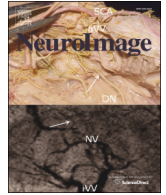




Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Q1 ConnectomeDB—Sharing human brain connectivity data[☆]

Q2 Michael R. Hodge^{a,*}, William Horton^a, Timothy Brown^a, Rick Herrick^a, Timothy Olsen^b, Michael E. Hileman^a,
 3 Michael McKay^a, Kevin A. Archie^a, Eileen Cler^a, Michael P. Harms^c, Gregory C. Burgess^c, Matthew F. Glasser^d,
 4 Jennifer S. Elam^d, Sandra W. Curtiss^d, Deanna M. Barch^c, Robert Oostenveld^e, Linda J. Larson-Prior^{a,f},
 5 Kamil Ugurbil^g, David C. Van Essen^d, Daniel S. Marcus^a

6 ^a Department of Radiology, Washington University School of Medicine, St. Louis, MO, USA

7 ^b Deck5 Consulting, Normal, IL, USA

8 ^c Department of Psychiatry, Washington University School of Medicine, St. Louis, MO, USA

9 ^d Department of Anatomy and Neurobiology, Washington University School of Medicine, USA

10 ^e Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands

11 ^f Department of Neurology, Washington University School of Medicine, St. Louis, MO, USA

12 ^g Center for Magnetic Resonance Imaging, University of Minnesota, Minneapolis, MN, USA

1 3 A R T I C L E I N F O

14 *Article history:*
 15 Accepted 21 April 2015
 16 Available online xxxx

17 *Keywords:*
 18 Human Connectome Project
 19 Neuroinformatics databases
 20 Open access
 21 Data sharing
 22 Connectomics
 23 Connectome Coordination Facility
 24 XNAT

A B S T R A C T

ConnectomeDB is a database for housing and disseminating data about human brain structure, function, and connectivity, along with associated behavioral and demographic data. It is the main archive and dissemination platform for data collected under the WU-Minn consortium Human Connectome Project. Additional connectome-style study data is and will be made available in the database under current and future projects, including the Connectome Coordination Facility. The database currently includes multiple modalities of magnetic resonance imaging (MRI) and magnetoencephalography (MEG) data along with associated behavioral data. MRI modalities include structural, task, resting state and diffusion. MEG modalities include resting state and task. Imaging data includes unprocessed, minimally preprocessed and analysis data. Imaging data and much of the behavioral data are publicly available, subject to acceptance of data use terms, while access to some sensitive behavioral data is restricted to qualified investigators under a more stringent set of terms. ConnectomeDB is the public side of the WU-Minn HCP database platform. As such, it is geared towards public distribution, with a web-based user interface designed to guide users to the optimal set of data for their needs and a robust backend mechanism based on the commercial Aspera *fastp* service to enable high speed downloads. HCP data is also available via direct shipment of hard drives and Amazon S3.

© 2015 Elsevier Inc. All rights reserved.

44 Introduction

45 ConnectomeDB (<https://db.humanconnectome.org>) is a database for
 46 housing and disseminating publicly available human brain connectivity
 47 data. It is a highly customized instance of the XNAT imaging informatics
 48 platform: an extensible, open source platform for managing and sharing
 49 imaging and related data (Marcus et al., 2007). ConnectomeDB is
 50 designed as the database and dissemination platform for the Human
 51 Connectome Project (HCP) consortium led by Washington University,
 52 University of Minnesota, and Oxford University (the WU-Minn HCP

53 consortium), and it houses and distributes data collected under the
 54 WU-Minn HCP (Van Essen et al. 2013). In 2014, it expanded to
 55 include diffusion data collected by the USC-MGH HCP consortium
 56 (Setsompop et al., 2013; Toga et al., 2012) and multimodal data
 57 collected under the WU-Minn HCP consortium LifeSpan Pilot Project,
 58 which is designed to provide information on the sensitivity of HCP
 59 methods to age-related differences. Current projects available in
 60 ConnectomeDB and details about their imaging data are found in
 61 Table 1. These projects have focused on providing normative data
 62 on healthy populations. However, ConnectomeDB will soon expand its
 63 portfolio to include data obtained through additional NIH-funded
 64 initiatives, including Connectomes of Human Diseases plus three
 65 Lifespan-HCP efforts. A Connectome Coordination Facility (CCF)
 66 centered at Washington University and also involving the Univer-
 67 sity of Minnesota is being established to operate the expanded
 68 ConnectomeDB repository and provide support for the groups
 69 running these projects.

[☆] Author note: Funded in part by the Human Connectome Project, WU-Minn consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 5U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. Additional funding was provided by 5R01EB009352 for XNAT support and 5P30NS048056 for the NIAC.

* Corresponding author.

