal strategies employed by the Free Software community, in suggesting frameworks neuroinformatics might adopt to reinforce the role of public-science databases, and propose a mechanism for identifying and allowing "open science" uses for data whilst still permitting flexible licensing for secondary commercial research.

Disclaimer:

This paper reflects the opinions and positions of the authors and is not an official policy or opinion of any national government or organization.

[*] Author to whom all correspondence and reprint requests should be sent.
E-mail: G.Egan@hfi.unimelb.edu.au

Introduction

The development of neuroinformatics promises to extend several important trends in scientific research into the practice of neuroscience. The most visible is the integration of data into large, international databases (and the accompanying creation of a culture of data sharing[1]); another is the growth of research projects beyond the scale of a few laboratories. A more fundamental change is the virtualization of the process of inquiry, enabling questions to be answered with reference to secondary data rather than first-order empirical investigation.

Aside from the direct implications that these changes may have for the nature and scope of scientific research, they also create important questions about policy and the organization of international collaboration.

Perhaps the clearest precedent for the transformative effect neuroinformatics might have on neuroscience can be found in the field of genomics. There, the Human Genome Project (HGP) and other international online collaborations have already come to fruition, and may carry important lessons for neuroinformatics policy.

One of the most visible features of the HGP was the entanglement between "open science" and the proprietization of information. The well-intentioned but simple strategy of placing all HGP data in the public domain has resulted in processes of proprietization—such as the patenting of genes[2] and the creation of privately-held databases—from which the original researchers receive little or no benefit and over which they have no control.

Certain "intellectual property" (IP) laws—or, more specifically, copyright, patent, and database control laws—have played a growing role in science over the last two decades (Dam, 1998). But the nature of electronic information access and exchange, in particular for large-scale, collaborative online research (such as neuroinformatics) makes questions of information privatization more pressing and potentially problematic.

Small research projects can afford to set rules through flexible negotiations which allow participants to obtain patents, for example, or to pass technology into the public domain.

Large international collaborations, on the other hand, need to have clear policy ground rules, for a number of reasons. Firstly, negotiation between all parties is highly impractical, and in a complex research area, a thicket of proprietary patents (for example) can create "anticommons" effects which hamper the progress of the field (Heller & Eisenberg, 1998). Furthermore, taxpayers and benevolent contributors require guarantees that the benefits of research will flow as widely as possible,[3] and funding agencies require clear policies for addressing these issues.

Whilst the short-term impact of patents and proprietary databases on open neuroscience research may be relatively small, in the long term, effects may be significantly larger.[4]

[1] Although universal data sharing has been adopted in other fields such as physics and genomics, the issue remains controversial in neuroscience; see, for example (Aldhous, 2000; Koslow, 2000; Nature, 2000; Koslow 2002).

[2] Either as nucleotide sequences themselves, or indirectly through the proteins they code for.

[3] For an extended discussion of the complicated interactions between the public interest and public policy on the proprietization of research, see (Eisenberg & Rai, 2001).

[4] An initial indication of the role of intellectual property privileges in the commercialization of neuroinformatics is the Brain Resource Company (http://www.brainresource.com), whose promotional material states that they have obtained software and a "large quality controlled database of normative subjects and with a range of clinical disorders," and that they provide fee-for-service analysis reports to clients including researchers, clinicians, and pharmaceutical companies, as well as for medico-legal purposes.

Therefore, anticipating issues arising from the relationship between public and private contributions to neuroinformatic resources, and constructing appropriate policy frameworks from the outset is essential.

In this paper, we consider what regulations, guidelines, and organizations might comprise suitable "policy frameworks" for international neuroinformatics projects, and how they might serve to balance the various competing interests in neuroinformatics research. So, rather than being technical in a computer science sense, our proposals include a degree of legal technicality—suggesting ways in which the law might be used to provide infrastructure to further the progress of research in neuroscience.

The Current State of Neuroinformatics Policy Development

The Neuroinformatics Working Party of the OECD Megascience Forum (OECD, 1999)[5] reported that, "the scientific goals of neuroinformatics are to accelerate the progress of neuroscience and informatics by: making better and more efficient use of neuroscience data using informatics-based, including computational, approaches; generating and evaluating new hypotheses and computational theories about brain function to drive further experiments; and developing and applying new tools and methods for acquiring, visualizing, and analyzing data important for understanding how the brain functions."

They proposed working towards these goals by the implementation of a number of critical steps including: "Establish the coordination, standardization and interoperability requirements needed for successful application, integration, stabilization and quality assessment of the distributed and local neuroinformatics facilities; and enhance collaborative opportunities in neuroinformatics, both nationally and internationally."

The successful development of neuroinformatics will be dependent on achieving these implementation steps. Establishment of appropriate software development and licensing models for sharing data and tools, as well as clarification of copyright issues pertaining to sharing publicly funded research results via inter-networked databases, is urgently required.

In the initial programs started in the United States under the aegis of the Human Brain Project (HBP), grantees were advised to protect their newly developed informatics capabilities, databases, and analytical tools through the copyright mechanism, and then, where possible, to openly share these resources. However, Grantees are also encouraged under this funding, under the provisions of the Bayh-Dole Act, to patent and sell the products that are developed from federally funded research.

In their recently published report on neuroimaging databases, the international society representing the functional neuroimaging community, the Organization for Human Brain Mapping (OHBM) identified the following questions (OHBM, 2001):

1. What rules should govern the use of data derived from a public database and who has the right to publish findings based on these data and within what time frames? How should credit be assigned?

[5] The Organization for Economic Co-operation and Development (OECD) assists in the development of market economies in its 30 member countries from the industrialized and emerging nations. The OECD established the Global Science Forum (GSF) to foster cooperation in global large science programs and issues, and recognizing the need for cooperative efforts in neuroinformatics, established a Working Group on Neuroinformatics (WG-NI). The WG-NI included scientists and policy officials from 24 of the OECD's member and observer governments, and the recommendations of the group were recently published by the OECD at http://www.oecd.org/pdf/M00033000/M00033112.pdf.

