1st INCF Workshop

on

Sustainability of Neuroscience Databases

December 13-14, 2007 - Stockholm, Sweden

incf International Neuroinformatics Coordinating Facility

www.incf.org

1st INCF Workshop on Sustainability of Neuroscience Databases

December 13-14, 2007
International Neuroinformatics Coordinating Facility Secretariat
Stockholm, Sweden

Authors

Jack Van Horn and Jaap van Pelt

Scientific Organizer

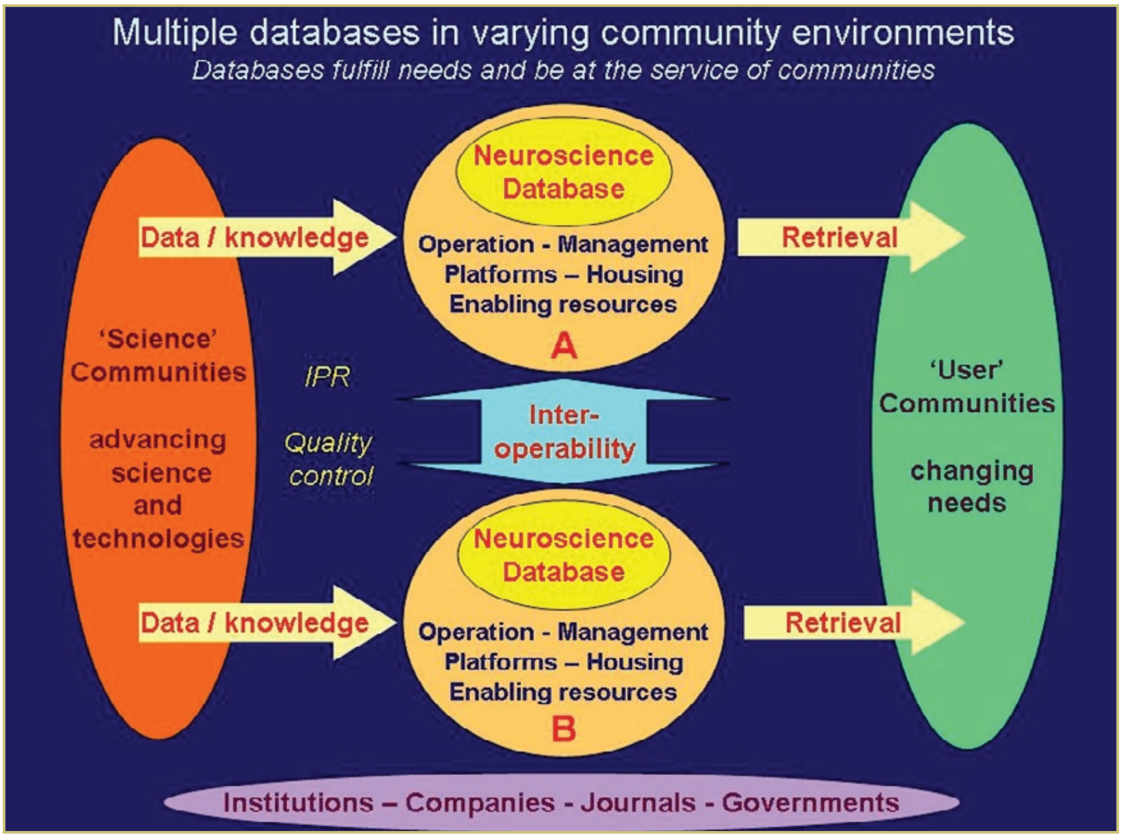
Jaap van Pelt, VU University Amsterdam, The Netherlands

Workshop Participants

Jan G. Bjaalie, INCF Secretariat, Stockholm, Sweden
Chris Emblow, Akvaplan-niva AS, Polar Environmental Centre, Tromsø, Norway
Sten Grillner, Karolinska Institutet, Stockholm, Sweden
Jack Van Horn, UCLA School of Medicine, Los Angeles, USA (Rapporteur)
Tadashi Isa, National Institute for Physiological Sciences, Okazaki, Japan
Martin Kersten, Center for Mathematics and Informatics, Amsterdam, The Netherlands
Wouter Los, University of Amsterdam, Amsterdam, The Netherlands
Alessandro Orro, CNR – Institute of Biomedical Technologies, Segrate, Italy
Roman Mouček, University of West Bohemia in Pilsen, Plzen, Czech Republic
Martin Nawrot, Free University Berlin, Berlin, Germany
Matias Palva, University of Helsinki, Finland
Jaap van Pelt, VU University Amsterdam, The Netherlands
Fiona Reddington, NCRI Informatics Initiative, London, UK
Shankar Subramaniam, University of California at San Diego, La Jolla, USA
Jostein Kandal Sundet, Oslo Central University Computing Center, Oslo, Norway
Shiro Usui, RIKEN, BSI, Wako, Japan
Ilya Zaslavski, University of California, San Diego, USA

Supported by the EU Special Support Action INCF, the INCF Central Fund, and the Swedish Foundation for Strategic Research

Contents



Multiple databases in varying community environments

Databases fulfill needs and be at the service of communities

Figure 1. Sustainability issues of interoperable databases in varying community environments.

1. Executive Summary

Neuroscience databases broaden and extend the scope of both published and non-published data by making them widely available to focused or open communities. Data organized in these fashions are searchable, viewable, and suitable for secondary exploration far beyond the purpose of their original collection. Governmental scientific agencies encourage the development and use of these resources but are not interested in long term support, per se. Once a database has been developed, questions arise as to who benefits from the resource, how it is being maintained, what data model underlies its organization, and whether it is interoperable with other resources. Failure in any of these areas may mean that the database cannot be sufficiently sustained as a resource. The sustainability of these endeavors is of paramount importance to the field of neuroscience and the INCF is uniquely positioned to examine how databases can maximize their long term sustainability, attract users, and provide linkages to other data resources in such a way as to make each one an indispensable component of a larger whole. This meeting sought to explore these issues and make recommendations for the INCF to pursue that would involve rating, ranking, and supporting database sustainability.

2. Introduction

Neuroscience databases provide a diverse collection of communities with access to raw and meta- data, analytical tools and computational models for use in new research, methods development, and science education. The development of these valuable resources often requires several years of active and focused effort in order to meet the needs of a particular research collaborator or to be able to be used by the neuroscience research community. Building the database architecture, populating it with content, and linking that content to content residing in other internet-enabled resources increases the overall richness of those resources. The mining, re-analyses, and visualization of data from archives of data from previously published research articles enables independent validation, new methods development, use in education, and novel neuroscience outcomes.

