The International Neuroimaging Data-sharing Initiative (INDI):
Next Steps for the 1000 Functional Connectomes Project

Background: The era of discovery science for human brain function was inaugurated by the collaborative launch of the 1000 Functional Connectomes Project (FCP) on December 11, 2009, by leading members of the functional magnetic resonance imaging (fMRI) community. Following the precedent of full unrestricted data sharing, which has become the norm in molecular genetics, the FCP entailed the aggregation and public release (via www.nitrc.org) of over 1200 resting state fMRI (R-fMRI) datasets collected from 33 sites around the world In just over 6 months, the release generated over 9000 downloads and ~32,000 page-views from 1,223 cities in 78 countries. The FCP was featured in Nature-Medicine and Nature-Methods and in the NIMH director's weekly blog. Also indicative of the FCP's scientific impact is the success of the initial publication demonstrating the feasibility of data-pooling and discovery science with the datasets (Biswal et al., 2010). This paper, published in PNAS on 2/22/10, was downloaded over 1000 times in its first two weeks and remained among the journal's most-downloaded articles in the following months. Additionally, the data processing steps employed to carry out the feasibility analyses were made available on 3/3/10, yielding almost 600 downloads from NITRC in the first three months. The first FCP-based paper by independent investigators has already been published in PNAS (Tomasi & Volkow, 2010) and many additional papers are in progress.

The overwhelming response to the FCP data release demonstrated the scientific community's hunger for data, and has shown how one of the greatest obstacles facing the neuroimaging community – access to large-scale datasets – can be overcome by open data sharing. The involvement of the broader scientific community (e.g., statisticians, mathematicians, engineers, computer scientists) in discovery science for human brain function is too often dissuaded by lack of access to large-scale imaging datasets. While the FCP data release represents a considerable step forward for the neuroimaging community, the sharing of comprehensive phenotypic information to accompany imaging data is a crucial next step that is yet to be realized. Phenotypic information is essential to the development of imaging-based approaches to the categorization of phenotypic diversity in brain function in clinical and non-clinical samples. In turn, such approaches are fundamental to the development of objective measures of brain function capable of detecting and monitoring the pathophysiological processes underlying neurological and psychiatric illness.

Having provided the first large-scale demonstration of the feasibility and scientific value of open sharing of R-fMRI data, the next major challenge is to make the aggregation and sharing of well-phenotyped datasets a cultural norm for the imaging community. In order to accomplish this, two major paradigm shifts will be required. First, comprehensive phenotypic information must be made available with imaging datasets to facilitate sophisticated data-mining – a process by which novel relationships between phenotypic and imaging data can be revealed. Scientists are often reluctant to release their phenotypic data, because of the concern that competitors will gain an advantage in answering specific scientific questions. Unfortunately, this parochial attitude leads to the perpetuation of highly under-powered studies and a plethora of false positive and false negative results, while potentially massive amounts of data remain locked-up in individual labs.

Beyond the sharing of phenotypic data, a second paradigm shift is from retrospective to prospective data sharing. That is, in contrast to the FCP release, which primarily comprised datasets that had already been published, prospective data sharing involves regularly scheduled (e.g., weekly, monthly, or quarterly) sharing of data collected at contributing sites, as the data is being collected. The notion of sharing newly acquired data, rather than waiting until those data have been published, is novel in the imaging community, but is common practice in fields such as genetics where discovery science has been successfully implemented. In order for such a shift in practice to occur, one or more imaging groups must take the lead, and set an example for the field.

Goal: To usher in a new era of discovery science for human brain function through active data-sharing by the neuroimaging community. Central to the attainment of this goal is the concept of modeling – large scale imaging groups must self-impose the requirements of prospective, phenotyped R-fMRI data-sharing,

providing a model for the broader imaging community, while simultaneously creating a public dataset capable of dwarfing those that most groups could obtain individually. We believe successful attainment of this goal will recruit widespread involvement of the broader scientific community (e.g., statisticians, engineers, computer scientists, mathematicians), accelerating the development and application of analytic techniques to the characterization and understanding of brain/behavior relations at an unprecedented rate.

Objectives:

1. To enhance the 1000 FCP by including comprehensive phenotypic data. Building upon the model defined by the initial 1000 FCP dataset, groups from around the world will be invited to share their R-fMRI datasets via NITRC (www.nitrc.org). Unlike the initial release, contributors will be asked to include additional phenotypic data beyond sex and age. Specific focus will be put on the attainment of contributions containing R-fMRI data obtained in typically developing and aging populations, as well as clinical populations. All datasets will be welcome, regardless of publication status (published or unpublished). As before, local ethical approval for data-sharing will be a requisite.

2. To establish a common protocol for sharing phenotypic/metadata via the 1000 FCP. Existing datasets on the 1000 FCP website will be redistributed on October 15th, 2010 as an XNAT (xnat.org) virtual machine, including both imaging data and available metadata (including extensive imaging parameters when available). Future contributions to the FCP will be distributed in an XNAT native format (.xar files) by default. The XNAT format has been selected to ensure compatibility between the FCP and Human Connectome Project platforms. The development of converters to allow usage of 1000 FCP datasets with other leading databasing packages will be both encouraged and facilitated.

3. To initiate open, prospective data-sharing for the neuroimaging community. Eight imaging sites have committed to share data prospectively, with each site having its own unique sharing schedule (e.g., weekly, monthly, or quarterly). The initial prospective data-sharing samples include: Baylor College of Medicine, Beijing Normal University, Berlin Mind and Brain Institute, Harvard-MGH, MPI-Leipzig, NKI-Rockland, NYU Institute for Pediatric Neuroscience and the Valencia node of the Spanish Resting State Network. We will actively seek the involvement of FCP contributors, and the broader imaging community. Prospective data-sharing will begin with regular weekly releases from the NKI-Rockland sample (5-10 R-fMRI datasets / week, accompanied by extensive phenotypic data and DTI) on October 1, 2010. We anticipate this will allow us to work out any unanticipated problems; initiation of sharing from other sites will begin by January, 2011.