2. What rules should protect the confidentiality of experimental participants, and how can these be kept in alignment with local institutional review board regulations and informed consent procedures?
3. What mechanisms should be implemented to prevent violations of these rules, what repercussions should ensue for infractions, and how can these enforced?

This discussion paper addresses these questions and related questions concerning software tools used for neuroinformatics research. We consider the particularities of neuroscience collaboration and draw on observations from the fields of genomics and free software development. We consider a number of licensing and software development models, and suggest possible legal and policy frameworks for the neuroinformatics community.

Models for Neuroinformatics Software Tool Sharing

As identified by this group (Amari et al., 2002), software tool sharing is one of the most important aspects of neuroinformatic collaboration, since software plays such a large and complex role in neuroinformatics research. Unless researchers share the same software tools, it becomes more difficult to make precise comparisons between results. This is not to say that all analyses must follow identical procedures, or that innovative new algorithms and techniques should not be encouraged; rather, development should occur with a maximal amount of transparency, extensibility, scalability, accuracy, and reproducibility. Software tool sharing models must navigate the legalities of copyright and liability in order to serve these goals.

The most well-known model for software development is the closed-source, proprietary approach. Characteristically, a firm employs programmers and funds the development of software; copyright in their work is appropriated by contract, or work-for-hire clauses in copyright legislation. The company then attempts to recoup its costs and make a profit by licensing their creations to customers. Although this model is very widespread and includes examples such as Microsoft Windows and Matlab, it has also been widely criticized amongst software developers.[6] Some of the biggest weaknesses of this model involve inflexibility, centrally determined functionality, and sophisticated users' loss of ability to examine and extend source code, and to fix defects in the tools they use. These problems are particularly relevant to applications in scientific research.

Other weaknesses in the proprietary model involve poor allocative and distributive economic efficiency (Shavell & van Ypersele, 2001), and a strong "network externality" effect[7] which frequently leads to monopolies. In economic terms, software is a naturally non-rivalrous, non-transparent product; it is excludable but only with significant loss of efficiency (DeLong & Froomkin, 2000). As a consequence, the software market is close to a classic example of "market failure." There are a number of other proprietary software development models, such as shareware or application service provider (ASP) approaches, but these also suffer from many of the weaknesses of the traditional "sales" model.

Proprietary software development models are also poorly suited to collaborative tool sharing, since research groups with lower levels of funding may be unable to afford all software that other researchers utilize. There is a strong economic incentive to specialize in the use of a few packages, thus minimizing the number

[6] For frequently cited examples, *see* Stallman, 1992 or Raymond, 1999.
[7] Where the commercial value of a product or tool increases with the number of other users.

of required high-cost purchases. Thus, when new techniques are developed, they receive a lower degree of testing, scrutiny, and duplication.

A different, but highly relevant, software development model is that of "free" or "open source" software. Free software is typically created through collaborations over the Internet, with an ethos that is in many ways similar to that of science, emphasizing the Mertonian norms of universalism, disinterestedness, communalism,[8] and skepticism (Merton, 1973). This culture plays an important (and growing) role in the information technology (IT) industry.[9] Examples of free software include GNU- and Linux-based operating systems, the Apache web server, and the GNOME and KDE desktop environments. A very large amount of scientific software, such as Statistical Parametric Mapping (SPM), Octave, or scientific Python extensions, have been developed using open collaborative models. Examples of relevant indices of scientific software are available from *Scientific Applications on Linux*,[10] *bioinformatics.org*,[11] and *Sourceforge*.[12]

The main disadvantages of free software development models are related to funding and incentives. Various theories are available to explain the widespread creation of free software, relating to the prestige of authoring a widely-used piece of software, and the opportunities for cross-promotion (such as for consulting services).[13] In general however, free software development is an economic "public good," and, as such, is subject to the "free-rider problem," where the number of people who use the good is much larger than the number who contribute (financially or materially) to its production.

In a scientific research environment where a substantial fraction of funding comes from government grants, this problem is circumvented. The status quo is that many groups develop free software tools as an aside to their main research focus; this practice benefits the entire international research community, and should be encouraged.[14] The limitation of the model is that researchers may have insufficient incentive to document, offer support, and integrate such tools.

The copyright and liability issues involved in this form of software development have already been thoroughly explored, and there are a range of free software licenses available which give contributions clear legal status. So-called "copyleft"[15] licenses may be the most desirable, in terms of guaranteeing that public funding for scientific software results in public benefit. There is also the additional option of dual-licensing software, which has

[8] Which Merton, perhaps confusingly, referred to as "communism."

[9] *See*, for example: Shankland, S., "Linux sales surge past competitors," CNET News, Feb 2000. (http://news.cnet.com/news/0-1003-200-1546430.html), or the Netcraft Web Server Survey (http://www.netcraft.com/survey/).

[10] http://sal.rising.com.au

[11] http://bioinformatics.org

[12] http://sourceforge.net/softwaremap/trove_list.php?form_cat=97

[13] On the question of an economic explanation of free software production, *see* (Lerner & Tirole, 2000; Kuan, 2000; Kelty, 2001).

[14] Indeed, while DiBona et al., (2000) suggest that there are many similarities between free software development and science, Kelty (2001) claims that the free software movement is really a part of the scientific enterprise.

[15] The term "copyleft" originated as a play on words from "copyright." It refers to the use of powers granted by copyright law, in order to provide benefits for the public, rather than copyright holders. These powers are chiefly used to constrain the private appropriation of derivatives of the copyleft work (*see* http://www.fsf.org/copyleft for further details).

potentially lucrative non-academic applications. Dual (or "two tiered") licensing, which has been used for successful software such as the QT GUI framework[16] and the MySQL database,[17] allows otherwise copylefted tools to be sold to firms that wish to develop proprietary systems with them. The success of the free software movement in creating widely used and successful software tools by collaboration over the Internet sets a useful precedent for the field of neuroinformatics.

