VIEWPOINT: NEUROSCIENCE

Neuroimaging Databases

The Governing Council of the Organization for Human Brain Mapping (OHBM)

These are comments written by the Governing Council of the Organization for Human Brain Mapping (OHBM), the primary international organization dedicated to neuroimaging research. The purpose of these comments is to identify and frame issues concerning data sharing within the neuroimaging community. Data sharing has become an important issue in most fields of science. The neuroimaging community is no exception, and it clearly perceives potential benefits in such efforts, as have been realized in other fields such as genomics. At the same time, such efforts can be costly (both in time and expense), and there are important factors that differentiate brain imaging from other fields and that pose specific challenges to the generation of useful neuroimaging databases. These include the rapid pace of change in brain imaging technologies; the complexity of the variables that must be specified to meaningfully interpret the results (such as the method of image acquisition, behavioral design, and subject characteristics); and concerns about participant confidentiality. These issues are outlined with the goal of framing and promoting a public discussion of the benefits and risks of data sharing, which can inform the field of neuroimaging as well as others that face similar challenges.

The following are comments prepared by the Governing Council of the Organization for Human Brain Mapping (OHBM)—the primary international organization dedicated to brain imaging research. These comments address the development of community databases in the field of neuroimaging. The need for data sharing within this community is compelling, ranging from the value of comparing data across laboratories to the construction of more sophisticated and complete models of brain function. OHBM realizes that data exchange at many levels and in a variety of forms is desirable, in accord with the wide spectrum of intentions within the neuroimaging community. OHBM is eager to help the field develop data-sharing efforts to meet these various needs, which, in collaboration with the relevant journals, will improve the quality, accessibility, and use of data in neuroscientific research. At the same time, data-sharing efforts within the field of neuroimaging face special challenges that differentiate it from other fields in which such efforts have already been successful. The purpose of our comments is to highlight these factors and the important issues that they raise. We hope that this will encourage members of the field, relevant experts (including those currently involved in neuroimaging database efforts), and other interested parties to express their views on these issues and that this will lead to a vigorous and informed public discussion. We expect that such an exchange will be of value not only to our field but to others that face similar issues. For example, the field of genomics now faces the challenge of cataloging gene function, a task that may be very closely related to the one our field faces in attempting to catalog brain function. We begin with a brief review of the relevant background, followed by a consideration of some of the issues we consider to be most important.

Background

Electronic data sharing has become an important tool in many scientific disciplines. This is especially true for those that work with large and complex data sets, such as astrophysics, proteomics, and, most recently, genomics. Recent successes in sequencing the human genome, in conjunction with efforts to disseminate those data, have provided a particularly visible and valuable example of how informatics can serve the interests of science and society at large.

The need. The value of data sharing has become increasingly apparent to scientists involved in neuroimaging, for several reasons. The volume of data generated with brain imaging techniques is striking, and continues to be one of the most rapidly growing areas in neuroscience. There are presently an estimated 1500 new brain imaging studies conducted per. year, comprising a total of about 10,000 subjects and about 100 terabytes of imaging data. Furthermore, these numbers are increasing rapidly as more scientists become interested in neuroimaging and gain access to neuroimaging facilities and as new facilities are being built. Published findings reflect only a small fraction of the data originally collected, often in a form that obscures their full complexity. The data themselves take a variety of forms and typically are not accessible for widespread sharing and use. Making neuroimaging data more accessible for sharing would facilitate the comparison of findings across laboratories, to allow better assessment of the reliability of methods and reproducibility of results; encourage meta-analyses that explore phenomena that are not apparent in individual data sets; and give investigators who do not have access to neuroimaging facilities the opportunity to conduct research using existing data. All of these are more efficient uses of neuroimaging data, which are relatively expensive to collect.

Some challenges. These potential benefits and the success of data sharing in other communities have inspired the neuroimaging community to consider ways of doing the same with brain imaging data. However, past experience and careful consideration of the issues involved make it clear that there are a number of factors that distinguish such an effort from similar ones in other domains.