Nevertheless, how these databases sustain their activities in the long term remains a question without a clear or satisfactory solution. Many variables come into play when considering database sustainability, not the least of which is ongoing governmental support for database curation and tools development. Others include an engaged process of curation, systems support, and active scholarly activity that draws from the resource. Moreover, the fact that people actually use these resources to conduct novel scientific discovery, and have come to rely on them, means that should they fail to maintain their sustainability, a certain segment of the neuroscience research enterprise (not just the database in question) will be affected. In addition, their usefulness in the training of the next generation of neuroscientists cannot be overstated. Government funding agencies must examine carefully the impact of getting such programs started, what it will take to continue their momentum following their initial construction, and who will be inconvenienced should they falter. Organizations such as the International Neuroinformatics Coordinating Facility (INCF) are well positioned to monitor international database activities, usage, and longevity.

In order to articulate its role in database sustainability, the INCF organized the 1st INCF Workshop on Neuroscience Database Sustainability at the INCF Secretariat, at the Karolinska Institutet in Stockholm, Sweden from 13-14 December, 2007. The goal of the workshop was to discuss issues related to the sustainability of neuroscience databases, identify problems and discuss solutions or approaches to these problems, and formulate recommendations to the INCF. Expert researchers were invited to the workshop from the international neuroinformatics community, as well as other disciplines within biology and clinical medicine where sustainability issues have already been approached and experience obtained.

3. Sustainability Issues – Summaries from Speaker Presentations

Each of the invited participants framed the issues of sustainability from his or her own unique perspective. Collectively, several general themes emerged: the pervasive role that databases are now taking in each domain of neuroscience; their linkage with peer-reviewed publication; the role of governmental funding agencies; the needs for minimal but useful and interoperable standards for data and meta-data file exchange; the needs for business models; the instruments for quality control, and the role of the community and other stakeholders in the endeavor of databases and data sharing—those contributing to and those drawing from databases. In the brief sections that follow, we outline the main points raised by each speaker.

I. Sten Grillner

Databases are important infrastructures, unique and competitive. They deserve commitments by funding agencies for their support, through regular reapplication for renewal and continuity. It is recommended that funding agencies allocate funds for their long-term continuity as they are doing, for example, for funding agency-associated research institutes, or international programs such as particle accelerator facilities (CERN), biodiversity programs (GBIF), or networks of astronomical observatories. Promotion of standards for data sharing will increase community involvement and use, resulting in financial benefits. INCF shall support and take action in web-based community interaction of databases of data, tools and models.

II. Jaap van Pelt – *Sustainability Issues of Neuroscience Databases*

The long-term success of neuroscience databases depends on a variety of factors:

- Sustainability of the database technical facility: The facility must implement interoperability standards (e.g., ontologies and benchmarks), follow technological developments through regular upgrades, and be designed (scale, granularity, central/distributed structure) to accommodate limited lifetime of platforms.

- Sustainability of database content: The scientific quality of database content must be maintained at the highest level, through instruments such as regular quality assessments, carefully evaluating the lifetime of the data and metadata, regular curation of the data, and protection against database degradation. The owner of the data has particular responsibility in this ongoing process. The quality issue also concerns tools for analysis, visualization, and modeling.

- Sustainability of database operation: The facility requires technical staff, maintenance, and user support. Data protection and security require appropriate measures for access rights, copyrights, and detection of misuse and unauthorized use. The facility should optimally meet the needs and requirements of the users and be integrated in scientific practice. Investments are needed in training, promotion and visibility, user friendliness, and documentation, ease of (meta-) data entry/retrieval. User groups must be involved in evaluations through measures of use, applicability, and needs. IPR and ethical issues need to be implemented as well as journal policies.

- Sustainability of database enabling resources: Organizations can have particular roles in providing enabling resources for the data base facilities such as representing user groups (Societies, Journals), application areas (Companies), support and implementation (Institutions, Funding Agencies). Appropriate financial resources are crucial to cover the costs of database accommodation, equipment, management and operation.

III. Shankar Subramaniam – *Interoperability and Data Integration in Neuroscience Databases*

Interoperability of databases sets requirements on data organization and presentation (ontologies, databases, query tools, interfaces), data interoperability (relational rules, data formats, disambiguation) and data integration (heterogeneous data integration, legacy knowledge integration, weighing knowledge association), data analysis (statistical tools, integration with biology, reducing complexity, models), and data curation. These Neuroscience Infrastructures are even more difficult than those in bioinformatics because of the curse of multiple dimensions (anatomy and atlases, models), the modality of measurements, the granularity of description, and shape and topology issues. Available databases may be distinguished from those containing neuroscience-related experimental data that are publicly accessible; those containing neuroscience knowledge, databases of tools and tool registries, or links to neuroscience information portals (e.g., Human Brain Project, BIRN, BrainInfo Sites).

Recommendations:

- Facilitate sharing of data and databases: INCF should encourage interoperable formats for data representation and presentation. The use of standard ontologies is recommended.

- Provide tools for interoperability between data and databases: INCF should encourage creation of data object porting capabilities (using markup languages) and mediators for interoperable querying of data between databases. INCF should encourage query infrastructures with APIs that will facilitate modular usability across diverse data.

- Make interfaces interoperable and uniform: INCF should encourage development of portable visualization interfaces (java applications) that can be used in modular and portable manner.

- Promote data integration in neuroscience: Given the importance of anatomy, INCF should encourage the use of standard brain atlases and mapping tools that will facilitate mapping of diverse data—molecular, cellular, tissue images—onto anatomy. This will provide a three-dimensional perspective of the data mapped to function.

- Emphasize the role of models: INCF should encourage the presentation of models in neuroscience in portable markup language formats that will make it easy to map models to other data. Also, annotation of models needs to be an integral part of the models.

- Encourage curation by community: INCF should create a global curation community using the auspices of journals in neuroscience/informatics for validation of data, models and anatomical correlates. INCF should encourage integration with the neuroscience clinical community to deal with pathology associated with the brain.

IV. Shiro Usui – *J-node Sustainability Scheme Including Government Support*

Sustainability of the INCF Japan-Node is supported by a well-organized scheme of coordination, management, promotion, and user and researcher involvement. This high level of organization also applies to the series of neuroscience topical and focused brain science platforms under governmental support. The Xoops scalable content management system XooNIps provides the technical support with tree-like links (http://xoonips.sourceforge.jp/) as an open source.

V. Tadashi Isa – *National Brain Research Project "Integrating Brain Research" from the Database Committee Point of View*

The Integrative Brain Research (IBR) Project Database distinguishes a top-down type (with neuroscientists and research outcomes), a bottom-up type (with a neuroscientists social networking service (IBR-SNS), and a mouse brain-behavior phenotype database. Sustainability issues concern the questions of how to get more data and more detailed information, how to encourage incentives for uploading the data from individual scientists, how to collect raw data for open access, and how to continue after the research term of the IBR project. Actions are planned to transfer the neuroscientists database and research outcome database to the XooNIPS @ NIJC in RIKEN BSI, and the SNS neuroscientists to some succeeding project of IBR, while collaboration should be maintained with the NIJC and Japan Neuroscience Society. The mouse phenotype database should be maintained in conjunction with the NIJC.

