**Data Transfer, QA, and Pre-processing**

In research, the single most important concern is data integrity. We can all help ensure data integrity by ensuring uniform procedures in data collection, prompt transferring of data, storing data uniformly, and promptly pre-processing and ensuring the quality of data to figure out any potential sources of error. Everyone works together in the lab to ensure that our data is of the highest quality.

This is a modular guide on data transfer, QA, and pre-processing steps, arranged by modality or scanning session type. At each step, this guide indicates who should be performing each step, and when.

**Contents:**

1. **Directory Structure Basics**
2. **Subject Numbers**
3. **MRI Scan Sessions**
   1. **Imaging Data**
   2. **Access Database Recording**
   3. **Physiological Data**
   4. **fMRI Task Data**
   5. **Simultaneous EEG during fMRI**
4. **MEG Scan Sessions**
   1. **MEG Data**
   2. **Access Database Recording**
   3. **MEG Task Data**
   4. **Headshape Data**
      * 1. **Directory Structure Basics**

On the NIH HPC systems (helix, felix, and biowulf), each user has their own home directory (throughout this document “username” is your username.

/home/username

Our group has a very large directory for data and analyses

/data/MoodGroup

Each of our studies, or protocols, have a directories for raw data and analysis, named by protocol number.

/data/MoodGroup/18M0001/

/data/MoodGroup/18M0001\_fmri\_analysis/

/data/MoodGroup/18M0001\_MEG\_analysis

Raw data should be placed in individual subject directories:

/data/MoodGroup/18M0001/STUDYID101\_sub-9999

Each study is given a short code identifying it, such as MOA for “mechanism of action” or RD for “repeat dose.”

Each subject in the study has an ID for that study: usually a number like 101, 102, etc.

Finally, each subject has an ID number that is unique to that individual, but does not vary by study. This is to be able to easily see if de-identified data comes from the same subjects or not.

Under each study/subject, data is divided by modality/imaging session:

/data/MoodGroup/18M0001/subj\_NEWSTUDY001/mri

/data/MoodGroup/18M0001/subj\_NEWSTUDY001/meg

PLEASE FOLLOW THESE FORMATS **EXACTLY**. IF THE FORMATS ARE INCONSISTENT IT IS **EXTREMELY** DIFFICULT TO EFFICIENTLY PROCESS RAW DATA WITH SCRIPTS.

* + - 1. **Subject Numbers**

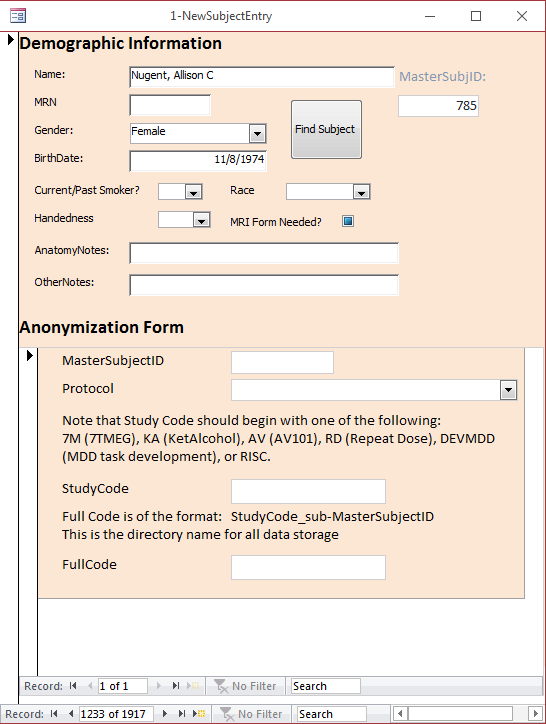
**As soon as a subject is enrolled in a study and scheduled for imaging procedures**, they should be entered into the Access Imaging Subject Database, along with basic demographic information.

Open up the SubjectData database on [\\nimhirp-ds1.nih.gov\snmdzarate$\Database](file:///\\nimhirp-ds1.nih.gov\snmdzarate$\Database)

You’ll be prompted for the password: Nellie4ever

On the left hand side, locate “All Access Objects”. There are tables listed first (these are like spreadsheets that are all linked together by common information). Next are queries, which are reports generated from data in one or more tables. Finally we have forms, which are more user friendly for entering data. Open the first form by double clicking on it – 1-NewSubjectEntry.