E-mail address: hodgem@mir.wustl.edu (M.R. Hodge).

Table 1
Imaging data currently available/planned for release in ConnectomeDB.

Project	Imaging type/scanner	Modality	Details	Processing levels
WU-Minn HCP (currently 526 subjects with imaging data)	3 T MRI	Structural	T1 weighted T2 weighted	Unprocessed, preprocessed
		Resting state fMRI	Two sessions of 2 scans (1200 frames each per scan). Each session contains a pair of scans, acquired with opposing phase encoding directions (LR/RL)	Unprocessed, preprocessed Group analysis (functional connectivity maps/timeseries, ICA-based parcellation and network matrices)
		Task fMRI	Tasks: <ul style="list-style-type: none">• Working memory (405 frames/scan)• Gambling (253 frames/scan)• Motor (284 frames/scan)• Language (316 frames/scan)• Social cognition (274 frames/scan)• Relational processing (232 frames/scan)• Emotion processing (176 frames/scan)	Unprocessed, preprocessed, analysis (individual subject and group analyses)
	7 T MRI	Diffusion MRI	One scan pair for each task with 1 scan in each phase encoding direction (LR/RL) 6 scans total (three gradient tables with 1 scan in each phase encoding direction per gradient table (LR/RL))	Unprocessed, preprocessed, analysis (to be released)
		Resting state fMRI	4 scans total. 900 frames/scan. Two scans in each phase encoding direction (AP/PA)	To be released
		Task fMRI	Tasks: <ul style="list-style-type: none">• Retinotopy (6 scans for mapping retinotopy, using rotating wedges, expanding/contracting rings, and drifting bars. Three scans in each phase encoding direction (AP/PA))	To be released
		Diffusion MRI	Movie (4 scans using 4 stimulus movies). Two scans in each phase encoding direction (AP/PA) 4 scans total (two gradient tables with 1 scan in each phase encoding direction per gradient table (AP/PA))	To be released
		MEG	Noise <ul style="list-style-type: none">• Empty room (one 5 min scan)• Patient noise (one or more 1 min scan)	Unprocessed
		Resting state	Three 6 minute scans	Unprocessed, preprocessed, source-level processed (time-series and connectivity data)
		Task	Tasks: <ul style="list-style-type: none">• Working memory (two 10 minute scans)• Story math (two 7 minute scans)• Motor (two 14 minute scans)	Unprocessed, preprocessed, source-level processed (averaged event-related and time-frequency responses)
WU-Minn LifeSpan Pilot (Currently 27 subjects with imaging data)	3 T MRI	Structural	T1 weighted T2 weighted	Unprocessed
		Task fMRI	Tasks: <ul style="list-style-type: none">• Working memory (405 frames/scan)• Emotion processing (199 frames/scan)• Social cognition (274 frames/scan)• Gambling (253 frames/scan)	Unprocessed
	Diffusion fMRI	One scan pair for each task with 1 scan in each phase encoding direction (LR/RL) 4 scans total (two gradient tables with 1 scan in each phase encoding direction per gradient table (LR/RL))	Unprocessed	

70 What is available?

71 WU-Minn HCP data—overview

72 Data collected under the WU-Minn HCP project includes multiple
73 modalities of imaging data, along with a large battery of behavioral data
74 spanning numerous physical, behavioral, and personality dimensions.
75 The final HCP dataset is expected to include data obtained from 1200

participants. The HCP population is a “healthy” population of twins and 76
siblings aged 22–35, where “healthy” refers primarily to absence of 77
conditions likely to affect brain structure and function or influence the 78
ability to successfully complete study protocol. The HCP sample is 79
described in further detail in Van Essen et al. (2013). As of January 81
2015, the consortium has shared data on over 500 subjects as part of its 81
“500 Subjects + MEG2” release. This includes high resolution 3 T MRI 82
session data on 526 subjects, MEG data on 67 subjects, and demographic 83

84 and behavioral data on 542 subjects. Imaging data and much of the
85 behavioral data collected under the WU-Minn HCP project are freely
86 and publicly available, subject to data use terms. Some sensitive
87 behavioral data elements are restricted to qualified researchers
88 under a more rigid set of data use terms.

89 WU-Minn HCP data—MRI

90 WU-Minn HCP MRI data released to date have been collected on a
91 customized Siemens MAGNETOM Connectom 3 T scanner at WU in
92 multiple imaging sessions covering four modalities: structural (T1w
93 and T2w), resting state fMRI (rfMRI), task fMRI (tfMRI, 7 tasks) and
94 diffusion MRI (dMRI). For a subset of 200 subjects, 3 T imaging data is
95 being supplemented by data collected on a 7 T MRI scanner. The 7 T
96 data will include resting state, diffusion, and two additional tasks
97 (retinotopy and movie clips).

