Specific Aims

This R24 proposal is submitted in response to RFA-MH-15-750, which invited an application to establish a Connectome Coordination Facility (CCF). The CCF will capitalize on recent successes of the Human Connectome Project (HCP), which has acquired, analyzed, and shared multimodal neuroimaging data and behavioral data on a large population of healthy adults. Major advances by the HCP include (i) the establishment of data acquisition protocols that yield consistently high quality data across multiple modalities (structural, functional, and diffusion MRI); (ii) the implementation of preprocessing pipelines that take full advantage of the high quality imaging data; and (iii) the establishment of a robust informatics and database infrastructure that has allowed widespread sharing of the HCP data within the neuroimaging and neuroscience communities. The CCF will build on these accomplishments and serve the human neuroimaging community in three ways.

Aim 1: Provide data acquisition methods and consultation services to the research community in order to facilitate harmonization of data generation across labs using HCP-style protocols. We will provide consultation and support services to the research community for the primary purpose of harmonizing image acquisition protocols with those of the HCP. The effort will establish a help desk for user support and provide several functions, including transferring data acquisition sequences and image reconstruction algorithms, providing updates and improvements for these sequences and algorithms, handling of material transfer agreements (MTA) and customer-to-peer (c2p) arrangements, harmonization of imaging protocols, image reconstruction support for different software platforms and versions, and consultation for potential problems (e.g. image artifacts).

Aim 2: Provide services to maximize comparability of data acquired by CCF contributors. These services will include pre-data acquisition guidance to contributors to ensure that each project's behavioral data are obtained using HCP-compatible methods. We will work with the contributors to develop mechanisms to streamline transfers of de-identified data from the study sites to the CCF database. The data from each study will include the unprocessed images (after deidentification), minimally preprocessed (MPP) data generated by each project's internal pipelines, and all data associated with the project's behavioral battery. We will implement manual and automated quality control procedures based on existing HCP methods to generate quality metrics that will be published with the data. Finally, we will run a standardized set of pipelines to produce minimally preprocessed data that is fully harmonized with the other data sets in the CCF database.

Aim 3: Maintain and expand the HCP informatics platform. We will maintain the existing ConnectomeDB data repository infrastructure for Human Connectome Data and expand it to include Connectome data from other research laboratories that are funded as U01 projects under the Connectomes Related to Human Diseases RFA. The hardware platform will be refreshed to include a new generation of servers and data storage system. The software platform will retain its existing capabilities for exploring, reviewing, and downloading richly documented data sets. Data access services will include the high-speed download tool within ConnectomeDB as well as shippable hard drives and cloud-based data sets hosted by Amazon's Public Data Sets program. The ConnectomeDB user interface will be adapted to support a broader range of data structures and will incorporate a rich set of data exploration tools, including dynamic charts and graphs, statistical toolboxes, and an interactive scripting console. The platform will be developed and operated following policies and procedures vetted by the HCP to ensure the privacy and security of the data it hosts.

Together, these three aims will establish the CCF as a central hub for connectomics data aggregation and sharing. The CCF's suite of harmonization services from data acquisition through data sharing will ensure an unprecedented level of compatibility across data sets. The resulting database will enable the scientific community to conduct novel analyses to better understand brain function in health and disease.

Research Strategy

Significance

Human neuroimaging is entering a new era that will include rapidly expanding amounts of data systematically analyzed and shared in order to accelerate progress in elucidating brain circuits in health and disease. The Connectome Coordination Facility (CCF) will contribute to this endeavor as a unique resource that enhances the acquisition, analysis, and sharing of high-quality neuroimaging and behavioral data related to brain ***connectivity*** and associated with a variety of brain disorders as well as healthy adults. The CCF will serve as: 1) a help desk for high quality MRI data acquisition; 2) a data sharing repository for projects funded by NIH under the Connectomes Related to Human Disease RFA (PAR-14-281); and 3) a platform that facilitates comparisons across projects, including the vast amounts of data on healthy adults generated by the Human Connectome Project (HCP).

In 2010, NIH awarded an HCP grant to a consortium led by Washington University and the University of Minnesota (the WU-Minn HCP consortium) to (i) enable methodical advances in MR scanning and data processing and (ii) to acquire and share multimodal neuroimaging and behavioral data from a target number of 1200 healthy adults (twins and their non-twin siblings). Data acquisition includes many imaging modalities, including two that provide invaluable (albeit indirect) information about human brain connectivity: resting-state functional connectivity (rfMRI) and tractography based on diffusion imaging (dMRI). The HCP datasets currently being acquired and shared have benefited from numerous improvements in data acquisition and analysis since the inception of the project. Many of these advances are detailed in a collection of 8 articles in a special 2013 issue of Neuroimage on Connectomics (see citations below). Here, it is useful to summarize these advances, because the projects to be supported by the CCF on brain diseases and disorders are poised to benefit from improvements attained by the HCP.

Data acquisition. On the data acquisition front, HCP efforts included several major improvements. Arguably the most important general advance involves multiband (a.k.a. simultaneous multi-slice) acceleration, which benefits fMRI (especially resting-state) and dMRI modalities by enabling larger k-space data sampling per unit time of data acquisition, which can be used in fMRI for higher spatial and/or temporal resolution, and in dMRI for higher spatial and/or angular (in q-space) resolution or simply shorter overall data acquisition times (Moeller et al., 2010; Setsompop et al., 2012). Piloting by the HCP enabled optimization of multiband pulse sequences for each modality, and also improved image reconstruction methods that are used to convert raw ('k-space') data acquired by the scanner into volume images (Ugurbil et al., 2013). Another advance involved a customized gradient insert with increased maximal gradient strength, which especially benefits the dMRI modality. The WU-Minn 3T 'Connectome' Siemens scanner has a maximal gradient strength of 100 mT/m compared to 40 mT/m on commercial scanners available at the time. Siemens subsequently introduced the 3T Prisma scanner, which provides 80 mT/m gradient strength (twice that of conventional scanners) and approaching that of the WU-Minn 3T scanner; the Prisma platform may be used by some of the Connectomes Related to Human Disease projects. Finally, the HCP demonstrated that high-resolution structural images (T1w and T2w at 0.7 mm isotropic voxels instead of conventional 1 mm isotropic voxels) can be routinely acquired and used as an accurate anatomical substrate for the other imaging modalities.