At the level of individual laboratories and research projects, we urge scientists to release any software they are developing under free software licenses (such as the GNU General Public License), and, where possible, to choose free software tools for their work.

At a broader policy level, we suggest that a good strategy for encouraging software tool sharing would be to allocate resources for the application of larger-scale software engineering practices and resources to pre-existing free scientific tools. Such a policy would result in a more standardized, more organized array of tools for international electronic collaboration. There is a role for both national science institutions and international neuroinformatics bodies to encourage and organize funding for these activities.

Models for Neuroinformatics Database Sharing

As observed by Amari et al. (2002), the complexity, heterogeneity, and contextual nature of neuroscience data means that the technical task of creating neuroinformatics databases is daunting. Nonetheless, there are also significant scientific cultural and political obstacles to primary data sharing that slow the construction of such databases. This is a conundrum, because the universal use of databases would provide more data and stronger incentives to accelerate neuroinformatics database research. Once shared database technology and content has matured, it is realistic to expect that policies such as those adopted by the genomics community (DOE-NIH, 1993) would be inevitable.

The experience of bioinformatics has also shown that the IP implications of research databases are quite profound. The promise of genetic engineering has seen public research data used as the starting point for numerous private commercial endeavors. Corporations have built proprietary datasets using public domain research databases to improve their information (Eisenberg, 2000); patents have been claimed on uses for genetic information that publicly funded researchers have placed in open databases. This environment has created significant consternation and debate amongst researchers, lawyers, and the wider public; it is uncertain whether a successful balance will be found to match incentives for R&D against the public interest in unfettered research and open markets.

Indeed, the potential for bioinformatic techniques to seed a "thicket" of proprietary rights that inhibit an entire field of research has been recognized not only by academics and policy makers, but also by large pharmaceutical firms. The work of creating a public domain database of Single Nucleotide Polymorphisms (SNPs) is being undertaken by the SNPs Consortium, which is funded by "big pharma" and which employs complicated and non-obvious strategies to ensure that as much information as possible remains unencumbered by patent laws (Marshall, 1997; Eisenberg, 2000; Sunder Rajan, 2002). It is possible that similar strategies to delineate the public domain in neuroinformatics will prove valuable.

[16] http://www.trolltech.com

[17] http://www.mysql.com

As with software, one of the most important policy instruments available for guiding the operation of collaborative databases are suggested or prescribed licensing models for access to these databases. There are a number of means by which such models can determine who is able to access what data, and how they can use that information as the basis for new works.

Database Licensing Models

Databases, in general, attract varying forms of protection under national copyright laws; the status of these laws is complicated, but in some jurisdictions, copyright may subsist in the aggregation of content in a database. In the United States, courts have ruled that there can be no copyright in a database of "mere facts" (Feist, 1991), while in Europe, regardless of the status of databases under national copyright laws, the EU has enacted specific *sui generis* (*Latin*: "in its own class") database rights. There is significant debate as to whether extensive IP privileges in databases are desirable.[18] For detailed discussions of the interaction between evolving database policy and open science, *see* Reichman & Uhlir, 2001 and David, 2001.

The situation for neuroinformatics databases is, however, quite different from that of databases in general. Unlike simple collections, such as the telephone directory in *Feist vs Rural Telephone* (Feist, 1991), which could be said to comprise "mere facts," neuroscientific data includes a great deal of material in which copyright would subsist directly, such as magnetic resonance images. As a result, copyright licenses applied to much of the content in a neuroinformatic database would be practically enforceable in a wide range of jurisdictions.

A second instrument of collaborative research policy is the requirement of a contractual agreement for use of a database. Celera uses agreements such as these to protect its commercial interests,[19] but researchers might choose to employ them in the service of an "open science" model; imposing certain requirements for non-privatization of knowledge might serve both the public interest and the scientific community.

Finally, there is the possibility of using electronic access controls to protect databases. In practical terms, such controls might only be used to support contractual access (requiring agreement before granting accounts) and to protect database quality (requiring authentication before allowing data to be deposited). Stronger access controls, such as Digital Rights Management systems would be both extremely difficult to implement and inappropriate for consideration in a scientific setting.

There are a number of broader objectives which all of the licensing models should attempt to achieve; allowing access to the research community, private corporations and the public; guaranteeing scientists credit for their contributions; preventing the uncompensated privatization of public knowledge; and maintaining incentives for commercial R&D. We now consider a number of database licensing models which might be adopted, and evaluate them by comparison with these goals.

Public Domain Databases

This is the traditional model adopted by "open science," which provides free public access to information and is suitable for scientific collaboration. However, it may be disadvantageous for researchers, because it does not guarantee them credit for their data under all circumstances, and does not provide them with any reward if extensions of their work are commercialized.

[18] *See*, for example, the report of the US National Research Council (NRC, 1999), (Steele, 1996) or (Benkler, 2000).

[19] *See* Celera Genomics, Terms & Conditions for use, http://www.celera.com/company/terms_conditions.cfm.

Databases Covered by Copyright Law (to the extent permissible by national laws, but with open access granted through a copyleft license)

This model retains all the benefits of public domain databases. It has the additional advantage that scientists will be in a position to negotiate for royalties in a situation where a private firm wishes to make proprietary extensions to a database (since they will need a special license agreement). However, it provides no "guarantee of credit" to the submitters of data, provides only weak protection against "patent thickets," and grants scientists no reward for patents on extensions to their work.