VI. Jack Van Horn – *Business Models for Neuroscience Database Sustainability*

Business models for neuroscience database sustainability need to specify their value elements. These include need and audience (see example of bioinformatics), models (is the database model particularly novel, general, or useful?), usage (does the data resource get used and in what way?), community (who does the database serve and how broadly?), literature (is the data supported by peer-reviewed publications and/or representative of the whole field?), sustainability (by what means are its ongoing activities supported?). Crucial for sustainability are the support models. Examples are databases driven by consortia (such as BIRN requiring membership, ICBM and caBIG™), or databases supported by consortia such as LONI, providing database and computational services to multiple consortia-driven projects (e.g., ADNI and BIRN) and individual labs (requiring collaborative agreements with preference for funded research). Such databases serve the needs of the members of the consortia, but may also serve larger communities with tuned access modes. An issue is what happens to a database when its consortium completes its mission.

Databases must be tied to scientific literature and based on agreed upon "classes of domain-specific metadata". Published articles appear to report methods with increasingly reduced detail, often relegated to "supplementary documentation", resulting in uncertainty on what was done and few if any replications (e.g., neuroimaging). No agreed upon guidelines, format, or framework exist for what should be provided in supplemental documentation. The core set of methods information needs to be identified and it should be expected to be fully disclosed in peer-reviewed articles. An example is the Minimum Information for Neuroimaging Description and Specification (MINDS) (Appendix A1).

Questions concerning sustainability of databases include their linkage to peer-reviewed research, their usage, their impact on generating new research (new papers from old data?), governmental support which is needed in the case of required deposition policies, society sanction (could SfN, INCF, or others sanction databases?), and level of service (basic or enhanced—dependent on free or paid access).

Recommendations:

- INCF can help databases clearly define the target audience benefiting from the resource and why there is a need.
- Encourage DB developers to be open, candid, and realistic about the data models.
- Showcase where and how data from the resource have been used and be prepared to back it up.
- Advocate professional society and organizational sanctioning.
- Encourage that database content be tied to the peer-reviewed literature whenever possible.
- Be vocal about governmental agencies making possible long-term options for funding coupled to level-of-scale user fee structure.
- Help database developers envision a clear "event horizon" for when it may be time to stop curating new data or for letting the database fade away gracefully should it be deemed necessary.

VII. Chris Emblow – *The Society for the Management of European Biodiversity Data and its Role in the Sustainability of Taxonomic Checklist Databases*

The Society for the Management of European Biodiversity Data plays an important role in the sustainability of taxonomic checklist databases. Experiences were obtained with the European Register of Marine Species (ERMS), covering four databases: (1) the register of species (over 30,000 marine taxa), (2) the bibliography of identification guides (840 publications), (3) the register of species identification/ taxonomic experts (600 individuals from 37 countries), and (4) the register of marine reference collections, which were all completed with the unpaid help of many individual specialists of Marine Species. This was made possible by specifying agreements from the contributors (i) to voluntarily provide data, information, opinion, or other expert assistance to the ERMS project, (ii) to retain the right to use and publish any data and intellectual property created by the contributor, (iii) to authorize the project to store, compile, modify, and disseminate data provided and derived by any means (e.g., electronic, World Wide Web, book), and (iv) to recognize that products of the ERMS are the copyright of the project and will not disseminate further ERMS publications or data without prior permission of the project Steering Committee. In addition, the project agreed to (i) acknowledge the contribution of the contributor in publications of the ERMS, (ii) provide the contributor with a copy of the ERMS publications, (iii) establish a new organization to manage the ERMS after the completion of the project on 31st March 2000, (iv) transfer all ownership of data and intellectual property collected as part of this project to the new organization by 31st March 2000, and (v) to ensure that the contributor has the right to elect individuals to the management committee of this new organization.

The Society for the Management of European Biodiversity Data (SMEBD) was established in 2000 as a "not-for-profit" company limited by guarantee and not having a share capital based in Ireland, and with contributors agreements transferred to SMEBD at the end of the project. Its role in sustainability was specified as (i) to act on behalf of its members to manage European Biodiversity data, including the European Register of Marine Species, (ii) to provide a legal basis for the protection of the members' contributed data, (iii) to facilitate communication and interaction between persons interested in biodiversity and its application to environmental management, (iv) to promote the publication and dissemination of information related to biodiversity, (v) to facilitate access to specialist knowledge and scientific opinion on biodiversity, and (vi) to raise funds to further the aims of the society.

In 2004, the Marine Biodiversity and Ecosystem Functioning was initiated as the first EU FP6 Network of Excellence with over 600 scientists, 56 member institutes, 36 associate member institutes, and 24 countries. As a new instrument of FP6 it provided appropriate funding to maintain and update ERMS. Further long-term commitment to host data was obtained from Flanders Marine Institute (VLIZ) in ERMS 2.0.

VIII. Wouter Los – *Maintenance, Sustainability and Management of Databases in the Environmental Sciences*

For data maintenance, management, and sustainability, ownership of distributed data is essential, as is the need to organize different expert communities (at the disciplinary, institutional and regional level). Data integration leads to peer-reviewed authority files.

Data e-infrastructure, including interoperability with other databases and applications, is important to support the data flows according to agreed standards. Biodiversity databases typically contain complex data records of diverse and heterogeneous data. Data are very often generated at distributed (monitoring and collection) sites and may depend on human interpretation. With parallel data taxonomies, data integration and data access require regional and global cooperation.

Recommendations:
Business plans for databases and services:
Increase and organize the ownership of databases

- Consider experts contributing as (co-)authors
- Organize peer reviews
- Track and publish ownership of data to provide proper credits
- Adopt Global Unique Identifiers (GUIDs) to trace uses
- Formalize institutional commitment

Databases are an asset: exploit these accordingly

- Provide mission statements that express such awareness
- Implement common standards and data management practices
- Promote actively international data sharing
- Develop/contribute to new services and products

Organize the components of the database development, maintenance, etc.

Infrastructure development

- Benefit from the potential of INCF

- Design the stratification and allocation of responsibilities in the chain of data flows

- Promote new infrastructure capabilities

IX. Fiona Reddington – *Multi-Disciplinary Data Sharing: A UK Perspective*

What is seen as the informatics problem can also be considered a consequence of success. In databasing and data sharing, one is faced with different types of data, large data volumes, different data standards, inaccessibility of much research data, multiple stakeholders, varying (dis-)incentives and willingness of researchers to share their data, lots of activities while little coordination, and little funding for standards, education or infrastructure projects.

In order to address these issues the UK National Cancer Research Institute (NCRI) initiated a partnership and developed a data sharing policy that could be a model for other disease-related data. This policy is shared among the partners of NCRI and builds on existing resources and standards, recognizing that this will be easier in some areas than others, leading to phased implementation. It empowers the research community and recognizes resource requirements.