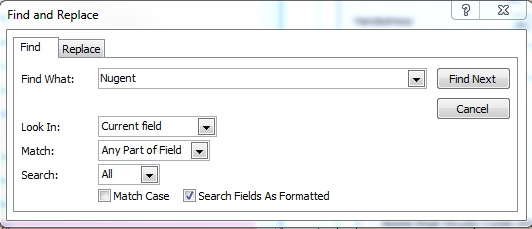
You’ll see this:



1. Figure out if your subject has been studied before:

Press the “Find Subject” button

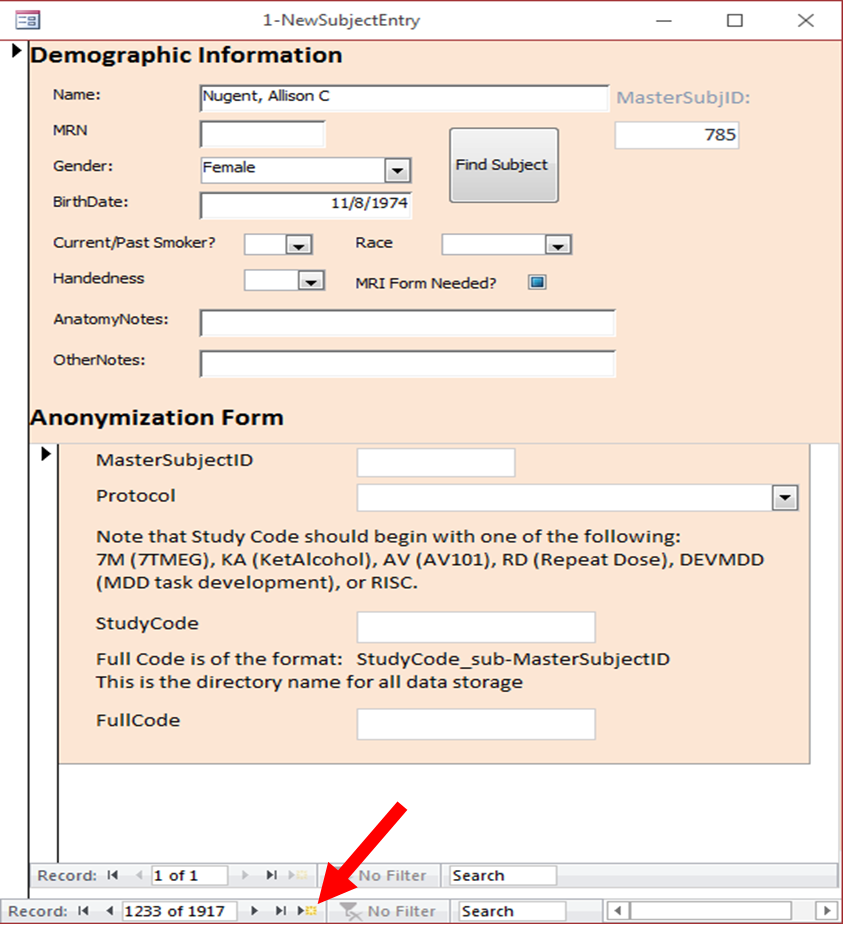
Under “Match” change “Whole Field” to “Any Part of Field”



Type in a portion of the name you are looking for, and press “Find Next”

If you typed in a relatively common name, (like Smith), you can keep pressing “Find Next” to cycle through all the “Smith”s in the database.

1. If the subject isn’t in the database, make a new entry. At the very bottom of the form press the yellow start to get a clean record under “Demographic Information.” Fill out all the information on the subject.



1. Fill out the anonymization form

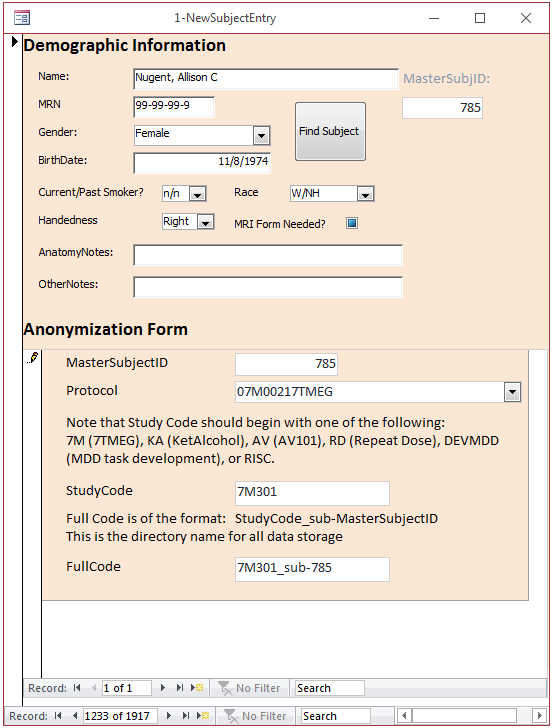
Enter the master Subject ID from the top portion of the form (you may get a read only error, ignore it and go ahead and enter the data).