Contractual Access to Copyrighted Databases

In this model, the database is covered by copyright with a copyleft license, but in addition, *access* to the database is provided only to account holders who have agreed to a contract with the database administrators. This contract would include a number of relevant provisions: a credit allocation requirement (*see* "Databases, Authorship, and Credit"). The user is prohibited from creating proprietary derivations, such as new databases, software or patents, using the database. If a user wishes to do these things, they may negotiate a special contract that allows them to do so. The user is prohibited from passing information from the database to other parties *for the purpose of circumventing the other terms of the contract*. Note that redistribution of the data is otherwise permitted by the copyleft license.

A contractual arrangement of this sort would, in practice, provide a very good guarantee of credit for researchers. It does not provide a watertight guarantee of reward in the case of patents arising from secondary research, since once a third party obtains the data, they are not actually bound by the terms of the agreement (this could only be achieved by a far more stringent contract which would significantly complicate some legitimate scientific uses). Nonetheless, it does provide a significant degree of protection, and commercial users would in general be unwilling to risk planned circumvention.

Administering Databases Using a Combined Contractual-Copyleft Model with Dual-Licensing

The "contractual access to copyrighted databases" model would add an additional (small) degree of complexity in administering database access, because under most jurisdictions, contracts are only enforceable where there is clearly demonstrated agreement, understanding, and negotiability (Burk, 2000). A scientific database might be able to meet these criteria by means of the following:

- Allowing a prospective user to request an account subject to the agreement;
- Asking a few simple questions about the nature of the contract;
- Offering an alternative channel for negotiations if they would prefer different terms (such as a fee-based agreement allowing use of the database in developing proprietary patents).

Such a system would be analogous to the "dual-licensing" model used by some free software developers and provides a very elegant mechanism for distinguishing between "public good" and commercial uses,[20] but there are still a number of unresolved issues to consider. Would proprietary licenses be available by negotiation with the organization administering the database, or only by direct discussion with the original researchers? In the former case, royalties given to the administrative body

[20] A recent paper by Reichman and Uhlir comes to many of the same conclusions about database licensing as we do; their work attempts to resolve the tension between open science and commercializability by analogizing the GNU Lesser General Public License (http://www.gnu.org/copyleft/lesser.html), a weak form of copyleft (see particularly (Reichman & Uhlir, 2001) p. 320). We believe, however, that the dual licensing model provides stronger protection for the public interest, clearer delimitation between use cases, and greater certainty that researchers will be remunerated when their work is commercial-

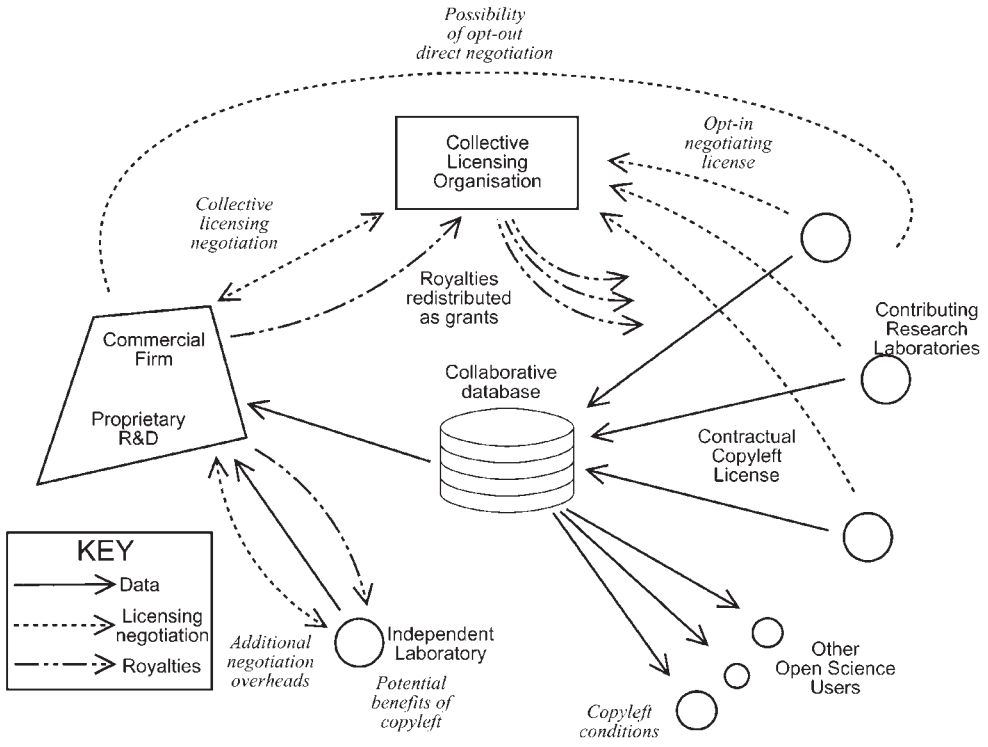


Fig. 1. Structure of the proposed copyleft/dual-licensing arrangement for neuroinformatics databases. Color figure available online.

would need to be redistributed back to the researchers, perhaps through some granting structure. In the latter case, unless client firms could find a practical way of negotiating with the researchers as a group, proprietary extensions of a large fraction of the database might be difficult to arrange. Whether such arrangements might be desirable is a complex policy question in and of itself.

Accountability for Collective Licensing Organizations

If the administrators of a database or their funding agencies are given responsibility for negotiating the terms for proprietary licenses (we will henceforth refer to such organizations as Collective Licensing Organizations, or CLOs), there are a number of potential "moral hazards"[21] to be addressed, and a need exists for effective guarantees of accountability. In particular, scientists need confidence that their interests will be securely represented in any such negotiation.

Guarantees of CLO accountability to scientists are unlikely to be robust if all rights to data are automatically assigned to the CLO. In order to make accountability systematic, we propose a mechanism in which there is collective licensing, but individual researchers or laboratories can withdraw if they believe their interests are not being represented.[22]

This "collective licensing with opt-out" scheme allows efficiencies of scale in negotiation, but preserves the sovereignty of each contributor (*see* Fig. 1). We believe this addresses

[21] Moral hazards are said to occur when actors have net incentives to act contrary to their appropriate role in an agreement (or organization).