98 MRI imaging data released under the WU-Minn HCP project includes
99 unprocessed, minimally preprocessed, and analysis data. For MRI, the
100 unprocessed data includes NIFTI files (with facial features anonymized;
101 Milchenko and Marcus, 2013) from session scans along with scan-
102 associated data (e.g. task timing and physiological monitoring files) in
103 text format. Preprocessed data contains the output of the minimal
104 preprocessing (MPP) pipelines, described in detail by Glasser et al.
105 (2013). These pipelines minimize spatial distortions in the images,
106 correct for subject motion, align data across modalities and bring the
107 data into a common atlas space, thereby preparing the data for further
108 processing. Preprocessed data also includes resting-state fMRI data
109 denoised by the ICA-based “FIX” method, which greatly reduces spatially
110 and temporally structured noise (Griffanti et al., 2014; Salimi-Khorshidi
111 et al., 2014). Currently available individual subject analysis data includes
112 results of the WU-Minn HCP task analysis pipeline. WU-Minn HCP
113 pipeline scripts and documentation are available on GitHub at [https://](https://github.com/Washington-University/Pipelines/wiki)
114 github.com/Washington-University/Pipelines/wiki. Additionally available
115 are more extensively analyzed group-average rfMRI datasets: full corre-
116 lation “dense” (grayordinate to grayordinate) functional connectomes,
117 rfMRI independent component analysis (ICA)-derived parcellations,
118 plus group-average and single-subject node time series and network
119 functional connectivity matrices.

120 Many of the files HCP releases are quite large, due to the high spatial
121 and temporal resolution attained using “multi-band” data acquisition
122 (Ugurbil et al., 2013), which enables 2 mm isotropic voxels and a
123 0.72 s TR (‘frame rate’) for fMRI (compared to conventional scans that
124 typically are ~3 mm voxel size and ~2 s TR) and 1.25 mm voxels for
125 dMRI (vs conventional voxel size of ~2 mm). The HCP MPP pipelines
126 are customized to handle such large datasets efficiently, for example,
127 by generating ‘grayordinate’ representations of fMRI data that include
128 only cortical surface vertices and subcortical gray-matter voxels
129 (Glasser et al., 2013), using a standardized ‘CIFTI’ data format ([http://](http://www.nitrc.org/projects/cifti)
130 www.nitrc.org/projects/cifti). The grayordinate-based files are much
131 more compact than standard NIFTI volumes (but both formats are
132 released in order to provide flexibility). Connectome Workbench
133 software ([http://www.humanconnectome.org/software/connectome-](http://www.humanconnectome.org/software/connectome-workbench.html)
134 [workbench.html](http://www.humanconnectome.org/software/connectome-workbench.html)) is customized for visualizing and analyzing HCP data
135 and it capitalizes especially on the grayordinates/CIFTI data representa-
136 tions (see <http://www.humanconnectome.org/documentation/tutorials>).

137 WU-Minn HCP data—MEG

138 MEG data released under the WU-Minn HCP project includes unpro-
139 cessed, anatomical and channel-level preprocessed, and source-level
140 processed functional data. Unprocessed data consists of 16-bit raw binary
141 c,rFDc files from the 4D scanner, supplemented by quality control (QC)
142 figures and ascii text files. Stimulus-response files derived from E-Prime
143 are included for task modalities. Anatomical preprocessed data includes
144 individual anatomical models for volume conduction and source
145 modeling. Coordinate transformation matrices are included for

translating between MEG-system coordinates and MRI-based individual
146 and normalized coordinate systems. Anatomical and channel-level
147 preprocessed data are represented in MATLAB (Mathworks, Natick,
148 MA) format. Source-level processed data are represented in CIFTI format
149 and contain source-reconstructed output from multiple processing pipe-
150 lines. All processed data are supplemented by QC figures and provenance
151 details. The “megconnectome” processing software is implemented in
152 MATLAB using the FieldTrip toolbox (Oostenveld et al., 2011) and is
153 made available along with the data in each release. MEG task details,
154 preprocessing and processing pipelines, and associated outputs
155 are described in detail in Larson-Prior et al. (2013) and in HCP
156 documentation. 157

WU-Minn HCP data—behavioral and other individual difference data 158

159 Along with imaging data, a wide array of behavioral and other non-
160 imaging data is obtained under the HCP, with an emphasis on obtaining
161 standardized measures that may covary with brain structure and
162 function. These data are available in ConnectomeDB and can be
163 downloaded as CSV files. The core components of these data are
164 implementations of the NIH Toolbox (<http://www.nihtoolbox.org/>)
165 and a modified web-based battery that includes components of the
166 Penn Neurocognitive Battery (Gur et al., 2001,2010) as well as additional
167 measures. These implementations assess many domains including
168 cognition, emotion, motor, sensory, visual processing and personality.
169 These core components are supplemented with additional instruments
170 supplying information on other areas such as psychiatric history,
171 substance use and family history. These non-imaging data are described
172 in detail in Barch et al. (2013, Table 2) and Van Essen et al. (2013) and in
173 project documentation. A list of categories of non-imaging data and
174 associated instruments are provided in Table 2. 174