First, unlike fields in which databasing efforts have been successful, there are no universally accepted standards for the structure and content of neuroimaging data sets. Data formats vary widely across different laboratories and neuroimaging methods. If this were a static phenomenon, then it might be a relatively straightforward technical challenge. However, the diversity of data and formats reflects dynamic factors, including the rapid and ongoing methodological developments within the field, the growing domain of phenomena to which neuroimaging techniques are applied, and rapid changes in our knowledge about brain organization and function itself.

Neuroimaging methods are evolving quickly, and changes are often accompanied by substantial changes in data content that directly affect format [for example, with the transition from anatomical to functional magnetic resonance imaging (MRI), data formats changed from three- to four-dimensional]. This problem extends to factors that define how the data were acquired and processed. New methods of image acquisition (such as for different MRI pulse sequences) and processing (for example, for alignment, noise reduction, and statistical analysis) often introduce new types and numbers of variables that together present a moving target for format definition of raw and analyzed data. Similar issues arise with regard to analyzed data. For

OHBM, Post Office Box 425464, Cambridge, MA 02142–0009, USA. URL: www.humanbrainmapping.org.

example, some methods of analysis produce maps of distinct regions of activity that can be summarized as maxima (or centroids) of activity and be assigned discrete anatomic coordinates. Other methods produce continuous-valued images. Methods for normalizing and/or comparing findings across subjects also vary widely, from three-dimensional morphing algorithms to surface flattening, all of which place different demands on data representation and formatting.

Finally, and perhaps most critically, imaging of brain function demands a clear specification of the behavioral conditions under which the data were acquired. Those are closely tied to the motivation and the scientific hypotheses to be tested, and often lead to highly specific experimental designs, the diversity of which may preclude predefinition. For example, the results of one study (such as a language study) may not depend critically on the visual intensity of a stimulus, and so this parameter may not be specified; whereas for another study, this may be an absolutely essential variable (as in a visual perception study). The number of potentially important behavioral variables is large and poorly defined, and our knowledge about which of these are most important is changing. Indeed, attempts to define them are at the very heart of most brain imaging research, and we are still in an early stage of development. A similar concern pertains to the equally rich set of subject characteristics (both demographic and clinical) that can influence brain imaging results. These factors present a profound problem for structuring and archiving the descriptors, or "metadata," that accompany imaging data; however, failure to do so can seriously compromise the interpretability of the imaging data themselves.

For these reasons, neuroimaging poses unique challenges to databasing efforts, and as yet there is no widespread consensus on the methods by which these challenges may best be addressed. By analogy to the field of genomics, we might think of efforts to catalog gene sequences as being comparable to building an atlas of brain anatomy rather than function. In contrast, creating databases of functional neuroimaging data is more akin to establishing a database of functional genomics: that is, the precise role of each gene, interactions among genes, and their relationship to phenotype. This is, of course, a greater and as yet unmet challenge. Nevertheless, there are clearly advantages to be gained from a discussion of, and well-informed efforts to begin developing data-sharing efforts within neuroimaging research.

What Can We Do?

Faced with these issues and recognizing the increasing need for data sharing within the neuroimaging community, OHBM asked its Neuroinformatics Committee to provide an outline of the critical issues relevant to the development of neuroimaging databases. A detailed outline of these issues appears as an appendix below. Several of these stand out as particularly challenging to the field of neuroimaging. Here, we highlight two general topics.