Enter the protocol you are enrolling the subject in

Enter the StudyCode for the subject, which includes their study ID number

Finally, enter the full code, which will be the directory name under which all raw data will be stored.

Here is a correctly filled out form:



**3. MRI DATA:**

**a. Imaging Data:**

**Imaging Data Should be downloaded within one business day of the scan by the IRTA attending the scan session**

Step 1: Make a directory for the subject’s data:

The example here shows a generic “username” where your username will appear. STUDYCODE will be replaced with the code for your study, 999 will be replaced with the study ID number for this subject, and SUBJNUM will be replaced by the unique subject ID number. We use a fictitious protocol here.

[username@helix tmp\_tar]$cd /data/MoodGroup/18M0001

[username @helix 18M0001]$ mkdir STUDYCODE999\_sub-SUBJNUM

[username @helix 18M0001]$ mkdir STUDYCODE999\_sub-SUBJNUM/mri

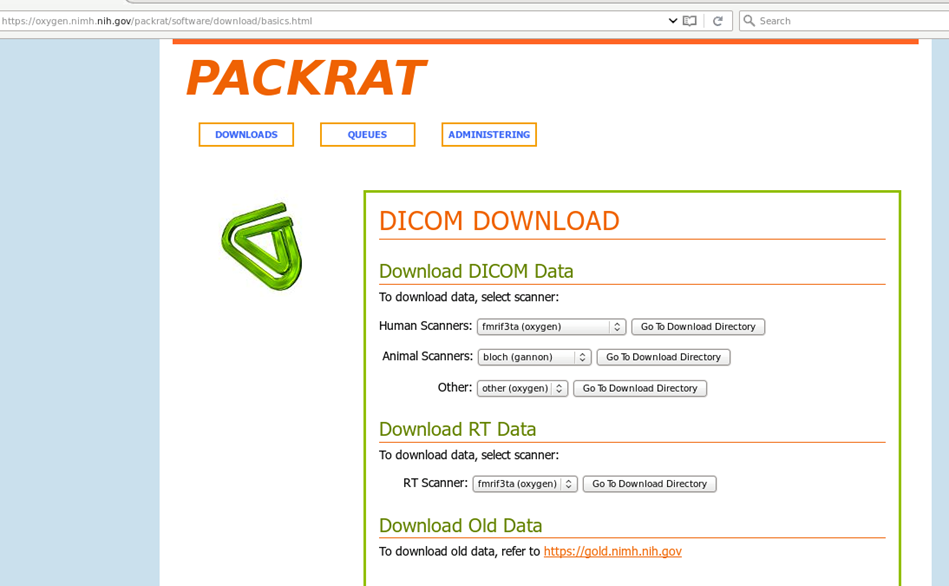
Finally, you’ll make a folder for the specific date of the scan – this directory name will also indicate if it was a 3T or 7T scan:

[username @helix 18M0001]$ mkdir STUDYCODE999\_sub-SUBJNUM/mri/20180101\_7T

Step 2: Download the MRI raw data:

Data from the scanner are pushed to a dicom server called Packrat:

<http://oxygen.nimh.nih.gov>



You can navigate the website by scanner, date, and subject identification.

Right click on the .tgz file, and select “Save link as…”

You can then navigate to the temporary tar file directory where we keep any identifiable information:

/data/MoodGroup/18M0001/tmp\_tar/

Save the data there.

Navigate to that directory

[username@helix ~]$ cd /data/MoodGroup/18M0001/tmp\_tar/

Do a directory listing to make sure it worked

[username@helix tmp\_tar]$ ls

Step 3: Unpack the data

Unzip and untar the raw data file:

[username@helix tmp\_tar]$ tar -xzvf LASTNAME\_FIRSTNAME-1111111-20180101-1111111-DICOM.tgz

[username@helix tmp\_tar]$ cd LASTNAME\_FIRSTNAME -1111111/20180101-1111111/

[username@helix tmp\_tar]$ ls

You should see folders for each separate image acquired in the session, as well as a file called README-Study.txt

[username@helix 20180101-1111111]$ cat README-Study.txt

This will print the contents of that file to the screen.