NCRI can help identify and promote standards (maintaining and promoting a "planning matrix"), deliver infrastructure to support data sharing, and support demonstrator projects and engage in/foster debate on key topics (such as hosting workshops, publications and conferences, presentations at external events).

Future actions focus on (i) the promotion of national and international collaboration (by linking the UK cancer research community, unifying cancer and non-cancer research teams, irrespective of size, frees up researchers to focus more on the science, integration of connected resources globally, and building a flexible but consistent infrastructure for future sources/ services), and (ii) fostering a data-sharing culture (by making resources more accessible, reusing resources (data and services) across multiple domains, reducing duplication thus saving time and effort, and adopting data standards and data sharing.

Summarizing, the NCRI Informatics Initiative is helping to work with the UK and international community to build reusable infrastructure, and to change culture to make the infrastructure usable. NCRI anticipates that its approach, being open and voluntary, is in everyone's best interest.

Recommendations:

- Focus on communication
- Share responsibilities with stakeholders
- Involve end users
- link NCRI with INCF and other international collaborators

X. Jostein Sundet – *Data Exchange in International Collaborations*

Success in data exchange in international collaborations depends crucially on the services available for the users. The research computing services offered at the University of Oslo facilitate researchers in providing access to (large) storage pools, in computation and visualization. It is important that services are offered at the IT-training and knowledge level of the researchers. Challenges in data exchange include redundancy of data, authentication, locality of data versus computational resources, bandwidth dependencies, common ontologies, useful toolboxes, publicly available data and results (journals), and apparent community agreement. Support is given to open-source policy and the philosophy that "the more databases the better (particularly for rare diseases)". University of Oslo offers storage services to the CERN grid, to the Accent EU NoE, and to NorStore as a collaboration between Universities. It also offers a virtual research environment, mostly for document exchange.

XI. Ilya Zaslavsky – *Long-Term Preservation of Spatial Information: Research Issues and Supporting Infrastructure*

According to Reagan Moore, preservation is an archival process through which a digital entity is extracted from its creation environment and migrated to a preservation environment, while maintaining authenticity and integrity information. The extraction process requires insertion of support infrastructure underneath the digital material and characterization of parameters, such as the authenticity and integrity, the digital encoding format, and the display operations. The goal is infrastructure independence, i.e., the ability to use any storage system, database, or access mechanism. In this sense, preservation is similar to information integration—but in the temporal domain.

A preservation strategy must provide emulation (to migrate the display application onto new operating systems, equivalent to forcing use of candlelight to look at 16th century documents), transformative migration (to migrate the encoding format to the new standard, migration period expected to be 5-10 years), and a persistent object for the characterization of the encoding format and the migration of the characterization forward in time.

A preservation environment must include a digital library infrastructure that supports the preservation metadata, the arrangement and description of items, and access mechanisms; and a data grid infrastructure that supports shared collections that are migrated forward in time, the management of technology evolution, and administrative metadata providing status of records.

In archiving and accessing spatial data, one is faced with:

- a variety of data types and models (open as well as proprietary data formats and management systems, and several emerging XML encoding standards), creating a need for some level of unification on a non-proprietary basis,
- different spatial registration mechanisms (e.g., spatial coordinates in a common (or well-described) system, labels in an agreed upon (or well specified) ontology, location relative to well-defined features, representative location—especially when navigating across scales),
- data structures with different amounts of "intelligence",
- the need of encoding data quality (spatial, semantic) dependent on the instruments, feature types, and transformations used,
- the need to specify metadata (the context, hyper-linked content, time stamps, etc.),
- the preservation of "look and feel" and
- graphical user interfaces for data integration, archive curation, and the retrieval and transfer of archived data into an operational environment.

As an example, the Smart Atlas is a grid-based Geographic Information System tool for spatial integration of multi-scale distributed brain data. It uses ontologies to delineate anatomic features, disambiguate label assignments, link features, and synchronize positions. The Smart Atlas tool is currently being developed by the BIRN-CC Data Mediation Team (www.nbirn.net/).

The following checklist of issues is recommended for archiving and preservation in the neurosciences:

- What are observation data preservation needs and challenges faced by neuroscientists?
- What are the appropriate archival forms for typical datasets in neuroscience? To what extent can available relational schemata, etc., be used as preservation formats, how can these approaches be integrated with existing standards such as the OAI (Open Archives Initiative) standards and protocols.
- How can archival and observations metadata be harmonized? The sharing and use of digital information in a community is different from the information and processing required for archiving in a long-term preservation environment. We need to incorporate the metadata standards adopted in the archival community without causing an undue burden on functioning systems.
- What are potential types of semantic mismatches, and controlled vocabularies and semantic reconciliation/mediation tools needed to support long-term archiving of neuroscience data? The neuroscience and archiving communities have different notions for some of the basic terms (e.g., a preservation record as understood by archivists may not match the notion of record as understood by domain scientists).
- What are the appraisal, accession, arrangement, description, preservation and access procedures to be implemented for neuroscience data? What is appraisal value of neuroscience records (and hence retention policies)?
- Once an archive is established, what is the minimal set of queries and instantiation requests it shall support? (and access restrictions)

- What are user interfaces for curating, updating, accessing, annotating and analyzing neuroscience data archives, appropriate for both archivists and domain scientists.

- What additional analytical services and navigation tools shall be implemented over large data archives?

- What is the software architecture of a neuroscience data preservation testbed, and infrastructure for sustainable distributed archiving of neuroscience data? Infrastructure-independence (to ensure that the archives can be migrated to new media, platforms and data formats as they become common) and replication of archives at different locations must be considered.

- What are institutional arrangements and "governance patterns" needed for a successful long-term archiving testbed implementation? Specifically, what steps are needed for integrating a preservation environment in an operational neuroscience data collection/dissemination system in a manner that establishes the archive as a trustworthy resource?

- What are the possible avenues for sustainability of a long term repository? (What is the business model?)

Recommendations:

- Engage with communities that deal with long-term preservation professionally.

- Have a "preservation testbed" to develop answers to the initial set of issues.

- Establish use cases for preservation.

XII. Matias Palva – *Data Format and Interface Issues for Sustaining Multi-Scale Databases*

There is a need for process and data management tools to solve problems encountered in current data analysis workflow. These problems arise from a combination of factors, such as the many sources for neuroscience data (such as MEG, EEG, EOG, EMG; eye gaze tracking; MRI, fMRI; psychophysics; quantitative trait loci (QTLs), genotype data, twins; neurophysiology; all resulting in a high dimensionality of the data); the manual work; the many operating systems, programs, and formats; the technical and programming skills required for the analysis process, and the incomplete documentation of process stages and details. As a consequence, processes are slow, cumbersome and irreversible, with irreproducible outcomes, waste of human resources, data integration challenges, and low return-of-payment from hardware investments. Data analysis becomes a bottleneck for productivity and quality within a hypothesis-driven science.