Data analysis. All modalities of unprocessed MRI volume data must undergo multiple transformations and preprocessing steps before they are appropriate for neurobiologically informative analyses. A wide variety of preprocessing approaches have been used in different laboratories around the world, as a consensus has been lacking on which approaches are most effective and appropriate. The WU-Minn HCP consortium carried out intensive refinement and optimization efforts that resulted in a set of 'minimally preprocessed' (MPP) pipelines for each modality. The overall objectives were to maximize spatial alignment, minimize spatial distortions, and avoid loss of information that is potentially useful at later analysis stages. For cerebral cortex, this includes improved methods for reconstructing the cortical sheet (capitalizing on the high-resolution T1w and T2w scans) and for intersubject alignment using a novel surface registration method (Robinson et al., 2014). We also introduced a novel method for generating cortical 'myelin maps' that are informative about functionally relevant cortical subdivisions in individuals and group averages (Glasser and Van Essen, 2011). Another advance is the introduction of a new 'CIFTI grayordinate' data format that combines cortical surface vertices and subcortical gray matter voxels into a compact representation that is valuable for many aspects of

data analysis and visualization (Glasser et al., 2013). The rfMRI data include outputs from an automatic denoising method based on automatic classification of signal vs. noise ICA components (Salimi-Khorshidi et al., 2014). The MPP pipelines resulting from these efforts (Glasser et al., 2013) are freely available (http://www.humanconnectome.org/documentation/HCP-pipelines/).

Quality control (QC). QC is another important aspect of HCP data processing. Many QC analyses are carried out automatically as part of the database infrastructure described below. However, since all imaging modalities depend critically on accurate mapping to brain anatomy, a trained QC expert assesses the quality of all structural images and cortical surface reconstructions. If, on initial evaluation, structural images do not meet QC criteria, rescans are scheduled whenever feasible.

Data sharing. The WU-Minn HCP data sharing effort has the overarching objectives of making data freely available at multiple levels of analysis via a user-friendly, robust database platform. To that end, the HCP developed and operates a sophisticated informatics platform to manage, distribute, and document its data. This database system, ConnectomeDB (Marcus et al., 2011, 2013) is built on the XNAT infrastructure (Marcus et al., 2007). Each data release includes both acquired, unprocessed data as well as minimally preprocessed data. The HCP data release process follows a formalized set of quality control, documentation, and organization procedures (Marcus et al., 2013) to ensure that the data are of highest possible quality and usability. Several dozen peer-reviewed studies have been published to date making use of the publicly shared HCP data from one or more modalities (e.g., http://www.humanconnectome.org/about/publications.html).

ConnectomeDB currently supports over 1600 users who have downloaded an aggregate total of over 850 terabytes (TB) of HCP data, plus 1250 TB shipped via hard drives to over 130 investigators. ConnectomeDB builds on the XNAT imaging informatics platform, adding a number innovative new features, including an extensible data dictionary service, tiered data access and project-specific data use terms, data dashboards with charting and filtering, and high speed data transfer technology.

To date, the HCP has distributed data in 5 major releases. A recent release (June, 2014) includes MRI and behavioral data acquired from over 500 subjects and also includes extensively processed resting-state and task-fMRI data. A 900 subject release is scheduled for spring of 2015, and a final 1200 subject release will be made available at the conclusion of the main HCP project in early 2016. In addition, an initial release of data acquired on the UMinn 7T scanner will occur in the winter of 2015, with the full 200-subject release in late 2015.

Support for multiple projects. The ability to support data sharing across multiple projects will be key to the CCF. The WU-Minn HCP has demonstrated progress on this important front through the release of data from two additional connectomics projects. One involves dMRI data from the MGH-UCLA/USC HCP consortium, which has acquired and shared dMRI data obtained using a customized Siemens scanner at MGH having a gradient insert with 300 mT/m maximal gradient strength.

Another project involves connectomic data across the human lifespan. In 2013, the WU-Minn and MGH/USC HCP consortia were each awarded “Lifespan Pilot” supplements to acquire and analyze pilot data across the lifespan. The WU-Minn consortium is focusing on six age ranges (4-6, 8-10, 14-15, 25-35, 45-55, 65-75 years) and has established data acquisition protocols that are approximately half the total duration of the main HCP protocols (2 hr vs 4hr) that can be consistently acquired across all but the youngest age group. These protocols are also well suited for studies of subjects with brain disorders, who in general are less tolerant of long periods in the scanner. An initial WU-Minn HCP Lifespan Pilot dataset was released in September, 2014. Additional Lifespan pilot data releases from both consortia are planned for the winter of 2015.

Given the success of its data acquisition, data processing, data releases and informatics platform, the WU-Minn HCP team is prepared to continue distributing its data and to expand its services to support awardees of the Connectomes Related to Disease U01 program through the Connectome Coordination Facility (CCF). In addition to distributing the data collected by these awardees, the CCF will provide consultation services to them at the front end of their studies to ensure that the data they collect is of highest quality feasible and as comparable as possible to the main HCP data.

Why is this the right team? The co-PIs for the CCF will be Drs. Dan Marcus and David Van Essen, who have worked very closely over the past 5 years to establish the HCP neuroinformatics infrastructure, which includes the ConnectomeDB database, the minimal preprocessing pipelines, and the Connectome Workbench visualization platform. For the CCF, a major focus will be on the expanded multi-project scope of ConnectomeDB. Dr. Marcus leads a highly experienced informatics team who will be able to move smoothly from the existing HCP effort to the needs of the CCF. Dr. Van Essen's experience as a PI of the WU-Minn HCP has provided a broad perspective on the interrelated aspects of data acquisition, analysis, and visualization that will aid in overall coordination across multiple projects.

The data acquisition help desk will be led by Drs. Kamil Ugurbil and Essa Yacoub at U Minn (CMRR). Dr. Ugurbil is a PI of the WU-Minn HCP consortium, and Dr. Yacoub has been a key investigator on the data acquisition and data analysis aspects of the HCP. They have both worked very closely and effectively with Drs. Van Essen and Marcus for the duration of the HCP. The CMRR team is highly experienced in the development as well as sharing of the multiband pulse sequences and image reconstruction algorithms that are central to the HCP data acquisition protocols.

In addition to the project leads at WU and UMinn, the CCF team includes several other key contributors to the HCP. Drs. Barch and Harms, who developed the behavioral data acquisition and quality control programs for the HCP, respectively, will lead these same efforts for the CCF. The project managers and outreach coordinator (Dr. Sandra Curtiss, Eileen Cler, Dr. Jennifer Elam) will continue in these roles for the CCF. Finally, the informatics technical team that developed the ConnectomeDB infrastructure will operate and expand ConnectomeDB for the CCF.

Innovation

There is growing recognition of the importance of data sharing for accelerating discovery and innovation across all domains of neuroscience (http://www.braininitiative.nih.gov/2025/BRAIN2025.pdf). However, there are numerous challenges in making data sharing efficient, robust, and scientifically useful in each domain of neuroscience. For human neuroimaging, the challenges include the complexity and size of commonly acquired datasets, plus the fact that various projects use diverse imaging and behavioral/clinical methods of acquisition and analysis.