[22] It is realistic to expect that different contributors will have quite different expectations about the licenses issued on their behalves. Some might seek a maximum rate or return from commercialization, while others might wish to guarantee flows of knowledge back to open science, or apply various ethical criteria to secondary research. For a CLO, the logical approach to this problem would be to categorize different positions, and record these as metadata annotations in the relevant databases.

many of the potential "institutional design" pitfalls associated with creating CLOs.

Databases, Authorship, and Credit

One of the concerns widely regarded as an obstacle for collaborative data sharing is that of credit for one's data. Many researchers feel that if they publish their primary experimental results along with their first paper, better resourced groups might pre-empt further publications. Advocates of data sharing have suggested that the additional expertise of the original researcher with regard to this data would be sufficient to prevent this from being problematic (Koslow, 2000; Nature, 2000). Furthermore, these analyses may underestimate the incentives that primary data publication would create for forming collaborations, to the direct benefit of the original author. The OHBM has identified credit allocation as a key question to be addressed for sharing neuroimaging data (OHBM, 2001). There are a number of different approaches that could be taken, including exclusion, citation, and co-authorship.

Exclusion

The original author has a fixed or negotiated "window of privilege" during which other researchers may not publish using that data. This has the serious disadvantage of slowing and complicating the research process, and would act as a disincentive for distributed collaboration.

Requirement of Citation

A citation must be made to the original researcher's paper. One problem is that some data may not have an associated publication to cite; this could be solved through the citation of special technical reports describing experimental techniques.[23]

Co-Authorship

Secondary data users could be required to offer co-authorship to the original researcher. Some journals have authorship requirements that might make this difficult, but in many cases, this approach could foster stronger collaboration as well as allocating credit and would enhance the re-analysis of the data if the original data producer was to be involved in the new interpretation.

Such policies could be made a condition of access through database licenses. On balance, it would appear that the third option, perhaps falling back to the second in situations where journal policies are incompatible, would best serve the interests of the neuroinformatics research community. Other options to be considered could be a separate citation byline on the title page listing the data source and the original authors.

Enforcing License Conditions

The OHBM (OHBM, 2001) identifies enforcement of access rules as an important issue to be addressed. Organizations such as the Free Software Foundation already exist in part to provide legal support for software developers who might need to prosecute violations of free software copyright licenses. Collaborative neuroinformatics software may obtain some degree of protection from such organizations.

If more specialized contract/copyleft access models were deployed for neuroinformatic databases, the organization administering the database would probably have to take on some responsibility for preventing misappropriation. The experience of the free software world has been, however, that when the community "cries foul" over license violations, the situation is often rectified before legal action is considered.[24]

[23] Preferably, a scientific paper should have been published describing all of the data which is inserted into neuroinformatics databases; in cases where this is not possible, at least a technical report describing experimental techniques should be published.

Privacy Issues in Neuroinformatics

Restrictions on the Transfer and Processing of Personal and Sensitive Data

Research projects have obligations (both legal and ethical) to guarantee the privacy of their subjects. This is especially true in fields where sensitive information about a large number of people is collected in well-indexed databases.

In some countries, privacy laws are relatively weak; the United States, for example, requires only that organizations collecting personal information provide citizens with the ability to "opt-out" of having those organizations share their data with third parties. At a global level, however, privacy regulations are expanding rapidly (White & Case, LLP, 2002), and it is probably wise to base policy analysis for international collaboration on the strongest applicable laws.

The European Union, for example, places strict constraints upon the sharing of "sensitive data," which includes data about racial or ethnic origin, and hence would probably include much of the information in neuroscientific databases.[25] In general, such information can only be processed by neuroinformaticians for non-clinical purposes if the subject has given their explicit consent to this processing.[26] The EU also restricts the export of data to states which do not have strong privacy laws, although the use of appropriate contracts will allow exports to parties which agree to abide by EU-equivalent regulations.[27]

Individual research groups, or national research administrations, could facilitate data sharing which is compatible with privacy laws worldwide, by obtaining informed consent from experimental subjects, for the use of (anonymized) sensitive data about them, for research purposes. The process of drafting and employing these consent agreements at the laboratory or even the national level may be slow, however, because of the need to ensure compatibility between privacy standards and thus enable the exchange of data.

Our suggestion is that steps should be taken to draft and employ consent agreements at an international level, which are compatible with strong privacy laws such as those found in the EU. Such a program dovetails well with our recommendations about database licensing, for a number of reasons. Firstly, there is already a contractual framework in place for access to the databases, within which privacy rules can be included. Secondly, the dual-licensing

[24] A small number of copyleft license violations have been identified by people examining proprietary executable programs and identifying similarities between these and GPLed software. For some examples, *see* http://slashdot.org/search.pl?query=GPL+violation&sort=2 . It seems unlikely that significant violations could go unnoticed for long periods of time, especially within a scientific research community. Moglen (2001) describes the process and success of GPL enforcement in more detail.

[25] There are slightly weaker requirements for data which is personal but does not fall into the "sensitive" category; however, considering the fact that neuroscientific data will often be very revealing, and that some nations may apply strict regulation to both classes of data, our analysis follows the more restrictive case.

[26] *See* Article 8 of EU privacy directive (European Parliament, 1995). The exception given in Article 8, § 3,which covers clinical uses of data, is probably not broad enough to allow neuroinformatic research. Also note that the directive does allow European states to enact national laws which both grant additional permissions for various uses of data, and laws which limit the extent to which individuals can give consent for certain uses.

[27] *See* the European Commission Decision on standard contractual clauses for the transfer of personal data to third countries,
http://europa.eu.int/comm/internal_market/en/dataprot/news/clauses2.htm

scheme fits well with the requirement of informed consent by subjects; if licensing is clearly separated between non-proprietary, research purposes, and separately licensed commercial uses, the subjects can grant consent to either or both of these classes of usage.