Quality control process 175

176 All imaging data collected under the WU-Minn consortium goes
177 through an extensive automated validation process. In addition to this
178 validation process, fMRI data goes through an automated QC process,
179 while structural MRI and MEG data are submitted to additional manual
180 QC processes. Furthermore, data collection staff follow detailed standard
181 operating procedures (SOPs); they are trained to identify excessive
182 movement and other issues that might affect data quality during scanning
183 and to attempt rescans when appropriate. The HCP acquisition and QC
184 process and pipelines are described in detail in Marcus et al. (2013). 184

185 Upon transfer of data from scanners to the internally-facing ‘IntraDB’
186 database (see below), data are sent through validation and QC pipelines.
187 Validation pipelines perform initial checks, with MRI utilizing information
188 from acquisition metadata embedded in the DICOM header (e.g., number
189 of slices, resolution, TR, TE, flip angle), to ensure that data was acquired
190 according to protocol. Structural scans are accepted for further processing
191 if they receive at least a good rating in the manual QC four point rating
192 scale (excellent, good, fair, poor; see Marcus et al., 2013). Following
193 validation, a second round of pipelines are run, performing a more
194 in-depth QC analyses on the fMRI data. These pipelines analyze signal to
195 noise ratios, search for motion outliers and compute other measures
196 affecting data quality, producing graphs and summary images to help in
197 the evaluation of image quality. However, the HCP rarely excludes fMRI
198 data from release solely due to motion (e.g., only in cases of extremely
199 bad motion). Users should be diligent in dealing with the impact of
200 motion in their analyses and are strongly encouraged to make use of
201 the aforementioned FIX-denoised rfMRI datasets or to carry out other
202 denoising strategies. 202

203 Data obtained under the WU-Minn LifeSpan pilot projects undergo
204 nearly identical acquisition and QC processes as those for the young
205 adult Human Connectome Project. The LifeSpan pilot projects involve
206 similar scanning protocols as the young-adult HCP but are shorter in
207 duration (~2 h total scan duration instead of ~4 hr). Owing to the 207

t2.1	Table 2		
t2.2	WU-Minn HCP non-imaging and behavioral data.		
t2.3	Category	Instrument	
t2.4	Subject information	Demographics (gender, age, twin status, zygosity, mother ID, father ID, race, ethnicity, handedness, employment status, household income, education, school status, relationship status)	211
t2.5	Study completion	Study completion: 3 T MR Image reconstruction version: 3 T MR Study completion: MEG Study completion: behavioral	212
t2.6	MR sessions	Session information	
t2.7	Health and family history	Physical health (height, weight, BMI, hematocrit, blood pressure, thyroid stimulating hormone levels, glucose levels, endocrine disorders) (self-report) Menstrual cycle information (in females) Parental history of psychiatric and neurologic disorders	
t2.8	Alertness	Cognitive status (mini mental status exam) Sleep: (Pittsburgh Sleep Quality Index)	
t2.9	Cognition	Episodic memory (picture sequence memory) Executive function/cognitive flexibility (dimensional change card sort) Executive function/inhibition (flanker task) Fluid intelligence (Penn progressive matrices) Language/reading decoding (oral reading recognition) Language/vocabulary comprehension (picture vocabulary) Processing speed (pattern completion processing speed) Self-regulation/impulsivity (delay discounting) Spatial orientation (variable short Penn line orientation test) Sustained attention (short Penn continuous performance test) Verbal episodic memory (Penn word memory test) Working memory (list sorting)	
t2.10	Emotion	Emotion recognition (Penn emotion recognition test) Negative affect (sadness, fear, anger) (self-report) Psychological well-being (positive affect, life satisfaction, meaning and purpose) (self-report) Social relationships (social support, companionship, social distress, positive social development) (self-report) Stress and self-efficacy (perceived stress, self-efficacy) (self-report)	
t2.11	FreeSurfer	FreeSurfer summary statistics Volume (subcortical) segmentation Surface area Surface thickness	
t2.12	In-scanner task performance	Emotion processing task Gambling task	
t2.13	(reaction time and accuracy)	Language processing task Relational processing task Social cognition task Working memory task	
t2.14	Motor	Endurance (2 minute walk test) Locomotion (4-meter walk test) Dexterity (9-hole pegboard) Strength (grip strength dynamometry)	
t2.15	Personality	Five factor model (NEO-FFI 60)	
t2.16	Psychiatric and life function	Life function (Achenbach adult self-report, syndrome scales and DSM-oriented scale) Psychiatric history	
t2.17	Sensory	Audition (words in noise) Olfaction (odor identification test) Pain (pain intensity and interference surveys) Taste (taste intensity test) Vision (EVA scores and Farnsworth test) Contrast sensitivity (Mars contrast sensitivity)	
t2.18	Substance use	Breathalyzer and drug test results Alcohol use 7-day retrospective Alcohol use and dependence Tobacco use 7-day retrospective Tobacco use and dependence Illicit drug use Marijuana use and dependence	
t2.19			
t2.20			
t2.21			
t2.22			
t2.23			
t2.24			
t2.25			
t2.26			
t2.27			
t2.28			
t2.29			
t2.30			
t2.31			
t2.32			
t2.33			
t2.34			
t2.35			
t2.36			
t2.37			
t2.38			
t2.39			
t2.40			
t2.41			
t2.42			
t2.43			
t2.44			
t2.45			
t2.46			
t2.47			
t2.48			
t2.49			
t2.50			
t2.51			
t2.52			
t2.53			
t2.54			
t2.55			
t2.56			
t2.57			
t2.58			
t2.59			
t2.60			
t2.61			
t2.62			
t2.63			
t2.64			
t2.65			
t2.66			
t2.67			
t2.68			
t2.69			
t2.70			