A growing number of data sharing efforts involving human neuroimaging have been supported by NIH and private foundations in recent years. Several factors will make the CCF a unique and innovative addition to the existing set of valuable resources, a non-exhaustive list of which includes ADNI (http://adni.loni.usc.edu); the NKI-Rockland datasets (http://fcon_1000.projects.nitrc.org/indi/pro/nki.html); and the Brain Genomics Superstruct project (https://thedata.harvard.edu/dvn/dv/GSP). (i) The CCF will support and share rich multimodal datasets that are consistent in acquisition and preprocessing and of demonstrably high quality. (ii) The datasets to be made available will likely encompass a wide range of brain disorders (though this obviously depends on which specific projects are funded over the course of the NIH Connectomes Related to Human Disease initiative). (iii) The CCF will provide access to data across these diverse projects via a robust database infrastructure that supports flexible queries within and across projects. Hence, we anticipate that it will spawn innovative analyses by the neuroscience and neuroimaging communities, as is already occurring for the HCP.

Approach

Overview

The Connectome Coordination Facility will provide services to maximize the quality, comparability, and sharing of the acquired and processed imaging and behavioral data collected by investigators intending to use HCP-style acquisition protocols, specifically the awardees of the U01 Connectomes Related to Human Disease program (**Figure 1**). The CCF will serve each U01 project during the project's initial optimization phase by providing consultation and guidance on how to acquire images that are maximally compatible with the data collected by the HCP (**Aim 1**). The CCF's helpdesk' will serve projects that use existing HCP scanners or similar platforms (e.g. Siemens Prisma) by providing up-to-date pulse sequences and verifying that the sites are obtaining high quality data. In addition, the CCF will provide services to harmonize data collected by the U01 awardees and ensure the secure import of these data into its database for broad distribution to the

research community (**Aim 2**). These services will include providing information on the HCP behavioral tests and guidance to ensure that each project's behavioral data can be shared in comparable ways. As each project moves from optimization to data collection, the CCF will work with the project teams to develop mechanisms to transfer de-identified data from the study sites to the CCF. The data from each study will include the unprocessed images (after de-identification), minimally preprocessed (MPP) data generated by each project's internal pipelines, and all data associated with the project's behavioral battery. Each project is expected to have its own quality control procedures to review and cleanse their data, and we anticipate that many projects will choose to implement HCP-style processing pipelines at their own sites. To complement these efforts, the CCF will execute its own procedures to generate quality metrics that will be published with the data and to produce minimally preprocessed data that is fully harmonized with the other data sets in the CCF database. In addition, CCF will run pipelines for any projects that elect to handle this processing through a subaward or other support of CCF personnel. Finally, the CCF will develop and operate a web-based informatics platform to make all U01-funded datasets available to the research community (**Aim 3**). This platform will build on the HCP's ConnectomeDB system, which includes data browsing, searching, mining, and high-speed download services. It is backed by a massively scalable, redundant data storage and high performance computing infrastructure. In anticipation of the CCF's needs, the HCP team has already begun adapting ConnectomeDB to support multiple data sets. The CCF will also leverage novel data distribution methods developed by the HCP to make data available via shippable hard drives and a public cloud computing service. To ensure the successful completion of these aims, the CCF will use formal project management, communication, documentation, and oversight practices established by the HCP (**Program Wide Activities**).

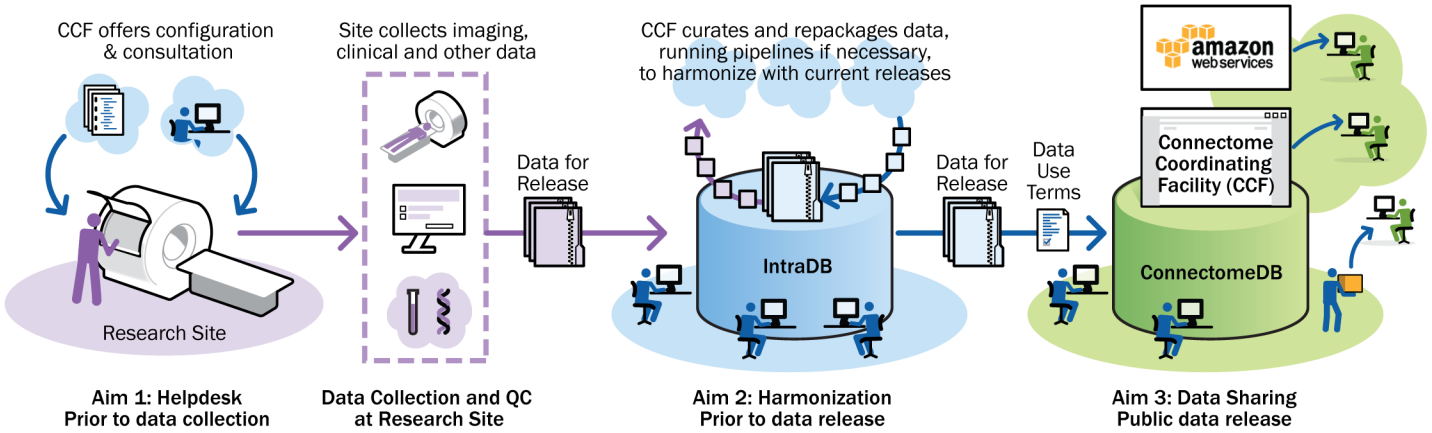

Figure 1. The Connectome Coordination Facility (CCF) will provide services from data acquisition through data sharing to enable the data collected by the NIH Connectomes Related to Human Disease program to be compatible, of highest quality, and readily accessible to the research community.

Specific Aim 1. Acquisition Helpdesk

A major goal of The Human Connectome Project (HCP) has been to acquire high spatial resolution images for functional and diffusion MRI while also optimizing temporal resolution and signal-to-noise ratio (SNR). This entailed building on the approach taken by HCP investigators at the University of Minnesota using 2D slice-accelerated technology (i.e. multiband (MB), or simultaneous multi-slice (SMS; Larkman et al., 2001)), which had been under development at the University of Minnesota's Center for Magnetic Resonance Research (CMRR; Feinberg et al., 2010; Moeller et al., 2010; Ugurbil et al., 2013; Xu et al., 2013). 2D slice accelerations permit many-fold reductions of the volume repetition time (TR), without significant penalties in image SNR. The MB technique was improved by incorporating controlled aliasing strategies (Setsompop et al., 2012). Extensive work undertaken by the WU-UMinn HCP consortium included further development, optimization and evaluation of the MB pulse sequences. These are detailed, along with the parameters finally decided upon for the HCP in Ugurbil et al. (2013) and Xu et al. (2013).