Provided the meanings of the different licensing cases are succinctly explained, and emphasis is placed on the fact that data will be anonymized and legally protected, it is likely that most people will grant permission for both forms of neuroinformatics. Furthermore, those subjects who are unwilling to allow private R&D with their data will still be able to contribute to open scientific research.

Anonymization of Neuroscientific Data

Any open, large scale, international databasing of neuroscientific data must necessarily be based on the principle that, although intimate and sensitive information is available about numerous individuals, this information is insufficient to be linked to their identities.

Thus, although the database might include anatomical, demographical, and pathological information, it should never contain names or links to other databases which contain identifying information.

It should be noted that some neuroscience data types, such as MRI images, contain inherently identifying information about facial and cranial structure.[28] Attempts could be made to perform cropping or transformation, so as to remove the unique link to the subject's identity. An extended approach might be to hold entire raw datasets in escrow, but make only defined subsets directly available through the database. If it transpired that the available subsets were insufficient for required analyses, then alternative access to the full dataset could be negotiated.

Beyond the point of removing bone structure and registering neuroanatomy, such a strategy, however, becomes more problematic because it is unlikely that transformations could absolutely conceal identities without virtually destroying all structure of scientific interest. Fortunately, it may not be necessary, because although transformed brain tissue images are inherently identifying, they only possess this property when the party seeking to identify a subject already has knowledge of the anatomy and pathology of particular individuals. This fact appears to constrain the privacy problems inherent in neuroinformatic research.

So then, the key challenges for guaranteeing anonymization in neuroinformatics databases are ensuring that no combinations of data are stored which together become identifying, and that any potential attempt at identifying a subject would have to go to inordinate lengths—such as obtaining brain scans from other sources—in order to recognize a particular individual. Once these steps are taken, subjects should be informed of these principles when consenting to the use of their data.

A recent example of a neuroimaging database partially addressing these issues is the fMRI Data Center which has been established, "to help speed the progress and the understanding of cognitive processes and the neural substrates that underlie them by: providing a *publicly accessible* repository of peer-reviewed fMRI studies; providing *all data* necessary to interpret, analyze, and replicate these fMRI studies."[29] This is a public domain

[28] Although the authors are not aware of any tools for identifying individuals based on MRI data, the significant body of biometrics literature and software for face recognition from video, would suggest that such identification would be achievable. In particular, it is likely that cranial structure could be linked to photographs, and that neuroanatomy could serve as a unique "fingerprint" for identifying individuals.

database, which requires users who wish to publish data to credit the authors of the original study and acknowledge the Center and the accession number of the dataset. The Center requires contributors to remove subject identifying tags, and the Center also "anonymizes" high resolution structural MR data by removing non-brain tissues. However, the Center does not require users to explicitly agree to the contractual terms of access.

Summary

Although the complexity and heterogeneity of neuroinformatics make the establishment of widespread collaboration a daunting task, the potential benefits to the field are huge. Sharing data and tools are perhaps the most important aspects of online collaboration; without this sharing, significant costs and effort recur in every research group.

The scientific R&D community has, in recent years, been pulled in two directions by opposing forces. On one hand, the traditional ethos of "open science" has suggested that information and tools should be shared, and that cooperation should be encouraged as widely as possible. On the other hand, the pressures of funding requirements, governmental policy, and market opportunities have increasingly required scientists to think in competitive terms, seeking to commercialize their work. The presence of commercial incentive structures at times appears to be irreconcilable with "open science." In this environment, the most valuable knowledge is kept secret or made proprietary.[30]

With sufficient foresight and planning, however, it is not impossible for the open scientific community to obtain the collaborative benefits of "open science" whilst simultaneously appropriating some of the economic value of its creations. This could be achieved by simultaneously adopting the copyleft principle in order to maximize the free flow of collaborative information and allow for negotiated alternative commercialization licenses for the private sector. Adopting policy positions that encourage the open sharing of data and tools, and the use of legal frameworks which support "open science" collaborations, are the most effective strategies available for nurturing the field of neuroinformatics.

Acknowledgments

Thanks to David Brennan, Jeff Bizzaro, Richard Stallman, and David Lindsay for their helpful information and feedback on some of the ideas discussed in this paper.

Appendix 1: Glossary of Terms

Application Service Provider (ASP): A firm which services the outsourced IT needs of another organization. This would typically include running software on the ASP's computers to provide services for their clients.

Copyleft: A *free software license* which requires modified versions of the software to have the same license. Thus, copyleft software in some sense belongs to the public, and private entities do not have the right to create privatized derivatives. The most commonly used form of copyleft is the *GNU General Public License*, or GPL.

Dual Licensing: The practice of releasing software under several licenses simultaneously. Copyright holders may choose to use any combination

[29] http://www.fmridc.org/about

[30] For further discussion of this process, and its relationship to IP privileges, *see* Rai, 1999 and David, 2001.

of free (copyleft or otherwise) and proprietary licenses.

Executable Code: The form of software that is used on a computer when a program runs. It is created from the *source code* by a compiler program. *Proprietary software* is usually distributed in executable form only.

Free Software: Allows users a significant number of freedoms including the right to access the software's *source* code, and the right to reproduce, modify, and distribute the software. Free software is typically developed in highly collaborative communities on the Internet. Free software comes in two varieties, *copyleft* software and *non-copyleft free software*. Unlike *public domain software*, free software is often protected by copyright laws.

Freeware: An ambiguous term for software that is available at no cost. Often, freeware is proprietary software available only in executable form. Thus, although freeware may be non-commercial, the author still retains control over the distribution and function of the program.

GNU General Public License (GPL): A strong form of *copyleft;* it applies to the entirety of larger works which are derived from the original. The GPL thus prevents libraries from being linked to proprietary applications.

Non-copyleft Free Software: Free software to which proprietary extensions can be made.