208 pilot nature of these projects, the inclusion of both young and older
209 subjects (often with concomitantly greater head movements), and a
210 desire to maximize sample size for pilot analyses, structural scans are

accepted for further processing if they receive at least a fair rating in
the aforementioned manual QC four point rating scale.

The WU-Minn database platform and data processing 213

ConnectomeDB is the public-facing side of the WU-Minn consortium
database platform. As such, it is designed for dissemination of the data
and contains only released data or those data currently being processed
for release. WU-Minn data acquisition and QC staff use a separate, internal
database, IntraDB that (like ConnectomeDB) is a customized version of
XNAT. It is in IntraDB that quality determinations are made to inform
what data will ultimately be transferred to ConnectomeDB for final
processing and dissemination. Pipelines in IntraDB convert images from
DICOM to NIFTI, while de-identifying subject data by masking facial
structures in the image and removing data such as session dates.
Identifying information like name, birthdate and address are never
stored in either database.

Once a subject's data collection is completed and quality determina-
tions are made, another set of programs are run to "sanity check" subject
data, checking mainly for data completeness, incomplete processing, and
corrupted files. Data must pass these checks before transfer from IntraDB
to ConnectomeDB, where usable data from all scan sessions are joined
into a single combined representation of the subject's imaging data. This
process includes modifications of file names and/or directory structure
from that which is optimized for the incoming data stream on the IntraDB
side to that which is optimized for preprocessing and data sharing on the
ConnectomeDB side. Separate session representations are built and
customized for data obtained from different scanners (i.e., 3 T MRI, 7 T
MRI and 4D MEG). Once data is in place in ConnectomeDB, processing
pipelines are run and data are packaged and prepared for distribution.

Accessing the data 239

All imaging-related data and much of the non-imaging data in
ConnectomeDB are freely and publicly available, subject to the user's
agreement to open access data use terms via their ConnectomeDB
user account. Accounts require validation by responding to a validation
link e-mailed upon account registration. Acceptance of terms for publicly
available data can be performed via the ConnectomeDB website immedi-
ately after account registration. Some non-imaging data are considered
"restricted access", because they are either "sensitive" or have potential
to reveal participant identity to family members or others. These data
are made available only to qualified investigators under more stringent
data use terms (Van Essen et al., 2013). Each project or dataset contained
in ConnectomeDB may have its own sets of data use terms required for
accessing public and restricted data. Before gaining access to restricted
data, a signed form sent to the WU-Minn HCP consortium administration
must be approved.

Upon login, users are directed to a unified page showing all publicly
available datasets in ConnectomeDB along with current access level
based on data use terms acceptance. From this page, users can initiate
acceptance of data use terms, launch project pages (containing dataset
resources and group average data), or launch a subject dashboard
from which they can engage with and/or download project data.

Imaging data 261

Imaging and associated data (e.g., task timing files) are made available
via data "packages" in ConnectomeDB. These packages organize data
into meaningful groups of manageable size. For each subject, separate
packages are created for each modality and processing level. Packages
are further divided for task fMRI by task paradigm and for resting state
fMRI by resting state scan pair. Packages are designed to contain all data
files necessary to perform the expected types of additional processing
and analyses that users might perform on those files while excluding
unnecessary processing intermediates that would just increase package

271 size. For example, unprocessed data packages contain the files necessary
 272 to run preprocessing pipelines, and preprocessed packages supply the
 273 files necessary to run additional HCP analysis pipelines. This approach
 274 does entail modest duplication of files in the supplied packages; for
 275 example, packages for different tasks may contain the same field map
 276 image files for distortion correction.