The platform developed by the OmniLyze Project aimed at providing generic solutions for data and process management, by developing a unified software system with graphical UI, centralized storage, services for computing, data management, process management, and entry through remote desktops or the web. This approach improves quality and productivity, decreases risks, facilitates data integration, makes the process reproducible and publishable, and improves human resource yield and hardware yield. In the view of the OmniLyze Project, process and data management should become a commonplace, everyday practise integrated into every step of the scientific workflow. Every form of data storage is considered to be a "database", and every database should offer data and process management. Databases exist on many scales such as personal, research groups, research centers, national, and global.

Sustainability of databases demands resources and continuity in both usage and development. Generic database solutions may be advantageous when resource income is proportional to the number of users, but require data to be transportable between databases across scales and functions, with standardized data formats and interfaces.

A standard data format (SDF) includes unambiguous definitions for legacy and metadata. It would promote database sustainability by user attraction, common UIs and usage, data and code reuse, sharing, modularization, co-development, avoid the need for specific database-database interfaces, support for database continuity, and decreased development and maintenance costs.

Recommendations:

A roadmap for INCF for data format and interface issues was suggested to include the following actions: (i) INCF arranges groups of experts to compile existing data and metadata definitions and format conventions, (ii) workgroups propose novel or amalgam definitions and formats, (iii) INCF compiles these and declares an SDF, (iv) database and equipment developers can produce the specific interfaces, and (v) INCF provides a portal to SDF resources and tools. SDF would be a basis for standardized process definitions, APIs, interfaces, and meta-interfaces.

At a technical level it was proposed to consider the HDF5 (Hierarchical Data Format; http://hdf.ncsa.uiuc.edu) as an attractive candidate to carry the role of SDF in neuroinformatics, as it handles data of any type, complexity, or size, it supports hierarchical grouping of high-dimensional datasets, and it has an efficient API providing fast data access including parallel I/O. It is usable in many computational platforms (AIX, HP-UX, IRIX, Linux, SunOS, Windows, Mac OS), is stable and sustained, widely used (EOS, 3 Pb, 1.6 M users) and is open source, costing nothing.

XIII. Martin Kersten – *Scientific Databases: The Story Behind the Scenes*

Key issues in the research of efficient and effective database technologies are the ultimate (virtual) machine architecture for database processing and the challenge of blending information retrieval with large-scale database processing. Thinking about neuroscience databases starts with an initial (data) exploration of the field followed by a series of stepping stones marking (1) the multimedia dimension, (2) the geometric dimension, (3) the lineage dimension, (4) heterogeneous databases, (5) GRID databases, and (6) the semantic search.

Stepping stone 1: Multimedia databases typically contain large volumes (>Tbyte, >Pbyte) of raw data with partitioning based on image, video segmentation, and indexing based on feature vectors. Query challenges concern proximity and probability based search, are CPU intensive with user-defined predicates and content-based information retrieval. For example, the database http://monetdb.cwi.nl/demo/bond consists of 100,000 images, with 25 patches extracted from each image, and a 14-dimensional feature vector derived for each patch, resulting in 2,500,000 images. The challenge is to find similar images based on Euclidian distance with sub-second response time. A novel database algorithm was developed to solve K-nearest neighbours (k-NN) search. An alternative scheme is to determine the probability that an image can be generated with a limited number of Gaussian mixtures. To this end, fix a limited number of Gaussian Mixture Models (GMMs), use an Expectation Maximization algorithm to fit the model over the image, and search similar images by comparison of the GMM model parameters.

Stepping stone 2: Any geometric abstraction of reality provides a good navigational map. Database storage and indexing support for 2D is mature with R-trees and Quad-trees, but commercial database vendors “do not like them”. Open research issue is to support 2D query embedding and scaling out towards 3 and 4 dimensions and temporal support. This approach is researched extensively in Geographical Information Systems. A lesson drawn from it is to avoid abundance of reference models.

Stepping stone 3: To ensure data lineage is a problem encountered in many scientific databases, namely the ability to travel back in time to understand, redo and judge the derivations. Challenges here are: (i) how to keep track of the complete context with data, software, parameter settings, etc., (ii) how to redo part of the analysis, and (iii) how to store and remember the lineage trails. An example is the AstroWise project in Groningen that keeps track of a complete workflow for telescope data analysis in a large Oracle database. All derivations are 5-line python programs.

Stepping stone 4: Sharing of heterogeneous information is the key problem in many databases. The standard approach here is to use commonly approved vocabulary and standard syntax, with XML being the standard language for self-descriptive data and its exchange between software systems. The database community is actively working on improving XML itself as well as XML-based tools and applications such as XQuery and Xupdate database engines, but it is not easy! The challenge is how to scale to large XML stores, how to efficiently search components, and how to realize structural information retrieval.

Stepping stone 5: GRID technology is partly a re-invention of distributed database technology, distributed programming, and high-performance computing with a focus on Authentication-Authorization-Access and data shipping over wide-area networks. Data distribution, replication, and parallel query processing has been well studied over the last three decades. World-spanning application deployment is centered around web-services, the middleware layers to publish and integrate different sources.

Stepping stone 6: Ontology integration is one of the most pressing challenges for the semantic web to take off, but integration of technology with databases is still immature. RDF and OWL are the leading paradigms; SPARQL is the first attempt to bridge the gap between traditional database management and semantic web technology. The integration of the cultural heritage field at http://e-culture.multimedian.nl/demo/search may serve as an example.

Recommendations:

- Manage sizeable scientific databases using open-source database technology in order to capitalize and steer expertise development

- Articulate the technical challenges in concrete requirements in order to know what you have to provide

- Benefit from progress in semantic web technology that might provide a good stepping stone requiring the development of ontologies and semantic bridges

- Make a dual community portal to mobilize the world, e.g., http://cas.sdss.org/dr6/en/

- Team up and make it a success—crossing the great divide between neuroscience and computer science; this is challenging and rewarding *if and only if* the building of the bridge starts from both ends, and parties recognize and respect each other's core business. The database community can provide knowledge on modeling, query processing, algorithms, data structures, scalability, persistency and flexible database systems.

Refs: http://monetdb.cwi.nl/, http://monetdb.sf.net/, http://www.monetdb.com

XIV. Roman Mouček – *EEG and ERP Records and Processing*

Data in EEG and ERP research originates from multiple electrodes and carries contextual information. A simple and flexible format for the exchange and storage of multichannel biological and physical continuous signals is provided by the European Data Format (EDF). Published in 1992 (Electroenceph. Clin Neurophysiol. 82: 391-393), it has become a standard for EEG recordings in commercial equipment and multicenter research projects. An extended version European Data Format (EDF+) was published in 2003 in Clin. Neurophysiol. 114: 1755-1761.

Recommendations:

- Work on data sharing in specific domains.

- Find the authorities, domain experts and technical experts.

- evelop domain standards (ontology, metadata), technical standards (file formats, database structures), technologies (programming languages, distribution, replication, encryption), and interfaces. A question is how to find the right level of domain differentiation.