The HCP effort resulted in a fully integrated, technician-friendly, fully tested and optimized package for slice accelerated fMRI and diffusion imaging, which runs on a Siemens clinical platform. The pulse sequences and reconstruction algorithms developed by the HCP have been installed and are currently the default sequence for both fMRI and diffusion imaging at more than 125 sites and for hundreds of investigators worldwide. In addition to implementation on Siemen's scanners, HCP investigators have also facilitated efforts in the

development of slice accelerated sequences on Phillips and GE scanners. The current implementations on non-Siemens scanners are not as fully integrated or as user friendly as the Siemen's implementation, requiring much more user know-how and, thus, significantly more user support. Currently, scores of NIH funded projects at many sites rely on the CMRR/HCP pulse sequences and subsequent support to generate and optimize their scanning protocols. Many of these projects use the sequence for improved data quality, while others aim to harmonize their protocols with the HCP protocol in order to produce HCP comparative data sets, such as recent applications for the NIH Connectomes Related to Human Disease program..

We will provide consultation and support services to the research community for the primary purpose of harmonizing image acquisition protocols with those of the HCP. The effort entails several components, including transferring the pulse sequence, harmonization of the protocols, image reconstruction support, and a help desk for user support.

Sequence Transfer and Usage. Logistical efforts include completion of a material transfer agreement (MTA), including a Siemens authorization process and an inter-institutional agreement. The transfer of pulse sequences to other investigators requires careful management of a process that includes legal signatures by the University of Minnesota, Siemens, and the institution to which the pulse sequences will be transferred. Once the sequence transfer has been approved by all parties, the receiving site is given access to download any available version of the sequence, along with documentation regarding the sequence and its usage. This process also has a web-based infrastructure in place, supported by the CMRR (http://www.cmrr.umn.edu/multiband/index.shtml).

Harmonization. Currently the CMRR MB "customer to peer" (c2p) process supports 10 different Siemens Software versions (VB17, VB17/F15, VB19, VB20, VD11, VD13A, VD13B, VD13C, VD13D), which are all currently in operation on Siemens scanners. Porting and harmonization of the pulse sequence itself from one software version to the next, while also continuing to develop and support previous versions, is a major logistical and technical undertaking. A typical Siemens software upgrade can involve fundamental changes to the sequence structure and reconstruction pipeline, requiring careful re-coding of the MB specific portions and version-specific compatibility issues. Such software upgrades can occur as often as once a year. Currently, only one software version runs on 3T Prismas, but this is likely to change in 2015. We have been providing the HCP-Prisma imaging protocols to requesting sites, allowing replication of what is run on the CMRR Prisma. When Prisma upgrades occur, we will provide updated protocols for sites that choose to upgrade. The CMRR will have 2 Prismas in operation and will upgrade one of the scanners upon availability. For sites running non-Prisma software versions, we will rely on collaborator sites to test and generate HCP-like protocols. For non-Siemens sites, we will continue to work with collaborators to facilitate versions running on GE and Phillips scanners. In doing so, we will facilitate generation of HCP-like protocols for any sites funded under the U01 Connectomes Related to Human Disease mechanism and for eventual broad distribution.

Image Reconstruction Support. An important component of the MB sequence support involves image reconstruction pipelines for online MB reconstruction on Siemens scanners, and Matlab-based versions for offline reconstructions on non-Siemens scanners. The need for offline reconstruction stems from unavailability of online reconstructions on non-Siemens scanners or inadequacy of online reconstruction computers on Siemens scanners. As part of the CMRR MB c2p pulse sequence package, the CMRR has also developed the infrastructure for sites to purchase their own high-end computers that can be interfaced directly (via a fast network card and using a scheme and software support provided by CMRR) with the Siemens image reconstruction system. This process requires substantial technical/network expertise and has been offered only on special request (http://www.cmrr.umn.edu/multiband/remote-recon.shtml). This functionality has proven to be critical to several sites that would otherwise not be able to reconstruct images online (in near real time) at the scanner, or would require several extra hours of scanner time to wait for the reconstructions to finish. As part of this aim, the CCF will provide additional support for sites limited by MB image reconstruction times on Siemens scanners. Off-line image reconstruction algorithms are also available for non-Siemens users, though in this case, CMRR does not provide support for integration of the off-line hardware to the scanner.

Help desk/User support. Our experience has been that many sites require practical support during the installation process or after the sequence is installed and becomes functional but prior to commencing scanning. Significant efforts have gone into the documentation and user friendliness of the CMRR MB

sequence, however, the MB sequence is still a novel pulse sequence requiring training and thus additional support. Further, the sequence installation process is not done directly by CMRR personnel but rather needs to be done by the user via an installation package. For various reasons related to Siemens specific issues, this process can fail from time to time. For non-Prisma sites, additional user support is needed, as any proposed protocols would not have been tested directly by developers at the CMRR. To facilitate support, because of the number of sites with varying expertise and experience, the CMRR has made available an additional website where support requests can be made directly (https://github.com/CMRR-C2P/MB). Finally, the CMRR will provide users with QC protocols run on the CMRR Prisma, which can be used to gauge the performance of the sequence and the data at a given site. CMRR has also been involved with solving site-specific problems, for example when there is suboptimal hardware performance, leading to image artifacts. CMRR has been, and, through funding from this grant, will continue to provide support to track the origins of these problems and even provide solutions in the form of modified image reconstruction algorithms.

Specific Aim 2. Intake and Harmonization Services.

The NIH is expected to fund a number of projects under the Connectomes Related to Disease U01 program. The data collected by these projects will be of greatest value if it is collected and distributed in a manner that maximizes comparability across projects. The CCF services for optimizing compatibility will include guidance prior to data acquisition, streamlined data intake procedures, uniform quality control metrics, and standardized HCP-based image processing pipelines (**Figure 2**).