Diversity of data, databases, and database models. As noted above, many types of data are generated in neuroimaging research. They include data from diverse modalities, providing a wide range of measures of brain structure and function, which vary across modalities in their raw and processed forms. A natural inclination is to establish different databases for different types of data, and this has already begun to happen. Different models also are being explored. Some use a centralized model that directly manages the storage and distribution of the data. Others use a distributed model, in which a centralized listing is maintained that describes the available data and their locations, whereas the data themselves are stored locally, under the control of their owner, and exchange occurs directly between the user and the owner. As in other fields, the development of neuroimaging databases is likely to be evolutionary, with a number of efforts exploring a variety of approaches. Early experiments will provide lessons that will inform subsequent efforts. As individual databases mature, they will provide the foundation on which federations of databases and metadatabases can be formed (as in marine science, meteorology, astronomy, and other fields). OHBM recognizes that this process is natural, as the field explores the value of different approaches. A critical factor in this process will be the exchange of information about these various efforts, coordination among them, and careful evaluation of their impact on and worth to the community. Indeed, consideration must be given to the potential costs, as well as benefits, associated with databasing efforts, including their expense in time and effort (on the part of both developers and contributors), and this must be weighed carefully against the quality of the science they promote. OHBM is committed to providing mechanisms for the exchange of information about such issues, the recruitment of relevant expertise from other fields that face similar issues, and the facilitation of coordinative efforts in ways the field deems useful. It is presently acting to promote discussions, in various forms, of issues such as how to improve data exchange across different platforms and the extent to which standards are a practical and desirable goal in this context.

Confidentiality, credit, and control. The incorporation of raw data into centralized databases has raised concern within the neuroimaging community (*1*). Property rights and credit must be given to those who originally generated the data. The neuroimaging community must decide on the appropriate means to secure such property rights and credits, and OHBM is willing to help with this process. It must also develop means for ensuring the confidentiality of those from whom the data were obtained. For example, it is possible from certain data sets (such as structural MRI) to reconstruct recognizable images of a participant's face. Ways must be developed to prevent unauthorized access to or distribution of such data. Furthermore, the submission of data to public databases must be kept in alignment with informed consent procedures, which should accurately reflect the potential widespread use of such data. These issues will require careful evaluation and close coordination with the funding and regulatory agencies involved in human brain research, and this is another area in which OHBM can be of assistance.

Conclusion

Neuroimaging is a burgeoning field. The pace of progress in both method development and data acquisition is truly staggering, and the opportunities inherent in these developments are inspiring. However, to realize their full potential, the neuroimaging community must begin to consider how it will promote and coordinate efforts at data sharing. OHBM, as the major international organization representing scientists performing brain imaging research, is committed to playing a central and constructive role in this process. The organization has set up a Web site for public discussion of the issues raised in this communication (www.humanbrainmapping.org). A part of the annual OHBM meeting (this year in Brighton, England) is reserved for presentation and discussion of neuroimaging databases. OHBM sees as one of its responsibilities the provision of electronic and meeting-based forums for the exchange of ideas and opportunities regarding data sharing. This includes forums for learning from other fields in which databases have been in use for a longer time. OHBM is also committed to facilitating communication between and cooperation among databasing efforts and scientists in the field. We hope that the comments offered here and the more detailed appendix that appears below will be useful in moving us toward these goals, by helping to promote an informed and public discussion of the issues and challenges surrounding data sharing within our exciting and rapidly developing field.

Appendix: Outline of Issues Related to Neuroimaging Databases

At present, OHBM as an organization does not endorse any particular approach to neuroimaging database development. The goal of the outline below is strictly educational: To frame and promote an informed public discussion of the issues related to data sharing within the neuroimaging community. The list of issues and questions in this outline should not be considered as definitive or complete, but rather as a point of departure for the initiation of a thoughtful and thorough consideration of the issues. It also should be recognized that the present focus is on databases. We recognize that there are equally important issues regarding the development and dissemination of tools for data analysis, and that these are closely intertwined with database development and use. We hope that the present discussion will lead naturally to a similar discussion regarding tools for image analysis.

1. Data Contents

A) Imaging Data

There are a variety of neuroimaging methods, each of which produces data with different characteristics, in different formats, and involves different forms of preprocessing and statistical analysis. This raises the following issues for a database:

1) What types of data should be archived [such as structural MRI, magnetic resonance spectroscopy (MRS), functional MRI (fMRI), positron emission tomography (PET), single-photon emission computed tomography (SPECT), electroencephalography/event-related potentials, near-infrared spectroscopy (NIRS), and microscopy, among others]?

2) Should these include raw data, analyzed data, or both? Even these terms may warrant further clarification. For example, MR data is acquired in the frequency domain and is then translated into the spatial domain before statistical analysis is done. Which of these should be considered the raw form? For analyzed data, the problem becomes more complex: What constitutes the end point of analysis (parametric statistical maps, thresholded regions of interest, coordinates of the maximum or centroid of activity within such regions, etc.)?