XV. Martin Nawrot – *Three Challenges: Data Complexity, Data Format Confusion, and Non-transparency of Data Analyses*

Exchanging and sharing data is important for scientific cooperation. Problems arise from the complexity of the data and the confusion from data formats. Commercial electrophysiological equipment often provides proprietary (closed source) file formats and analysis software with very limited "built-in" data analysis and poor script languages. As a consequence the experimenter is often "stuck" with his company/format, and faces very limited data exchange across formats (format communities). It leads to (expensive) in-house hacks for data export/import, making the process fragile, error prone and vulnerable (no docs, no compatibility).

Suggested measures are to define open-source "unified" file formats, provide free services and expertise (for import / export functions, professional documentation, data base for data storage/exchange, teaching and training), enforce open-source standards, improve quality and flexibility (development community), and provide single interfaces to programming environments (Matlab, Python,…). Questions arise as to whether such a "unified data format" is realistic, as it supposes no competition on the market. An intermediate step in electrophysiology, adopted by a variety of companies, may be the application programming interface (API) Neuroshare http://neuroshare.sourceforge.net), with access to a variety of proprietary formats. It is, however, platform specific (Windows only), dependent on closed-source DLLs by industry, and no common standard is proposed.

A third problem arises from the non-transparency of data analyses. Data analysis links experiment and theory and high quality analyses and rigorous testing of theoretic models are crucial for the scientific advancement. Published analyses are generally in-transparent, while sophisticated tools are not easily accessible. Commercial software packages and poor data access are limiting factors. Suggested measures are to provide a code repository for data analysis (open source, multi-lingual), a tested toolbox for neural data analysis (open source), professional support and documentation, and teaching and training.

XVI. *Alessandro Orro* *– Data Management in the EGEE Grid Infrastructure: The BioinfoGRID Experience*

Analysis of biological data often goes with computational intensive tasks on huge amount of data obtained from custom databases with indexed flat files or local data from local file systems. A way to handle this is by using biological databases organized in a Grid, a network of clusters. Main issues are the installation and maintenance of the databases in the Grid Environment, the optimization and scalability (i.e., minimizing waiting time, storage and transfer costs), and the handling of versions.

Recommendations:

- Make database for analysis mainly read-only
- Relational/Object (for LIMS, for interoperability) and custom database (for analysis)

4. Workshop Discussions

Following the individual presentations of the expert participants, three discussion groups were formed to discuss sustainability issues from the community, the technical and the standards points of view, respectively. On the 2nd day of the workshop, a discussion group was formed on “additional issues of sustainability”. The next paragraphs summarize the main outcomes of these group discussions.

4.1 Community Issues of Sustainability

Clearly define the community (e.g., who is the audience for the resource?)

- Generators, Organizers, End Users (including clinicians), Funders, Journals, Libraries, Others
- Identify roles and needs of each

Communicate

- Engage all stakeholders and articulate simple but long-ranging goals
- Re-assure users outside of consortia that their needs are also being considered
- Involve journals and editors in the process
- Raise the awareness that the enterprise of neuroscience is bigger than any single sub-community
- Provide mechanisms for incorporating feedback on roles/needs/wants from users (wiki pages, bulletin boards, etc.)

Formulate (Standards)

- Develop focused but flexible standards
- Follow best practices where available (guideline document)
- Make standards decisions open to stakeholders and community input
- Review regularly and when needed

Educate

- Lower barriers to community availability/access

- Operational transparency should be high and it should be easy to get started with data sharing

- Empower next generation of neuroscientists by encouraging use of databases in dissertation meta-analyses and quantitative literature reviews

Evaluate

- Develop and utilize unambiguous/unbiased database usage statistics

- Give credit to database developers

- Give credit to those who have shared their data (for professional advancement)

4.2 Technical Issues Involved in Sustainability

Limitations of the technical domain

- Technical issues are secondary to community and standards development

- Understanding how neuroscience community is organized and works with data is critical in developing infrastructure for data sharing and sustainability

- To what extent shall INCF focus on technical solutions, beyond cross-walks and bridges across frameworks (metadata, ontological, spatial)? What would be an "INCF stamp of technical approval" for data: Standards-compliant (which standards?) with complete metadata, Cross-walk-ready?

Query languages

- Still early to embrace newest query languages—typing scheme of XQUERY is too complex and incompatible with SQL. Data-centric is still the focal point (not document-centric)—most of the data can be safely expressed in relational schema.

- Need a comprehensive data model, integrating datasets, documents and annotations.

- Isolate large neuroscience datasets: strategies for serving and querying them are different (perhaps OLAP, image services, etc.).

Analysis and benchmarking

- Target analysis environments: Matlab / Octave / R / Python. Open source is preferable but Matlab is overwhelming. There is a need to figure out use cases and scenarios, as well as recognized benchmarking.

- INCF is recommended to actively promote/solicit open-source solutions, develop case studies supporting open source/comparisons with proprietary systems, develop understandable benchmarks (and perhaps awards for reaching such benchmarks).

Towards a complete solution

- A vision of a complete solution/scenario sees the task broken down into manageable pieces, with well-defined metrics and a mapping to recognized user/developer roles, as well as to an infrastructure/service-oriented architecture (SOA).

- Look at experience of NIST organizing advancement of technology for informational retrieval. Define target problems, explicate queries, and come up with independent solutions to be compared at a workshop: This is a potential model for INCF.

Grid and web services

- The task here is to survey the scope and opportunities of neuroinformatics integration: find who or what tasks/scenarios need integration, determine what sources are likely to be queried within most scenarios, and support interesting integration testbeds scenarios (testbedding experience of OpenGIS may be relevant).

- Published service signatures are needed for web services: some XML standard for output (BrainML or anything), and web service wrapping tools for relational and O-O sources.

INCF Data Portal and Annotation Environment

- INCF may work on a NeuroWiki of BrainWiki (off INCF domain?), integrated with neuroscience publications/journals. But who will be the editors and which social networks will be addressed (off INCF)? Easy annotation and voting/ranking tools are needed (But what is the annotation model?), as well as a centralized gateway to all sorts of organized resources, neuroscience clearinghouse (à la kdnuggets.com).

Long-term preservation

- Often used datasets could be replicated at the central site, with centralized archival management/curation, and tools for format conversion/migration/annotation. Neuroscience data may have cultural heritage value, for instance, future generations of researchers may be interested in the first digital atlases of the brain, requiring long-term preservation strategies for selected atlases, but how do we select atlases and other neuroscience data to archive, what is the archiving method, and where would future researchers find such atlases? UNESCO may provide resources for long-term preservation. *This might be a unique long-term role for INCF*

Privacy/de-identification

- Recommendations need to be formulated on ethical and patent/copyright issues, and de-identification requirements for integrated datasets.

- Technical issues include grid and web service security, access control, single sign on, etc.

What can we learn from related communities?

- How are similar complexities managed in neighbor communities?