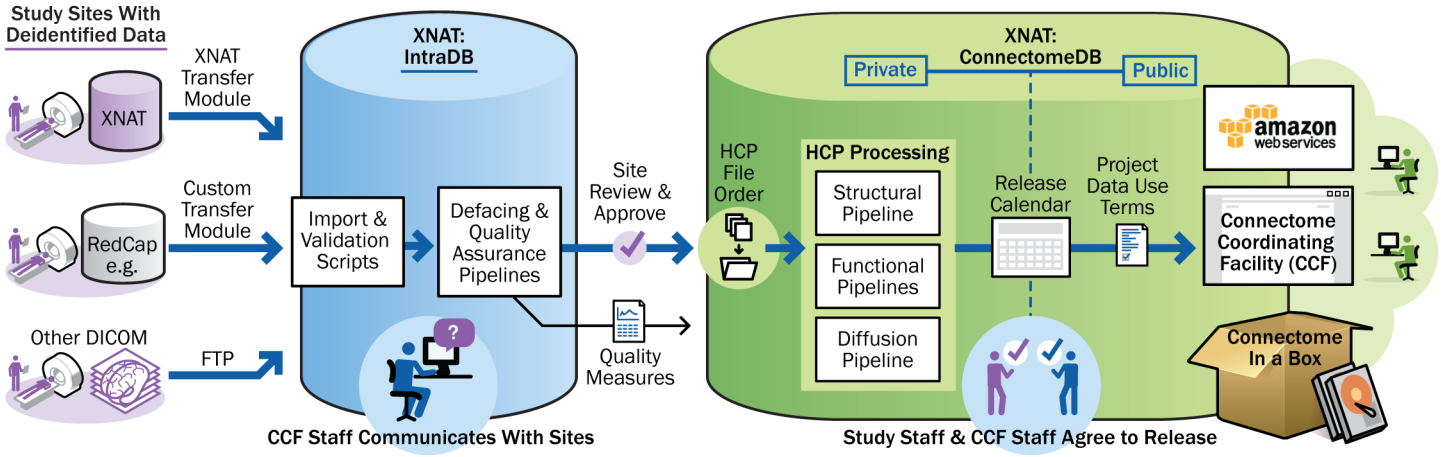

Figure 2. The CCF intake procedures cover import from study sites through data sharing and include validation, quality control, and multi-prong distribution process.

Behavioral Battery Harmonization. As with image acquisition protocols, the HCP has developed a broad-based behavioral battery expected to be used and adapted by the U01 projects. The battery includes a spectrum of domains, including cognition, alertness, emotion, motor, personality, sensory, psychiatric and life function, substance use, and medical history. It is implemented on the freely available NIH Toolbox (http://www.nihtoolbox.org) and University of Pennsylvania (Gur et al., 2010) testing systems. The CCF will provide guidance to the U01 projects in implementing the battery for their studies, under the leadership of Dr. Deanna Barch, who spearheaded the development of the HCP behavioral battery. While the HCP battery will provide a suitable basis for a "standard" battery, the unique aspects of each U01 project, including age ranges and disease characteristics, will necessitate that the HCP battery be adapted to best suit each project. The CCF will work with the PIs of each U01 project to harmonize behavioral data collection, with the aim of: 1) retaining all instruments in the battery and to administer them according to HCP's standard operating procedures; 2) using the same testing platforms (NIH Toolbox and University of Pennsylvania) used for the HCP or the widely used REDCap platform, thus ensuring compatibility and continuity in data collection formats; 3) using standardized assessments such as the NIH Common Data Elements (http://www.nlm.nih.gov/cde/); and 4) selecting common instruments for projects in overlapping domain areas, coordinated through the CCF.

Project Intake. Data will be imported from contributors using methods designed to streamline the process for each project, with the goal of minimizing the impact on contributors while providing ample opportunity for the CCF to review and curate the data prior to public release. Based on communications with many U01 applicants, we expect that different methods will be used by the various projects to manage their internal data

collection and storage. Many projects will likely use instances of the HCP's XNAT-based IntraDB system (or HCP's IntraDB itself). For these projects, we will implement a module that automatically transfers data at specified intervals from the study's IntraDB instance to the CCF. Other projects may use informatics systems similar to XNAT (e.g. REDCap) for which automated transfer procedures can similarly be developed. Finally, some projects may not use an informatics system at all or may use a system that is not easily adapted to automate transfers. For these projects, the data will be transferred to CCF using a secure FTP site. Data transferred using any of these methods will first be deidentified by the originating site by removing identifying information (including acquisition dates) from the data file metadata (see *Privacy Policies* below). Prior to regular transfer, a test data set containing a single subject will be transferred to CCF and reviewed to verify the deidenfication and transfer procedures. Imaging data will be expected to be transferred in its original DICOM format (after deidentification) to allow the CCF to review and verify acquisition protocol details embedded within the DICOM metadata.

The timing of data transfers will be coordinated individually with each contributing project. We expect that studies may choose to transfer individual acquisitions soon after data collection while others will prefer to transfer their data in batches. For projects that choose batch transfers, we will require an initial batch within 1 month of study data collection to allow us to review the data for compatibility with the HCP protocol and to ensure that the data are of sufficient qualify. If the data are determined to be of insufficient quality, the Acquisition Helpdesk will assist the site in improving their procedures and adapting their acquisition protocol as needed.

Communication with data contributors will be led by the CCF PIs through the CCF Steering Committee (see below) and project-specific meetings. Project coordination will be led by the CCF project managers using web-accessible applications (e.g. Jira, Basecamp) to document timelines, deliverables, and action items. For each project, a Data Transfer Plan will be developed and approved by all stakeholders in order to formalize the data types and formats to be transferred, scan naming conventions, acquisition session structure, the method of transfer, the timing of transfers, deidentification procedures, and review procedures for the data. Technical aspects of data exchange will be implemented by CCF staff programmers, and a dedicated CCF data curator will execute and review all data transfers.

Data flow. When the CCF receives data from a contributing project, it will be imported into the HCP IntraDB informatics system, a private system accessible only to internal staff. The CCF's data curators will review the data within IntraDB, using a combination of automated and manual procedures. An automated procedure will inspect the DICOM headers to verify that the acquisition protocol was followed. The HCP "defacing" program will be executed in order to remove potentially identifying image details from the face and ear regions in high resolution anatomic scans, and a data format conversion program will generate a NIFTI-formatted file for each scan. The anatomic scans will then be inspected and rated by a highly experienced image analyst. Finally, automated image quality control procedures will execute to generate quantitative quality metrics to be published with the data (see *Image Quality Metrics* below).

Once these procedures have been completed, the data will be transferred to a private project within ConnectomeDB dedicated to a specific data release and to which only personnel from the CCF and the contributing project have access. During the transfer process, the data will be organized to comply with the standard structure and naming schemes developed by the HCP. The HCP's standard processing pipelines (see 'Pipelines' below), adapted as needed for the study's acquisitions, will then execute, and the output will be reviewed by the CCF's image analyst and relevant project personnel. The project will be released for public access after the dataset has been reviewed and approved by both the CCF and contributing study. Specific subjects and acquisitions that are deemed unusuable will be flagged and removed from the release.

Image Quality Metrics. The HCP has implemented automated quality control procedures that generate quantitative metrics for individual resting state and task fMRI scans, including temporal signal-to-noise ratio (tSNR), absolute and relative motion, and DVARS (Power et al 2012; Marcus et al 2013). Similar metrics will be implemented for diffusion scans. These quantitative metrics will be published with each data set, including data set-wide averages and individual subjects values and a user interface will be included in ConnectomeDB that allows users to filter data by quality ranges. We anticipate that the connectomics community will use the HCP data and similar data sets to develop a consensus of quality ranges for tSNR, motion, and other metrics

that we can use to help guide users in their selection of quality data sets. Such guidance will need to be cognizant of the likeliness that disease affected cohorts may have very different quality profiles than healthy controls. This is an evolving and active area of methodological research in which the CCF will be highly engaged. Our QC methods will be updated as improved tools and practices emerge.