277 To facilitate robust, high speed downloads, ConnectomeDB uses a
 278 commercial UDP-based data transfer technology called Aspera *fast*TM
 279 (<http://asperasoft.com>). This technology allows transfer rates much
 280 faster than those achievable by TCP-based technologies such as FTP and
 281 HTTP and enables reliable transfers with retry-and-resume capability. A
 282 downloadable plugin is required to enable Aspera-based transfers. This
 283 client-side software allows for user controlled management of the
 284 download process and is available to users free of charge.

285 *Non-imaging data*

286 Non-imaging data can be searched and filtered in ConnectomeDB
 287 and is downloadable in CSV format. Links to download public
 288 non-imaging data are available in ConnectomeDB for all users who
 289 have registered and accepted the appropriate open access data use
 290 terms; links to download restricted CSV files are available only to
 291 users approved for restricted access. Users with restricted access
 292 permissions can switch between “Open Access”, “Restricted”, and
 293 “Sensitive” views in the ConnectomeDB user interface (UI). Once in
 294 the “Restricted” or “Sensitive” views, restricted data, associated fields,
 295 and download links become viewable in the UI.

296 In addition to the behavioral data collected, metadata about the
 297 imaging and non-imaging data are available in the DB. These data mainly
 298 provide information about the existence of and completeness of the
 299 various types and modalities of data available for individual subjects.
 300 These can be useful when filtering for subjects having complete data in
 301 the desired modality or domain (see next section).

302 *Groups, filtering and the data dictionary*

303 The HCP data stored within ConnectomeDB are immense and
 304 growing. The current “HCP S500” data release includes 17 terabytes
 305 of MRI data and approximately 3 terabytes of MEG data, including
 306 data from all processing levels. The size of these datasets will grow as
 307 additional subjects are added and additional processing and analysis
 308 output is made available.

309 Because of the size of the growing HCP database, it is important to
 310 provide users with tools to identify and obtain the minimum dataset
 311 required to meet their needs and conduct follow-up analyses. The
 312 organization of data into packages helps in this regard by facilitating
 313 selection of subsets of data within each subject, but it is also important to
 314 enable users to identify the optimal group of subjects for which to obtain
 315 data. To accomplish this, ConnectomeDB uses groups and filters. Groups
 316 are subsets of subjects, and ConnectomeDB supports both pre-defined
 317 and user-defined groups. ConnectomeDB has identified several meaning-
 318 ful subsets of subjects and has organized them into pre-defined groups.
 319 One example is the “Single Subject” group, which enables users to get a
 320 good feel for what the data contains by obtaining a representative subject
 321 with complete data. Another is the “100 Unrelated Subjects” group, which
 322 provides users with a practical amount of data for analysis without the
 323 need to correct for family structure.

324 In addition to preset groups, the UI supports filtering of subjects
 325 based on any project data field (e.g., gender, completeness of imaging,
 326 motor assessment results). These filtered sets of subjects may be
 327 saved as “user-defined groups” which the user can save and return to
 328 in later sessions. The filtering functionality is enabled and supported
 329 by the ConnectomeDB data dictionary (Herrick, et al., 2014).

330 The ConnectomeDB data dictionary contains metadata about all
 331 non-imaging data fields in ConnectomeDB as well as many fields that
 332 are derived by analysis of the imaging data (e.g., FreeSurfer results).

333 These metadata, along with additional metadata used within the
 334 application, are assembled into the data dictionary used by the DB.
 335 Dictionary entries include information about data type, differing levels
 336 of description, expected values and other useful information about
 337 ConnectomeDB data.

338 **Multiple methods of access**

339 As noted above, HCP data is immense and growing. It is expected
 340 that many users of HCP data will be able to work with one of the
 341 predefined groups or a filtered subset of subjects. Furthermore, such
 342 users will likely be interested only in some subset of the packages
 343 (e.g., preprocessed diffusion data). ConnectomeDB and the Aspera
 344 UDP-based download mechanism serve these users well. Other users
 345 will prefer to have access to all of the available data. While it might be
 346 possible to eventually download all the data via ConnectomeDB, these
 347 users are generally better served by one of the alternative access
 348 methods, Connectome-In-A-Box and Amazon S3 cloud storage.

349 Connectome-In-A-Box ([http://www.humanconnectome.org/data/
 connectome-in-a-box.html](http://www.humanconnectome.org/data/connectome-in-a-box.html)) is a process that sends a set of hard drives
 350 containing the entire HCP imaging database from a specific project or
 351 subproject to requesting users who have accepted data use terms.
 352 Currently available is a set of drives containing the 3 T MRI data for
 353 the entire HCP 500 subject release, or a single drive option with all the
 354 3 T MRI data for the U100 group. Charges for this service recover only
 355 the costs of the physical hard drives and delivery.