3) What format should be used? There are a variety of formats specific to particular hardware vendors and software platforms (for analyzed data), as well as some industry standard formats (for example, DICOM). Not all of these are isomorphic with one another (there can be differences in resolution, dimensionality, and so on), and thus translation between them can involve data loss. For analyzed data, these problems become particularly difficult (for example, what coordinate system to use). If imported data is translated, how should discrepancies between formats be handled and what measures should be taken to ensure the fidelity of the archived form with respect to the original and to document any differences?

B) Metadata (Descriptors)

Neuroimaging data are associated with a rich set of descriptive information, or metadata, that defines the methods and conditions of data acquisition (such as device characteristics, imaging protocol and parameters, behavioral paradigms, and subject characteristics) and, for analyzed data, the statistical procedures that were used. The problem in defining and archiving such data is compounded by the rapidly evolving nature of the methods and the scientific applications. These issues raise the following questions:

1) How should metadata be organized and made accessible with the primary imaging data? How should these metadata be structured to handle the large diversity, and in some cases complexity, of existing descriptors?

2) How can the database adapt to handle new forms of metadata as new methods and/or applications emerge?

C) Data Import and Export

As already noted, a variety of data formats exists (for both imaging and metadata), and new ones are likely to emerge as methods develop. This poses challenges for the exchange of data between the archive and its users, as well as with other databases.

1) What formats should a database accept? Who should be responsible for conversion to or from the required formats: the database or the user?

2) If a database uses its own native format, should this be proprietary or should its full specifications be available for inspection and use by the community at large?

3) Should efforts be encouraged within the community to establish format standards and/or translation utilities that can facilitate data exchange? If so, who should be responsible for doing so and ensuring that these are kept up to date?

D) Data Quality

A critical factor for any database is ensuring the quality of its data along a number of critical dimensions, including validity (its veracity, or "truthfulness"), accuracy (the precision with which the archived copy approximates the original data), completeness (the availability of all data and relevant metadata contents), durability (the enduring availability of the data in a valid, accurate, and complete form), physical integrity (resistance to corruption or degradation), and logical integrity (the internal consistency of the archive's contents). Criteria for submission also are critical for defining the scope and value of the data.

1) Which quality factors are most critical for neuroimaging data? For example, how accurately must imaging data be stored (that is, with what precision) to ensure their validity, and how can such standards adapt as methods evolve (for example, as improvements occur in spatial or temporal resolution)? Given that similar studies often produce conflicting results, how should logical integrity be evaluated? Questions such as these raise the more general issue of how the quality of neuroimaging data should be defined and how such definitions can adapt to changes in the underlying methodology.

2) Should a database ensure the ongoing availability and completeness of its data so that at any time and even when the originally acquired data become unavailable, all information can be retrieved solely from the database? The issue of completeness is closely related to issues concerning metadata noted above. If a study involves critical factors that cannot be specified as metadata in the database, then how can completeness and durability be ensured?

3) Should data submissions be required to meet certain standards with regard to the method of image acquisition and/or statistical analysis? Should publication status be considered (for example, are all data eligible; should they have been published in some form; or should more stringent criteria apply, such as publication in a peer-reviewed journal or select set of approved journals)?

2. Data Access

All databases require some interface, running on the user's computer, that allows the user to enter data into the database and/or query its contents and retrieve the results.

1) What kind of queries should the database support? Should it simply provide access to the data in its stored form, or should it also support procedures for analysis and data reduction (for example, "for a given region of the brain, compute other regions that are coactivated with this region, given a set of experimental paradigms and a measure of the degree of coactivation")?

2) What mechanisms for user interaction should the database support? Should these be format-based or graphical, and should they be Web-based or involve platform-specific applications? If the latter, what platforms should be supported (Unix, Macintosh, Windows, etc.)? Should the underlying protocols be standardized (such as SQL or XML) or is it necessary to develop customized ones?