Pipelines. The HCP has implemented a series of minimal preprocessing pipelines for structural, functional, and diffusion MRI to accomplish many low level tasks, including spatial artifact/distortion removal, surface generation, cross-modal registration, and alignment to standard space (subcortical volume space and cortical surface space; Glasser et al., 2013). These pipelines are specially designed to capitalize on the high quality data offered by the HCP-style acquisition protocols. The final standard space makes use of a recently introduced CIFTI file format and the associated grayordinate spatial coordinate system. This allows for combined cortical surface and subcortical volume analyses while reducing the storage and processing requirements for high spatial and temporal resolution data (Glasser, et al., 2013). These pipelines will be adapted as needed to suit any variations in the acquisition protocols of the U01 awardees and executed on all of the data contributed to the CCF. The output of the pipelines will be distributed through ConntectomeDB using the same organization and file naming schemes as established by the HCP.

Privacy Policies. The HCP has worked hard to ensure that HCP data are maximally accessible to the scientific community while still insuring appropriate privacy protection for our participants. We believe that the HCP model of tiered data—Open Access and Restricted Access, depending on how likely the data in combination are to identify a participant and whether the data would be embarrassing if made known—will also work well for U01 Disease Connectome awardees. Data that are within the Restricted Access tier require additional levels of user authorization and vetting, including assurances that the data will be managed in a secure environment and not further distributed. The CCF will work with the Wash U IRB to ensure that all CCF data sharing is in accordance with the privacy provisions of HIPAA and the Common Rule (45 CFR 46). Specific privacy approaches will include redacting instrument-generated testing dates from data files, obfuscation of the voxels around faces and ears in high-resolution structural scans ("defacing") (Milchenko and Marcus, 2013), and binning of subject ages into 5-year bands for Open Access data.

Data Sharing. We will coordinate data releases in close consultation with the U01 contributors and NIH program staff. We anticipate supporting 2-3 public releases for each project, with an initial release approximately one-quarter of the way through data collection and a final release at the completion of data collection. The exact frequency and timing will be dependent on the specific characteristics of each project. Project-specific Data Use Terms will be implemented in ConnectomeDB and users will be required to review and sign these prior to gaining access to the project data. Documentation similar to the structure and content of the HCP documentation (http://www.humanconnectome.org/documentation/) will be produced in collaboration with each contributor.

Specific Aim 3. Informatics Platform

The HCP informatics platform includes broad functionality to store, manage, share, and track the data collected by HCP and other programs. ConnectomeDB, an XNAT-based enterprise data repository, sits at the core of the HCP's platform (Figure 3). It includes a range of features – a searchable database, a high performance file archive, an ergonomic web application, and a high-speed data transfer system – that enables users to find, mine, and download HCP data. The platform also includes high performance redundant data storage and highly scalable computing systems. The CCF will assume responsibility for the operation and ongoing development of the HCP's informatics platform in order to continue the distribution of the HCP data and expand its capabilities to support import and sharing of data obtained by the U01 Connectomes of Human Disease program awardees.

Architecture. ConnectomeDB builds on the widely-used XNAT open source imaging informatics platform (**Figure 3**) (Marcus et al., 2007). XNAT is a web-based application designed to facilitate common management and productivity tasks for imaging and associated data. It consists of an image repository to store raw and post-processed images, a database to store metadata and non-imaging measures, and user interface tools for accessing, querying, visualizing, and exploring data. XNAT supports all common imaging methods, and its data model can be extended to capture virtually any related metadata. XNAT's web application provides a number of productivity features, including data entry forms, searching, reports of experimental data,

upload/download tools, image processing pipelines, and an online image viewer. A fine-grained access control system ensures that users are restricted to accessing only authorized data. XNAT also includes a web services API for programmatic access and an open plugin architecture for extending its core capabilities. ConnectomeDB uses XNAT's native extensibility to incorporate new web pages and other functionality.

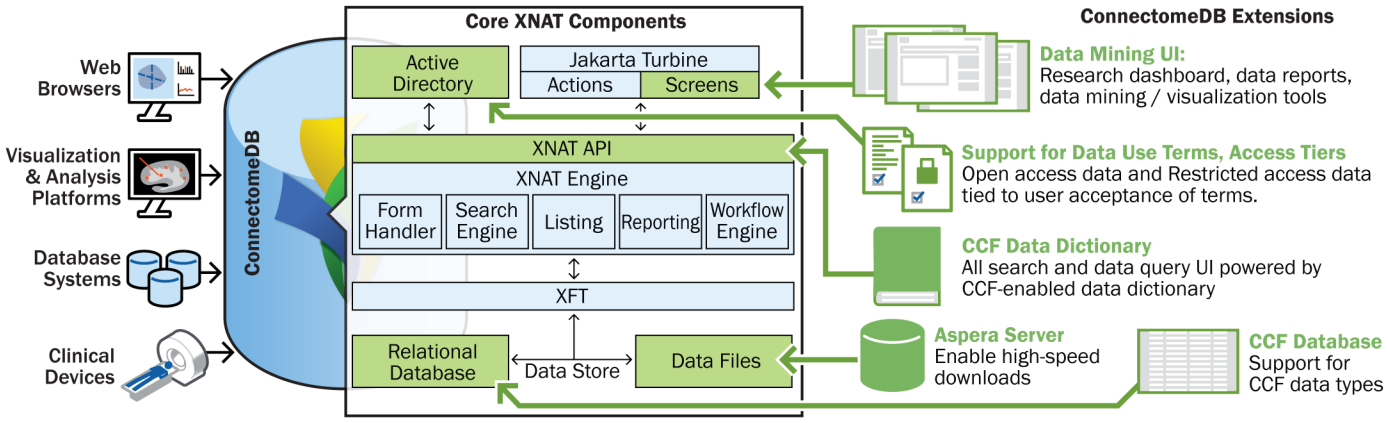

Figure 3. The ConnectomeDB system is built on the XNAT informatics platform by incorporating a customized user interface, custom data types, data mining tools, and a high speed download engine.

ConnectomeDB runs on a virtualized computing infrastructure that includes web and database servers, a compute cluster, and development machines as detailed in Resources component of this proposal. The data storage system currently includes over 1 PB of redundant capacity that will be expanded to host all CCF data. An offsite disaster recovery (DR) system hosted at the University of Minnesota provides backup of raw data and other files that cannot be easily regenerated. This system holds a matching 1PB of high-capacity storage and receives nightly backups from the primary storage system. In addition to the HCP's own computing capacity, the Washington University Center for High Performance Computing (CHPC), operated by the MIR's Electronic Radiology Laboratory, provides massively scaled compute resources to WashU investigators. The CCF will have privileged access to the CHPC's computing resources and through HCP operations we have experience running many thousands of jobs on it. Additionally, the CHPC recently received funding to expand their resources which will double its current compute capacity by adding approximately 1000 CPU cores and 8.0 TB of RAM along with 50 additional GPUs and a large shared storage pool. This expansive computing infrastructure will provide sufficient capacity to execute quality control and the adapted HCP pipelines on all of the data contributed to the CCF..