Create the nifti files, using the command README2nii2.py, which is a python script written by Jen Evans. Usage is as follows:

python README2nii2.py --file README\_FILE --outdir OUTPUT\_DIR --pref OUTPUT\_PREFIX

[username@helix 20180101-1111111]$ python README-Study.txt /data/MoodGroup/18M0001/ STUDYCODE999\_sub-SUBJNUM/mri/20180101\_7T/nii --pref subj002

The output prefix should be the subject number for the study

If the python command doesn’t run, you may have to load the python module:

[username@helix ]$ module load python

Look at the directory to make sure things ran properly:

[username@helix 20180101-1111111]$ cd /data/MoodGroup/18M0001/ STUDYCODE999\_sub-SUBJNUM/mri/20180101\_7T

[username@helix 20180101\_7T]$ ls

You should see a list of all the images acquired in the session

Step 4: Copy over and redact the README file

You’ll want to copy over the README-Study.txt file, but we’ll need to remove the identifying information from it.

[username@helix 20180101\_7T]$cp /data/MoodGroup/18M0001/tmp\_tar/LASTNAME\_FIRSTNAME-1111111/20180101-1111111/README-Study.txt .

That’s a space then a period at the end. The period basically tells the operating system to copy the file “here” or the current directory.

Since the README file has names, we’ll need to edit it.

[username@helix 20180101\_7T]$ cut -d ‘,’ -f ‘2-6’ README-Study.txt

You can use cat to print the file and make sure that worked.

[username@helix 20180101\_7T]$ cat README-Study.txt

Finally, you’ll need to remove the directory of raw dicom data.

[nugenta@helix 20180101\_7T]$ cd /data/MoodGroup/18M0001/tmp\_tar/

[nugenta@helix tmp\_tar]$ rm -r LASTNAME\_FIRSTNAME-1111111

**b. Access Data Recording**

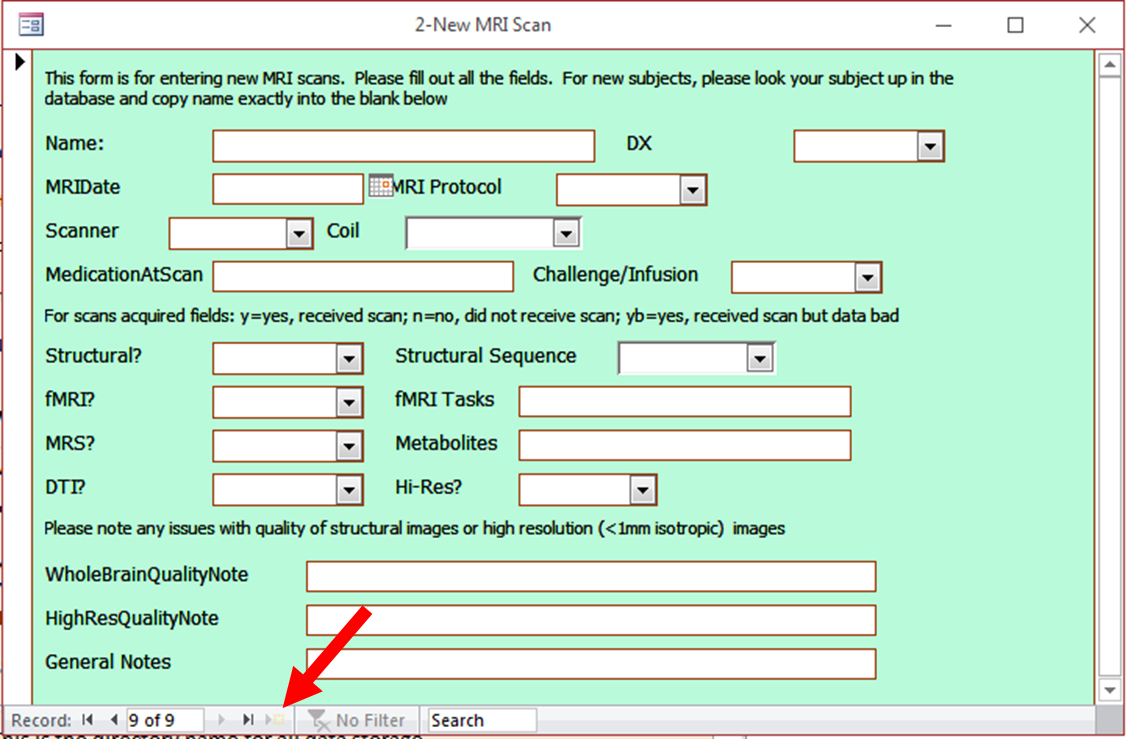
**Physio data should be downloaded within one business day of the scan by the IRTA attending the scan session**

Open the 1. New Subject Entry form. Locate your subject using the “Find Subject” button.