- What are the limits of applicability of standard Service-Oriented Architecture (SOA) solutions, and how do standards organizations function in other communities (e.g., OpenGIS in the geospatial field, emerging standards in astrophysics, biodiversity)? And to what extent are these experiences relevant? For instance, in GIS there were several known vendors, and it was a matter of having them support a few common exchange mechanisms—but in neuroscience?

Policy to support sustainability

- INCF could identify the data resources with highest information value, and the interconnections between these resources. Then, INCF can specify which resources shall be preserved and at which schedule, which resources are not sustained, and which resources have low information value and do not need to be sustained.

4.3 Standards Issues of Sustainability

The motivation of the working group was to identify issues for standardization, to provide a framework and rules for adopting standards, to discuss the role of INCF vis-à-vis other Standard Bodies and the question whether INCF should be a certifying body for standards.

Standardization issues – were identified as data, databases, ontologies, tools, models, and interfaces.

For standards to be established – they need to be adopted by several groups/organizations/agencies, be portable across systems, and be translatable into exchange formats. Standards should have a minimal core but be extensible, accommodate as many types of experimental hardware as feasible, and be compatible with existing standards.

Standardization topics – should cover a range of granularities: genomes and gene products (no need for developing new standards), cells and tissues, networks and systems, anatomy and images, brain regions and atlases, functional mapping, diseases, dynamical data including development, and models.

Data – should have a markup language with metadata info for formats, experimental information, granularity, description of terminology, and minimal standards. It should be portable, scalable and extensible, and needs an ontological framework on which the data is based. Examples for such standards are:

- for genomes and genes: NCBI/EBI,

- for gene expression and proteomics: MIAME and HUPO standards,

- for network data: NeuroML and extensions,

- for image data: DICOM,

- for anatomy: Standard atlases, Neural names.

A major issue is also to find common representations or cross-mappings of representations for the different fields and convergence of languages.

Databases – should be based on defined ontologies and schemata that are portable (in visible formats). They should allow for import/export of database data in exchange formats. Query engines must be integral to databases and be defined explicitly. Languages and source code specifications must be provided for database applications.

Models and Tools – Models should be in exchange formats with full specification, be portable, with defined granularity, and input/output specifications. Conceptual and analytical models must be defined explicitly, minimally, and where possible a computational model must be provided. Graphical representations are recommended. Languages, compilers and interfaces need to be specified for tools. Tools are needed for format conversion that allow for interchanging between different data formats.

Interfaces and APIs (standards) – The Web is taken as a standard for interfaces (user interface). Each interface must have a defined API, with specifications for graphical interfaces, portability, query, and use cases.

4.4 Additional Issues of Sustainability

Teaching and training – applies to writing papers (all components must be of high quality), to providing tools for good laboratory practice (tools and standards), use of well-designed databases and metadata, to process management, collecting and sharing data, E-learning courses, practical courses, exchange of students and lab visits, teaching resources (lecture notes, course scripts, citable lecture notes with review editorial boards), and taped lectures.

Software development – Open-source code policy was seen as an important issue. "Standardized" open source code / toolboxes should be developed by open-source toolbox communities, and open-source web dissemination (e.g., via SourceForge). INCF can be helpful in initiating these communities based on already existing initiatives in member countries, (e.g., XooNIps under the Japan node, pilot project with Bernstein center projects and CARMEN for electrophysiology, other communities for image analysis toolboxes, for data mining and machine learning (Weka open source in Java; http://www.cs.waikato.ac.nz/ml/weka/), and for signal analysis tools). Examples of open-source tool boxes are Numerical Recipes, OpenG (the LabVIEW open-source community), Python/SciPy and R. Toolbox communities can also support convergence of metadata descriptions (at all their levels). Adherence to standards (ISO) is important. A tracking system was suggested for authors/papers having used particular data.

5. Recommendations

Following the plenary presentations of the discussion group reports at the second day of the workshop, the participants discussed and outlined a series of recommendations to the INCF of actions to support sustainability of neuroscience databases.

i. INCF establishes a moderated web-based infrastructure with specific issues for discussion by the community

This infrastructure will enable community discussions and documentation on choices of data, databases, etc. for minimal information standards, discussion of models for broad utilization, and access to reference data, ontologies, atlases, markup dictionaries, and other relevant objects.

ii.INCF engages peer-reviewed journals in the process of identifying domain-specific minimal information recommendations for the sharing and sustainability of neuroscience data.

- Journals remain the principle means of scientific communication and provide motivation for focused groups of neuroscience researchers to collaboratively engage in distilling minimal information required for data exchange and utilization in defined domains of neuroscience. In addition, journals have the ability to reach the largest segment of the neuroscience research community. INCF can engage the journals to seed the process of providing domain expertise for minimal information standard definition to support experimental methods reporting which is also in the interest of the journals themselves.

- INCF encourages journals to publish special issues/section with articles discussing minimal standards for dissemination of neuroscience data, methods, tools, and models.

iii. INCF identifies specific types of data/databases and a set of researchers who are generating and disseminating these data to form a special interest group that will develop the minimal information standards for that data/database.

- Neuroimages, microscopic images, electrophysiological recordings, EEG/MEG, histology, and optical recordings are examples of mature types of data that INCF can begin to explore.
- INCF identifies and engages experts who disseminate such data and who are well motivated towards community oriented approaches.
- INCF develops a mechanism for accrediting and acknowledging experts who contribute to the above objectives of INCF.
- Process of development is transparent, public, and feedback is solicited and welcomed.
- Minimal information standards should have broadest applicability in order to avoid unnecessary granulation.

iv. INCF identifies specific types of models/tools and a set of researchers who are generating and disseminating these theoretical/computational models to form a special interest group that will develop the minimal information standards (in appropriate exchange formats, I/O, GUIs, etc.) for those models/tools.

- Network, anatomical, and disease models are examples for INCF to begin to explore.
- INCF identifies and engages experts who disseminate such models and who are well motivated toward community-oriented approaches.
- INCF develops a mechanism for accrediting and acknowledging experts who contribute to the above objectives of INCF.
- Process of development is transparent, public, and feedback is solicited and welcomed.
- Minimal information standards should have broadest applicability in order to avoid unnecessary granulation.

v. INCF investigates existing neuroscience data/tools/models clearinghouses and examines how they can engage in coordinating dissemination activities

- NITRC, NIF, SfN, and various research consortia are examples of data/tool/database clearinghouses for the INCF to investigate.
- Examine how this is done in other science disciplines.

vi. INCF examines how to serve as an accreditation body

- Develops the criteria and process for evaluating data/databases/models/tools.
- Possible mechanism is via application for accreditation to the national nodes.
- Recommendations then passed to INCF Secretariat.
- Posted on web, certification.
- Informs journals and societies of DB accreditation so as to help them know which resources meet/beat expectations for best practices.

vii. INCF can facilitate grass-roots recognition of need for data/database sustainability

- Satellite meetings/workshops at international meetings on database sustainability.
- Engage member-state funding agencies to extent possible involvement.