Data Dictionary. The HCP behavioral battery and study metadata include hundreds of individual data fields, and the U01 projects will likely add many more disease-specific fields. ConnectomeDB includes a data dictionary service that enables the meaning of these terms to be conveyed to users (Herrick et al., 2014). The dictionary includes a hierarchical structure with separate levels for categories, assessments, and individual data fields. The dictionary service enables client tools, such as the web interface, to query ConnectomeDB for details at each level of the dictionary. The dictionary is currently used primarily on the Data Dashboard (see below), where it is used to generate filters for searching the data. Within the scope of the CCF, we will implement a richer dictionary-based user interface that allows users to explore the data associated with each term, including through dynamic charts and graphs, statistical models to identify covariates, genetic association models, and links to external resources associated with the term. A critical feature of the dictionary service is that it provides a common ontology to harmonize the various data obtained across the distinct data sets that will be managed within ConnectomeDB. The CCF will work with the funded U01 projects to ensure that the data definitions developed for each project are harmonized to a common ontology. Thus, a user will be able to search for and merge data sets that include common behavioral and imaging measures. Whenever possible, the CCF will use standard ontologies developed by the Neuroscience Information Framework (http://www.neuinfo.org) and other organizing bodies.

Data Packages. The HCP preprocessing pipelines generate a large number of files, many of which are not essential for typical analyses but are retained in the database for quality control and review purposes. To streamline distribution of the high interest files, a file package generation service was developed that allows specific subsets of the files hosted in ConnectomeDB to be compiled into zip-formatted distributable file

archives. Packages are typically defined to include contents for a single subject, but group level data can also be packaged. The vast majority of HCP data distributed through ConnectomeDB has been in the form of these prebuilt packages. The CCF will continue using the package service to distribute data provided by contributors.

User Interface. ConnectomeDB includes a modified version of the standard XNAT navigation structure. A landing page details the available data and guides the user through a process to review and accept project-specific data use terms. Once the user agrees to the data use terms for a project, he is automatically granted access to the open access components of the data. For each project to which the user has obtained access, links are provided to navigate to a data set-specific Data Dashboard screen, which provides a searchable view of all subject metadata, and the Package Download screen, which provides a streamlined interface for selecting and downloading relevant data packages. Users can also access the standard XNAT reports and navigation elements to explore individual subjects. These pages have been customized to reflect the specific metadata and data organization of the HCP. They will be further adapted to meet the additional needs associated with the new U01 Disease Connectome projects.

ConnectomeDB's *Data Dashboard* provides an interface for dynamic exploration of the HCP dataset (**Figure 4**, **left**). Interactions on the dashboard are built around the concept of subject groups – subsets of the HCP data set who match some specified criteria. The dashboard includes several components for interacting with subject groups: a tool ribbon provides access to various actions, including saving subject groups, opening previously saved groups, and downloading data for the current subject group; data filters for defining the subject group criteria; and a tabular display of demographics, behavioral measures, subject data and metadata for the currently selected group. The dashboard components use the data dictionary, with the filter selectors and tabular data displays mapping directly to the dictionary hierarchy elements and the data tables drawing from the dictionary categories.

ConnectomeDB's *Download Packages* interface utilizes the Data Package service described above to enable users to download individual subject or subject group data selected through the Data Dashboard (**Figure 4, right**). The interface presents users with a filterable list of packages for unprocessed and preprocessed versions of each of the modalities included in the HCP protocol, including the structural, diffusion, fMRI resting state, and each fMRI task. For each package type, details of the package contents, including number of files and total size, are displayed along with information about the availability of the files for each subject in the selected subject group. Once the set of packages has been selected, the user is directed to a download interface that uses the commercial Aspera fasp high speed data transfer technology. Aspera's fasp is a proprietary transport built on the datagram-oriented Internet standard protocol UDP and designed to maximize throughput on high latency networks. In our testing, Aspera fasp proved to be 10-12 faster than standard HTTP- and FTP-based data transfer from Washington University to typical remote academic centers. To date, over 850 TB of HCP data have been distributed using fasp. CCF will leverage this proven technology to enable high-speed download of project-specific data packages.

Security. ConnectomeDB is designed to protect the privacy of study participants and to restrict access to authorized users. Users are required to create password-protected accounts and to login to the site each time they visit. Each data set within ConnectomeDB is contained within an XNAT project, which permits access only to those users who have been explicitly granted access to the project. ConnectomeDB extends the standard XNAT access control levels by incorporating multiple tiers of access within a project to enable particularly sensitive data fields (e.g., history of drug use) to require a secondary level of authorization. Project-specific use terms can be specified for each tiered level of access.