Open Source Software: *see free software.*

Proprietary Software: Software that is owned (and controlled) by a single, usually commercial, entity.

Public Domain Software: Software which is not covered by copyright law; thus, anybody is able to do anything with it.

Shareware: A form of *freeware* where a restricted version of the software is made available at no cost; a fully-featured version is available from the author for a fee.

Source Code: The form in which programmers create software. It is written in a particular programming language, such as C++, Pascal or Java. It is generally feasible for programmers to change a program if they have access to the source code.

References

Aldhous, P. (2000) Prospect of data sharing gives brain mappers a headache. *Nature* 406:445–446.

Amari, S., Beltrame, F., Bjaalie, J., et al. (OECD Neuroinformatics Working Group) (2002). Collaborative Neuroscience: Neuroinformatics for Sharing Data and Tools. *J. Integ. Neuro.* 1:117–128.

Benkler, Y. (2000) Constitutional Bounds of Database Protection: The Role of Judicial Review in the Creation and Definition of Private Rights in Information. *Berkeley Technology Law Journal* 15:535–595.

Burk, D. L. (2000) Intellectual Property Issues in Electronic Collaborations. In: *Electronic Collaboration in Science* (Koslow, S. H. and Huerta, M. F., eds.) Lawrence Erlbaum Associates, Mahwah, NJ, pp. 15–44.

Dam, K. (1998) Intellectual Property and the Academic Enterprise.John M. Olin Law & Economics Working Paper 68, University of Chicago Law School, Chicago, IL.

David, P. (2001) Will Building "Good Fences" Really Make "Good Neighbours" in Science? In: *IPR (Intellectual Property Rights) Aspects of Internet Collaborations* (Granstrand, O., Foray, D., & David, P., eds.) European Commission Research Directorate General. Brussels, Belgium.

DeLong, J. B. and Froomkin, A. M. (2000) Speculative Microeconomics for Tomorrow's Economy. *First Monday* 5(2) http://firstmonday.dk/ University of Chicago, Chicago, IL.

DiBona, C., Ockman, S., and Stone, S., eds. (2000) Introduction, Open Sources: Voices from the Open Source Revolution, O'Reilly & Associates, Sebastapol, CA.

DOE-NIH Guidelines for Sharing Data and Resources (1993) Human Genome News 4(5) 4.

Eisenberg, R.A. (2000) The Public Domain in Genomics. Proc. "A Free Information Ecology in the Digital Environment" conference, New York University School of Law Information Institute, New York, NY.

Eisenberg, R. A. and Rai, A. K. (2001) The Public and the Private in Biopharmaceutical Research. Proc. Conference on the Public Domain, Duke University, Durham, North Carolina, pp. 157–175.

European Parliament and of the Council of the European Union (1995). Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data. European Commission, Brussels.

Feist Publications vs. Rural Telephone Service Company Inc., 499 U.S. 340, 1991.

Heller, M. A. and Eisenberg, R. S. (1998) Can Patents Deter Innovation? The Anticommons in Biomedical Research. *Science* 280: 698–701.

Kelty, C. (2001) Free Software/Free Science. *First Monday* 6(12) http://firstmonday.dk/; University of Chicago, Chicago, IL.

Koslow, S. H. (2000) Should the neuroscience community make a paradigm shift to sharing primary data? *Nature Neuroscience* 3:863–865.

Koslow, S. H. (2002) Sharing primary data: a threat or asset to discovery? *Nature Reviews/Neuroscience* 3:311–313.

Kuan, J. (2000) Open Source Software As Consumer Integration into Production. Unpublished, available from http://freesoftware.mit.edu/online_papers.php

Lerner, J. and Tirole, J. (2000) The simple economics of open source. NBER working paper 7600. National Bureau of Economic Research, Cambridge, MA.

Marshall, E. (1997) Snipping away at genome patenting. *Science* 277:1752–1753.

Merton, R. K. (1973) The Sociology of Science: Theoretical and Empirical Investigations, Chicago University Press, Chicago, IL.

Moglen, E. (2001) Free Software Matters: Enforcing the GPL, parts I & II. Linux User, August-September 2001, Cheshire, UK, pp. 66–67.

Nature Neuroscience editorial (2000) *Nature Neuroscience* 3:845–846.

OECD (1999) Final report of the OECD Megascience forum—Working Group on Biological Informatics, OECD Publications. http://www.oecd.org/pdf/M000014000/ M00014759.pdf, Paris, France.

Governing Council of the Organization for Human Brain Mapping (OHBM) (2001) Neuroimaging Databases. *Science* 292:5522:1673–1676.

Rai, A. K. (1999) Regulating Scientific Research: Intellectual Property Rights and the Norms of Science. *Northwestern University Law Review* 4(1): 77–152.

Raymond, E. S. (1999) On management and the Maginot Line. In: The Cathedral and the Bazaar. http://www.tuxedo.org/~esr/writings/cathedral-bazaar/cathedral-bazaar/x340.html, O'Reilly & Associates, Sebastapol, CA.

Reichman, J. H. and Uhlir, P. F. (2001) Promoting Public Good Uses of Scientific Data: A Contractually Reconstructed Commons for Science and Innovation. Proc. Conference on the Public Domain, Duke University, Durham, NC, pp. 239–322.

Shavell, S. and van Ypersele T. (2001) Rewards versus Intellectual Property Rights. *J. Law. Econ.* 44(2):525–548.

Stallman, R. M. (1992) Why Software Should be Free. http://www.gnu.org/philosophy/shouldbefree.html. Free Software, Free Society: Selected Essays pf Richard M. Stallman (2002), Guy, J. ed. GNU Press, Boston, MA.

Steele, S. (1996) Electronic Frontier Foundation commentary on the "Sui Generis Protection of Databases" treaty proposed for the Diplomatic Conference on Certain Copyright and Neighboring Rights Questions in Geneva. Unpublished but available at http://www.eff.org/Intellectual_property/eff_wipo_19961122.comments.