3) How fast and efficient should queries

and transfers be? Should data also be made available in "hard" form (as tapes, magneto-optical disks, DVDs, etc.)? How well should the structure and performance of the database scale with rapid growth? These considerations are particularly important, given the steadily increasing size of neuroimaging data sets and the remarkable rate at which they are being generated.

3. Data Ownership, Credit, and Confidentiality

Often, neuroimaging data sets contain more information than is reported in initial publication. Furthermore, they can contain information that reveals the identity of experimental participants.

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?

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?

4. Database Structure

A fundamental issue concerns the nature of the structure of the database itself. At present, there are two primary models for large-scale data sharing: centralized and distributed. Centralized databases maintain all data in a common centralized archive. All control of the data is administered by the database manager. Distributed databases maintain only an index of data sets, with descriptors and pointers to the location of the actual data, which reside with their owners who control their access. Once a set of data has been identified, data exchange occurs through direct interactions between the user and the owner.

1) How can the benefits of each model be exploited to meet the various challenges posed above? Centralized databases seem well suited for ensuring data quality (once this has been defined). However, it can be difficult to maintain their scalability and flexibility in the face of rapidly growing and evolving data forms. They also require careful negotiation with those who provide the data, regarding the issues of control, credit, and confidentiality noted above. Distributed databases seem to be well adapted for rapid growth and offer owners full control over their data. They also permit greater flexibility with regard to metadata, because the actual data maintained by the owner and provided to the user can extend beyond what is indexed in the database. However, this model can have problems with data quality and access.

2) Is it possible to build hybrid systems that provide the benefits of each model, or does it make more sense to support parallel efforts using each, encouraging them to meet complementary needs?

5. Interactions with the Community

A number of databasing efforts have already been initiated within the field of neuroimaging. These are at various stages of development and have somewhat different scopes and goals. Unfortunately, for the most part these have been isolated efforts, with little or no interaction among them. At the same time, a number of organizations and agencies have begun to recognize the importance of and express an interest in data sharing within the neuroimaging community. These include funding agencies, journals, and members of other scientific disciplines. These developments raise questions about how databasing efforts should relate to one another and to the community at large.

1) How can interaction and cooperation between different databasing efforts be encouraged? Should it focus on technical issues, such as data formats and interoperability, or should it extend to social factors such as data ownership and privacy?

2) Should other entities play a role in shaping such efforts? For example, should community-based organizations (within the field of neuroimaging, such as OHBM, or within the field of informatics, such as the Object Management Group) play a role in helping set standards? Should funding agencies or journals promote such efforts by considering mandatory submission of data to databases or helping establish standards for such submissions? If so, how will the databases be selected and who will do so?

3) How can the field assess the merits of particular databasing efforts? What measurable criteria can be used for evaluating and comparing efforts? Should quality and/or performance standards be developed, with which databases must comply to be eligible for endorsement and/or support? If so, how would these standards be developed and enforced, and what entities should be responsible for doing so?

References and Notes

1. M. Chicurel, *Nature* **406**, 822 (2000); *Nature* **406**, 443 (2000); P. Aldhous, *Nature* **406**, 445 (2000); *Nature Neurosci.* **3**, 845 (2000); S. Koslow, *Nature Neurosci.* **3**, 863 (2000).
2. This report was prepared by the Neuroinformatics Committee [Jonathan Cohen, Princeton University; Anders Dale, Massachusetts General Hospital (MGH); Alan Evans, Montreal Neurological Institute and Hospital; John Mazziota, University of California, Los Angeles (UCLA); and Per Roland, Karolinska Institute] and reviewed and approved by the OHBM Council [Nancy Andreasen, University of Iowa; Peter Bandettini, National Institutes of Mental Health; Randy Buckner, Washington University, St. Louis; Jonathan Cohen, Princeton University; Anders Dale, MGH; Karl Friston, Wellcome Department of Cognitive Neurology; Helen Mayberg, Rottman Institute; John Mazziota, UCLA; Marsel Mesulam, Northwestern University; Aina Puce, Swinburne University of Technology; Anna (Kia) Nobre, University of Oxford; and Per Roland, Karolinska Institute].