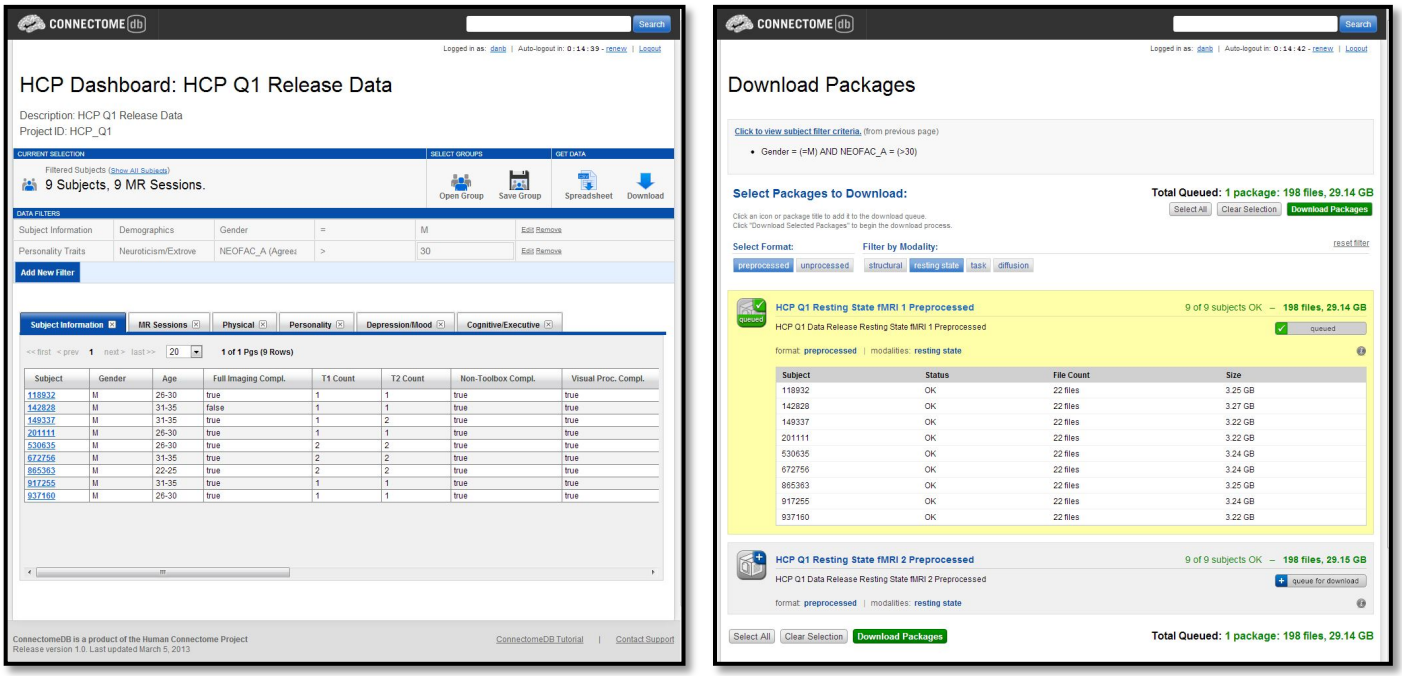

Figure 4. Left. The ConnnectomeDB project-specific Data Dashboard allows users to navigate and filter the data. **Right**. The Download Packages interface allows users to select specific packages to download and links out to the Aspera high-speed download service.

Data Mining and Visualization. The HCP informatics team is currently developing new features on the data dashboard to dynamically display HCP measures in configurable charts and graphs. In addition, a service is being piloted that allows users to submit analytic scripts to be executed on the HCP servers. The current pilot links to the SOLAR-Eclipse software package to conduct heritability analyses on any of the HCP variables. These features allow users to mine the HCP data without necessarily downloading the data set or even having access to sensitive and restrictive data fields at the individual subject level. The CCF will expand on these features to include display of connectomics data on an embedded web-based surface viewer and improved linkage to Connectome Workbench.

Public cloud computing. In addition to distribution of data through ConnectomeDB, the HCP's data packages are also available through Amazon Web Services' Public Data Sets (PDS) program. The data hosted on PDS can be accessed by users via Amazon services such as Elastic Compute Cloud (EC2) and Simple Storage Service (S3) at no cost to users. These data are ideally situated for users who utilize EC2 to execute image processing and analysis routines. Through the availability of pre-configured EC2 virtual machines such as NeuroDebian and the NITRC Compute Environment that are being adapted to include the HCP pipelines, we anticipate that more neuroimaging researchers will move to the EC2 cloud to accomplish their work. We therefore will host CCF packages via the PDS program (see letter of support from Jamie Kinney, Sr, Manager of Scientific Computing at Amazon). Amazon has agreed to partner with the CCF to provide training opportunities for CCF users to learn how to use the Amazon cloud for connectomics research. All data hosted on PDS will only be accessible to users after they have signed the project-specific data use terms through ConnectomeDB.

Connectome in a Box. Owing to the massive size of connectomics data, downloading large portions of the data over the Internet is at the limits of current network bandwidth and latency constraints, particularly for international locales. As an alternative to network-based downloads, we have developed a physical data transport mechanism that we refer to as Connectome in a Box (CinaBox). CinaBox is based on high capacity (currently 4 TB) consumer grade hard drives on which the HCP unprocessed, preprocessed, and analysis data are loaded. The current HCP 500 Subject data release includes 17 TB of imaging data and spans 5 hard drives. Users with limited processing resources can order a single drive starter kit containing the HCP's Unrelated 100 Subjects (UR100). Using a custom built duplication system and program built around RSync, we are able to generate and verify replication accuracy of up to 15 simultaneous drives in approximately 40 hours. Copies of CinaBox can be ordered via the HCP website along with an optional external drive enclosure

for the UR100 and are delivered at cost to customers via FedEx. CinaBox data are organized in the same directory structure as data downloaded directly from ConnectomeDB.. The CinaBox program has proven to be very popular with the research community. To date over 1250 TB of data have been shipped to 78 domestic and 53 international customers. The CCF will expand the CinaBox program to allow users to order project-specific drive shipments.

Program wide activities

Project Administration and Project Management.
Most of the individual project activities of successful U01 Disease Connectome applicants will be administered and managed by the awardees. However, it is vital that the CCF include strong project administration and project management components, in order to ensure that there is 1) program-wide (i.e., Disease Connectome-wide) understanding of the CCF and how it operates; 2) program-wide agreement on specific details such as data ontology and file naming conventions that are needed to make the data's comparability apparent to users; and 3) cross-project scheduling of data releases by CCF to ensure that each project is able to share their data as closely as possible to their desired data sharing schedules.

Once the first round of U01 awardees is announced, the CCF will establish and convene a CCF Steering Committee that includes each U01 PI and project manager (PM). The Steering Committee will also include Drs. Van Essen (chair), Marcus, Ugurbil, Yacoub, and will meet as needed to work through program-wide issues. We anticipate that the Steering Committee will decide on necessary harmonization-related policies and will then task the PM group to implement the policies, facilitated by the CCF senior PM (Dr. Sandra Curtiss) and IT PM (Eileen Cler). This CCF Steering Committee will also serve as a focus group to ensure that the CCF is serving the needs of its U01 partners.

Outreach/education
It will be critically important that the neuroscience community becomes aware of and is able to appreciate the nature of the data available through the CCF from each of the U01 awardees. Therefore, CCF outreach efforts will focus on (i) documentation of the types of available data and their comparability; and (ii) informing the neuroscience community about the CCF and ConnectomeDB via webinars and through outlets such as the Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC) website, the HCP's very active email discussion, and the HCP's public wiki pages.

Timelines
In our experience with the WU-Minn HCP, it is easy to underestimate the amount of work and time necessary to process, document, and publicly release high-quality imaging and behavioral/clinical datasets. Therefore, as described above, we will begin data acquisition and release discussions with U01 awardees soon after the awards are announced. We anticipate that some U01 projects may be ready to share data as early as the beginning of Year 2, i.e. in the summer of 2016. The CCF will strive to meet the data release timelines desired by the U01 awardees, while maintaining the high data quality and excellent documentation for which the WU-Minn HCP is known.

Institutional support The original WU-Minn HCP grant submission included major institutional support from multiple sources. This included $900,000 of support from the Dean of Washington University School of Medicine to be allocated during the time of the grant, plus additional support to be allocated for an additional five years after the conclusion of the grant in order to support and maintain the HCP informatics platform, at an "estimated additional encumbrance of $937,000". We are very pleased that the Dean has reaffirmed this commitment of $937,000 over the period 2015-2020 (see December 4, 2014, letter from Dean Larry J. Shapiro).