356 In addition to local storage and availability via ConnectomeDB, a
 357 copy of the entire Human Connectome Project database has recently
 358 been made available via the cloud through Amazon S3. Users can access
 359 the data in Amazon S3 through ConnectomeDB in a process that links
 360 their ConnectomeDB credentials with their Amazon S3 account
 361 credentials. Users can then access HCP data directly from the cloud.
 362

363 **Documentation, mailing lists and feedback**

364 All WU-Minn HCP data available in ConnectomeDB are extensively
 365 documented. Reference manuals are compiled for each major HCP
 366 data release and are available at [http://humanconnectome.org/
 documentation](http://humanconnectome.org/documentation). These manuals contain detailed information specific
 367 to each release about accessing the data, hardware and protocols, SOPs,
 368 directory structure and file information for downloaded data. They also
 369 contain detailed information about data collection procedures, task
 370 procedures, pipelines and more. In addition, an HCP wiki ([https://wiki.
 humanconnectome.org/display/PublicData/Home](https://wiki.humanconnectome.org/display/PublicData/Home)) is maintained
 371 containing additional documentation and updates. Most notably, this
 372 site contains a “Known Issues and Planned Fixes” page that details
 373 recently discovered issues, including information regarding fixes and/
 374 or estimated timelines for fixes between data releases. For backwards
 375 compatibility of data, the HCP preserves an archive of previous data
 376 releases (including data with known issues) that are accessible upon
 377 request. (However, the user interface only provides access to the current
 378 version of the data). Similarly, previous versions of HCP software are
 379 available via GitHub
 380

381 In addition to documentation, the HCP has established mechanisms
 382 for announcements and user feedback. The WU-Minn HCP consortium
 383 maintains a website portal (<http://www.humanconnectome.org>)
 384 where users and the public can find announcements, links to documenta-
 385 tion, and general information about the project. Through the website and/
 386 or upon registration to ConnectomeDB, users are invited to subscribe to
 387 an announcement list and e-mail user forum. The announcement list is
 388 used to make announcements about data releases, software updates,
 389 events and issues. The forum is an open discussion group, actively
 390 monitored by project investigators and staff, where users can submit
 391 issues, requests and bug reports.
 392

393 In addition to these groups, HCP administration has set up support
 394 and feedback e-mail addresses that are answered by HCP staff. Links

to these are made available in “Contact” sections of ConnectomeDB and the HCP website. In addition, the software platform or scripts developed by HCP (ConnectomeDB, Connectome Workbench, FieldTrip, and the HCP Processing Pipelines), have user-friendly feedback mechanisms (bug report forms, user support email, wiki forums), built in for users to submit bug reports and feature requests. Submitted requests are transferred automatically or manually entered into issue tracking software to be managed by project staff and developers. This approach allows for user feedback from all sources of HCP information, data, and software to be organized and managed in one system.

Future directions

The neuroimaging community has expressed high interest in the HCP effort, and many investigators have started using HCP data in their work. As of April 2015, over a million package files totaling nearly 1.5 petabytes of data have been served to over 1900 users in 57 countries. Additionally, 168 users have requested data via Connectome-In-A-Box, and 431 drives have been shipped. Nearly 350 researchers have requested and been approved for access to restricted data. Over 50 publications using released HCP data and methods, including over 30 by investigators completely independent from the WU-Minn consortium, have been published since the initial HCP Q1 Data Release in March 2013.

Much has been learned during the course of the WU-Minn consortium HCP project and much remains to be done. The consortium will continue to release data as new subject data is acquired and processed. Further analysis outputs will be made available as they are made ready for release, including the release of diffusion tractography analyses and parcellated task fMRI data. Additional work on the ConnectomeDB application and platform is also planned, including the expansion of data mining tools, in order to provide a better within-DB platform for data exploration and analysis.

ConnectomeDB will become the foundation for the NIH-supported Connectome Coordination Facility (CCF) and the primary dissemination platform for NIH-funded HCP-style data acquisition and analysis supported by the Connectomes of Human Diseases (<https://grants.nih.gov/grants/guide/pa-files/PAR-14-281.html>) and Lifespan-HCP (<http://grants.nih.gov/grants/guide/rfa-files/RFA-AG-16-004.html>) funding mechanisms. An important function of the CCF will be to facilitate comparison and aggregation across datasets. Consequently, ConnectomeDB is being developed to provide an improved unified-yet-customized interface to support data from many sites and studies and to handle the corresponding increase in traffic these studies are expected to generate.