Appendix A

A1 Databases/Portals/Other References Used in the Report

Neuroscience

NIF – http://neurogateway.org/ – Neuroscience Information Framework

NDG – http://ndg.sfn.org/ – Neuroscience database gateway

CARMEN – http://www.carmen.org.uk/ – Code Analysis, Repository and Modelling for e- Neuroscience

NITRC – http://www.nitrc.org/ – The Neuroimaging Informatics Tools and Resources Clearinghouse

ICBM – http://www.loni.ucla.edu/ICBM/ – International Consortium for Brain Mapping

Neuroshare – http://neuroshare.sourceforge.net/index.shtml – Open data specifications and software for neurophysiology

Other areas

NCRI – http://www.cancerinformatics.org.uk/ – National Cancer Research Institute Informatics Initiative

caBIG™ – https://cabig.nci.nih.gov/ **– C**ancer **B**iomedical **I**nformatics **G**rid™

Norstore – http://www.norstore.no/ – Norwegian Storage Infrastructure

Biology Workbench – http://workbench.sdsc.edu – web-based tool for biologists with Sequence/structure Databanks

Microarray Resource – http://genome.ucsd.edu

Proteomics Resource – http://www.cellularsignaling.org, http://www.mitoproteome.org

Metabolomics Resource – http://www.lipidmaps.org

Cellular Pathways – http://www.cytoscape.org, http://www.biopathwaysworkbench.org

Modeling Tools Statistical Tools – http://www.modelingworkbench.org

MonetDB – http://monetdb.cwi.nl/ - open-source database system for high-performance applications in data mining, OLAP, GIS, XML Query, text and multimedia retrieval

SkyServer – http://cas.sdss.org/dr6/en/ - Sloan Digital Sky Survey

EGEE – http://www.eu-egee.org/ – Enabling Grids for E-sciencE

BioinfoGRID – http://www.bioinfogrid.eu/ – Bioinformatics Grid Application for life science

SourceForge – http://web.sourceforge.com/ – global technology community's hub for information exchange, open-source software distribution and services

Weka – http://www.cs.waikato.ac.nz/ml/weka/ – collection of machine learning algorithms for data mining tasks (open source written in Java, issued under the GNU General Public License)

Biodiversity Websites and Links

SMEBD – http://www.smebd.eu – Society for the Management of European Biodiversity Data

MarBEF – http://www.marbef.org – network of Excellence & ERMS 2.0

Fauna Europaea – http://www.faunaeur.org

ERMS – http://www.marbef.org/data/erms – The European Register of Marine Species (ERMS)

EurOBIS – http://www.marbef.org/data/eurobis – European Node of OBIS

MedOBIS – http://www.medobis.org/ – Regional Repository of Marine Biodiversity Data

OBIS – http://www.iobis.org/ – Ocean Bibliographic Information System

COML – http://www.coml.org/ – Census of Marine Life

Species 2000 – http://www.sp2000.org/ – Federation" of database organisations working closely with users, taxonomists and sponsoring agencies

GBIF – http://www.gbif.org/ – Global Biodiversity Information Facility

IFREMER – http://www.ifremer.fr/anglais/ – French Research Institute for Exploitation of the Sea

VLIZ – http://www.vliz.be/ – Flanders Marine Institute

ICES **–** http://www.ices.dk/indexfla.asp **– C**oordination and promotion of marine research in the North Atlantic.

Life Watch **–** http://www.lifewatch.eu **–** e-Science and Technology Infrastructure for biodiversity data and observatories

A2 Minimum Information for Neuroimaging Description and Specification (MINDS)

Several critical elements contributing towards MINDS include:

1 The raw data and full details for each MR acquisition type (e.g., T2, EPI, T1, DTI files; TR, TE, FOV, etc.).

2 The final processed (normalized) data for the set of acquisitions in the experiment/study (e.g., the version of the data just prior to statistical analysis).

3 Full subject demographic and diagnostic details (e.g., age, gender, clinical group, etc.).

4 The experimental design including sample data relationships (e.g., functional time course design matrix, conditional info, etc.).

5 Sufficient annotation of results with mappings to formalized neuroanatomical reference.

6 The essential data processing protocols and workflow (e.g., what normalization method has been used to obtain the final processed data—provenance).

7 The details of how statistical modeling was performed and inferences have been made (e.g., GLM, GRF, ICA, etc.).

Appendix B Workshop Program

General goal:
To discuss issues related to the sustainability of neuroscience databases, to identify problems, to discuss solutions or approaches to these problems, and to formulate recommendations to the INCF.

Program components:
1. Presentations by participants of issues related to the sustainability of neuroscience databases (20 minutes per presentation including discussion).

2. Discussion on identification of key problems, on approaches to these problems and on the role of the INCF (plenary and in small workgroups).

3. Formulation of recommendations and concluding report.

December 13, 2007

09.00 – 09.15	Introductions (Bjaalie and Van Pelt)
09.15 – 18.00	Scientific presentations and discussions
Jaap van Pelt	Introduction to the workshop – Sustainability issues of neuroscience databases
Shankar Subramaniam	Interoperability and Data Integration in Neuroscience Databases
Tadashi Isa	National brain research project “Integrating Brain Research” from the database committee point of view
Shiro Usui	J-node sustainability scheme including government support
Jack Van Horn	Business Models for Neuroscience Database Sustainability
Chris Emblow	The Society for the Management of European Biodiversity Data and its role in the sustainability of taxonomic checklist databases
Wouter Los	The Maintenance, sustainability and management of databases in the environmental sciences of partial correspondences in a single species
Fiona Reddington	Multi-disciplinary Data Sharing: A UK Perspective
Jostein Kandal Sundet	Data exchange in international collaborations
Ilya Zaslavski	Long-term preservation of spatial information: research issues and supporting infrastructure
Matias Palva	Dataformat and interface issues for sustaining multi-scale databases.
Martin Kersten	Scientific databases, the story behind the scene
Roman Moucek	EEG and ERP records - storage and processing
Martin Nawrot	Three challenges: data complexity, data format confusion and non-transparency of data analyses
Alessandro Orro	Data management in the EGEE Grid Infrastructure: the BioinfoGRID experience
16.30 – 17.30	Meetings of the Discussion Groups on “Community Issues”, “Technical Issues” and on “Standardization Issues”
19.00	Dinner

December 14, 2007

09.00 – 16.00	Discussions
10.00 – 12.00	Meetings of the Drafting Groups on “Recommendations” and on “Sustainability Issues”

design | easy.no

www.incf.org

INCF Secretariat
Karolinska Institutet
Nobels väg 15 A
SE-171 77 Stockholm
Sweden

Tel: +46 8 524 87 093
Fax: +46 8 524 87 150
E-mail: info@incf.org

incf International Neuroinformatics
Coordinating Facility