Stewart, J. E., Mangalam, H., and Zhou, J. (2001) Open Source Software meets Gene Expression. *Brief Bioinform* 2:319–328.

Sunder Rajan, K. (2002) Banking (on) Biologicals: Commodification and Global Circulation of Human Genetic Material. *Sarai Reader* 2:277–289.

US National Research Council (NRC) (1999) A question of balance: Private Rights and the Public Interest in Scientific and Technical Databases, National Academies Press, Washington, DC.

White & Case, LLP (2002) Global Privacy Law: A Survey of 15 Major Jurisdictions. Global Privacy Forum. http://www.whitecase.com/report_global_privacy.pdf.

OECD Neuroinformatics Working Group Members' Affiliations:

Australia

Gary F. Egan
Howard Florey Institute
University of Melbourne
Parkville 3010, Australia
Email: G.Egan@hfi.unimelb.edu.au

Belgium

Erik De Schutter
Born Bunge Foundation
Univ Antwerp-UIA
Email: ERIK@BBF.UIA.AC.BE

China

Tang Yiyuan
Institute of Neuroinformatics
Dalian University of Technology
Email: yy2100@163.net

Czech Republic

Mirko Novak
Faculty of Transportation
Czech Technical University
Institute of Computer Science
Academy of Sciences
of the Czech Republic
Laboratory of System Reliability
Email: mirko@fd.cvut.cz

Vaclav Sebesta
Institute of Computer Science
Academy of Sciences of the Czech Republic
Email: vasek@cs.cas.cz

European Commission

Line Matthiessen
Scientific Officer
European Commission
DG Research B-II-03 - Neurosciences
E-mail: line-gertrud.matthiessenguyader @cec.eu.int

Finland

Iiro P. Jaaskelainen
Massachusetts General Hospital-NMR Center
Harvard Medical School
Email: Iiro@NMR.MGH.Harvard.EDU

Dr Ulla Ruotsalainen
Tampere University of Technology
DMI/ Signal Processing Laboratory
Email: ulla.ruotsalainen@tut.fi

Germany

Andreas V. M. Herz
Humboldt-Universitaet Berlin
Innovationskolleg Theoretische Biologie
Theorie neuronaler Systeme
Email: a.herz@biologie.hu-berlin.de

Klaus-Peter Hoffmann
Universitaet Bochum
Allgemeine Zoologie und Neurobiologie
Email: kph@neurobiologie.ruhr-uni-bochum.de

Raphael Ritz
Humboldt-Universitaet Berlin
Innovationskolleg Theoretische Biologie
Theorie neuronaler Systeme
Email: r.ritz@biologie.hu-berlin.de

India

Viji Ravindranath
Officer on Special Duty
National Brain Research Centre
ICGEB Campus
Email: vijir@vsnl.com

Italy

Francesco Beltrame
Delegate from the Italian Ministry
of University
and of Scientific and
Technological Research
Email: francesc@dist.unige.it

Japan

Shun-ichi Amari
Vice Director
RIKEN Brain Science Institute
Laboratory for Mathematical Neuroscience
Email: amari@brain.riken.go.jp

Shiro Usui
Biol. & Physiol. Eng. Lab.
Dept of Inf. & Comp. Sci.
Toyohashi Univ. of Technology
Email: usui@tut.ac.jp

Korea
Soo-Young Lee
Department of BioSystems
Director, Brain Science Research Center
Korea Institute of Science and Technology
Email: sylee@ee.kaist.ac.kr

Netherlands
Jaap van Pelt
Netherlands Institute for Brain Research
Meibergdreef 33, 1105 AZ Amsterdam
The Netherlands
Email: j.van.pelt@nih.knaw.nl

Norway
Jan G. Bjaalie
Neural Systems & Graphics Computing
Department of Anatomy
Institute of Basic Medical Sciences
University of Oslo
Email: j.g.bjaalie@basalmed.uio.no

Poland
Andrzej Wrobel
Professor of Neurosciences
Nencki Institute of Experimental Biology
Email: wrobel@nencki.gov.pl

Portugal
Fernando Mira da Silva
INESC - Grupo de Redes Neuronais
Email: Fernando.Silva@inesc.pt

Spain
Carmen Gonzalez
Departamento de Farmacología
Facultad de Medicina
Universidad Miguel Hernández
Email: gonzalez@umh.es

Sweden
Sten Grillner
Nobel Institute for Neurophysiology
Dep. Neuroscience
Karolinska institute
Email:Sten.Grillner@neuro.ki.se

Switzerland
Paul Verschure
Institute of Neuroinformatics, ETH-UZ
Email: pfmjv@ini.phys.ethz.ch

Turkey
Turgay Dalkara
Professor of Neurology
Institute of Neurological Sciences
and Psychiatry
Hacettepe University and Advisory
to the President
Scientific & Technical Council of Turkey
Email: dalkara@tr.net

United Kingdom
Rob Bennett
Board Programme Manager
MRC Neurosciences & Mental Health Board
Medical Research Council
Email: robert.bennett@headoffice.mrc.ac.uk

David Willshaw
Institute for Adaptive and Neural Computation
Division of Informatics
University of Edinburgh
Email: david@anc.ed.ac.uk

United States
Stephen H. Koslow (Chairman)
Associate Director
National Institute of Mental Health, NIH
Director, Office on Neuroinformatics
Email: koz@helix.nih.gov

Perry L. Miller
Director, Center for Medical Informatics
Yale University School of Medicine
Email: perry.miller@yale.edu

Shankar Subramaniam
Professor of Bioengineering
Chemistry and Biochemistry, UC San Diego
Email: shankar@ucsd.edu

Arthur W. Toga
Laboratory of Neuro Imaging
4238 Reed Bldg
Department of Neuorology
UCLA School of Medicine
Email: toga@loni.ucla.edu