References

- Barch, D.M., Burgess, G.C., Harms, M.P., Petersen, S.E., Schlaggar, B.L., Corbetta, M., Glasser, M.F., Curtiss, S., Dixit, S., Feldt, C., Nolan, D., Bryant, E., Hartley, T., Footer, O., Bjork, J.M., Poldrack, R., Smith, S., Johansen-Berg, H., Snyder, A.Z., Van Essen, D.C., for the WU-Minn HCP Consortium, 2013. Function in the human connectome: task-fMRI and individual differences in behavior. *Neuroimage* 80, 169–189. 437
- Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., et al., 2013. The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* 80, 105–124. 441
- Griffanti, L., Salimi-Khorshidi, G., Beckmann, C.F., Auerbach, E.J., Douaud, G., Sexton, C.E., Zsoldos, E., Ebmeier, K.P., Filippini, N., Mackay, C.E., Moeller, S., Xu, J., Yacoub, E., Baselli, G., Ugurbil, K., Miller, K.L., Smith, S.M., 2014. ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. *Neuroimage* 95, 232–247. 444
- Gur, R.C., Ragland, J.D., Moberg, P.J., Turner, T.H., Bilker, W.B., Kohler, C., Siegel, S.J., Gur, R.E., 2001. Computerized neurocognitive scanning: I. Methodology and validation in healthy people. *Neuropsychopharmacology* 25, 766–776. 445
- Gur, R.C., Richard, J., Hughett, P., Calkins, M.E., Macy, L., Bilker, W.B., Bressinger, C., Gur, R.E., 2010. A cognitive neuroscience-based computerized battery for efficient measurement of individual differences: standardization and initial construct validation. *J. Neurosci. Methods* 187, 254–262. 452
- Herrick, R., McKay, M., Olsen, T., Horton, W., Florida, M., Moore, C.J., Marcus, D.S., 2014. Data dictionary services in XNAT and the Human Connectome Project. *Front. Neuroinform.* 8, 65. 456
- Larson-Prior, L.J., Oostenveld, R., Penna, S.D., Michalareas, G., Prior, F., Babajani-Feremi, A., Schoffelen, J.M., Marzetti, L., Pasquale, F., Pompeo, F.D., Stout, J., Woolrich, M., Luo, Q., Bucholz, R., Fries, P., Pizzella, V., Romani, G.L., Corbetta, M., Snyder, A.Z., 2013. Adding dynamics to the human connectome project with MEG. *Neuroimage* 80, 190–201. 457
- Marcus, D.S., Olsen, T.R., Ramaratnam, M., Buckner, R.L., 2007. The Extensible Neuroimaging Archive Toolkit: an informatics platform for managing, exploring and sharing neuroimaging data. *Neuroinformatics* 5, 11–34. 462
- Marcus, D.S., Harms, M.P., Snyder, A.Z., Jenkinson, M., Wilson, J.A., Glasser, M.F., Barch, D.M., Archie, K.A., Burgess, G.C., Ramaratnam, M., et al., 2013. Human connectome project informatics: quality control, database services, and data visualization. *Neuroimage* 80, 202–219. 463
- Milchenko, M., Marcus, D.S., 2013. Obscuring surface anatomy in volumetric imaging data. *Neuroinformatics* 11, 65–75. 464
- Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.M., FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data, *Comput. Intell. Neurosci.* vol. 2011, Article ID 156869, 9 pages. 465
- Salimi-Khorshidi, G., Douaud, G., Beckmann, C.F., Glasser, M.F., Griffanti, L., Smith, S.M., 2014. Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage* 90, 449–468. 466
- Setsonpop, K., Kimmlingen, R., Eberlein, E., Witzel, T., Cohen-Adad, J., McNab, J.A., et al., 2013. Pushing the limits of in vivo diffusion MRI for the Human Connectome Project. *Neuroimage* 80, 220–233. 467
- Toga, A., Clark, K., Thompson, P., Shattuck, D., Van Horn, J., 2012. Mapping the human connectome. *Neurosurgery* 71 (1), 1–5. 468
- Ugurbil, K., Xu, J., Auerbach, E.J., Moeller, S., Vu, A., Duarte-Carvajalino, J.M., Lenglet, C., Wu, X., Schmitter, S., Van de Moortele, P.F., Strupp, J., Sapiro, G., De Martino, F., Wang, D., Harel, N., Garwood, M., Chen, L., Feinberg, D.A., Smith, S.M., Miller, K.L., Sotiropoulos, S.N., Jbabdi, S., Andersson, J.L., Behrens, T.E., Glasser, M.F., Van Essen, D., Yacoub, E., 2013. Pushing spatial and temporal resolution for functional and diffusion MRI in the Human Connectome Project. *Neuroimage* 80, 80–104. 482