Highlight the Name of the subject, so it is spelled and formatted exactly the same, and right click or press Ctrl-C to copy.

Open the 2. New MRI Scan form

On the New MRI Scan form press the yellow star to enter a new record:



Right click in the name field or press Ctrl-V to paste the name you copied from the subject form.

Medication at Scan refers to any psychiatric meds the subject is taking daily at the time of the scan, this will be “None” for most baseline studies. This does not include any recent infusion.

Please enter all the information, using the drop down boxes when possible.

**c. Physio Data**

**The Imaging record should be entered in the database within one business day of the scan by the IRTA attending the scan session**

Step 1: Data Transfer

After scanning, you should have put the physio data in the /data/MoodGroup/18M001/tmp\_physio directory. Now you can move it to the proper place.

[nugenta@helix]$ cd /data/MoodGroup/18M0001

[nugenta@helix 18M0001]$ cp tmp\_physio/ACN001/\* subj-NEWSTUDY001/mri/20180101\_7T/physio/.

[nugenta@helix 18M0001]$ cd subj-NEWSTUDY001/mri/20180101\_7T

Step 2: Data Preprocessing

Next, you’ll need to convert the physio data from the Biopac format to the afni .1D format, then to physiological regressor files. This will use the python script phys2slibase.py

Usage is as follows:

python phys2slibase.py --fmri\_fn IMAGE\_FILE.nii --phys\_fn PHYSIO\_FILE.acq

Depending what directory you are in, you will want to make sure that you use the correct path for each file. Using tab-complete is a good way to make sure the operating system is finding the file as you type in the command.

[nugenta@helix 20180101\_7T]$ python phys2slibase.py

--fMRI\_fn nii/subj002676b\_sin\_epi\_2iso\_ee\_run1.nii --phys\_fn physio/20180101\_ACN11110001.acq

[nugenta@helix 20180101\_7T]$ ls physio

This should show you several new .1D files, including ones labeled slibase, card, resp, and trig.

[Additional QA Information – under construction]

**d. fMRI Task Data**

[Retrieval of fMRI Task data – under construction]

* 1. **Simultaneous EEG data**

[Retrieval of EEG data – under construction]

**3. MEG**

1. **Imaging Data**

**Data should be downloaded within 1 business day of the scan by the IRTA attending the scanning sessions.**

Step 1: Make a directory for the subject’s data (there may already be a directory for the subject):

[username@helix tmp\_tar]$cd /data/MoodGroup/18M0001

STUDYCODE will be the unique code for your study/protocol, 999 will be the subjects study code, and SUBJNUM will be the unique subject number.

[username@helix 18M0001]$ mkdir STUDYCODE999\_sub-SUBJNUM

[username@helix 18M0001]$ mkdir STUDYCODE999\_sub-SUBJNUM/meg

[username@helix 18M0001]$ mkdir STUDYCODE999\_sub-SUBJNUM/meg/20180101

Step 2: Download the data:

[username@helix 18M0001]$ cd STUDYCODE999\_sub-SUBJNUM/meg/20180101

[username@helix 18M0001]$ ssh -X [mfurey@tako.nimh.nih.gov](mailto:mfurey@tako.nimh.nih.gov) (for RISC)

[username@helix 18M0001]$ ssh -X [gilbert@tako.nimh.nih.gov](mailto:gilbert@tako.nimh.nih.gov) (for Repeat Dose)

Find the data by navigating to the correct date

[mfurey@tako ]$ cd data/20180101/

[mfurey@tako 20180101]$ ls

There should be several datasets, ending in .ds, one for each recording. Copy them to felix.

[mfurey@tako 20180101]$ scp -r \*.ds /data/MoodGroup/18M0001/subj\_NEWSTUDY002/meg/20180101

[mfurey@tako 20180101]$ exit

Make sure the datasets arrived safely

[username@helix 18M0001]$ cd subj\_NEWSTUDY002/meg/20180101

[username@helix 20180101]$ ls

**Data should be QA’d within one week of data collection by the IRTA who attended the imaging session**

Step 3: QA the data:

All QA will require the use of the CTF module of software tools, which you’ll have to load:

[username@helix 20180101]$ module load ctf

(you only have to do that once per terminal)

Also, you will need to have a copy of the runsheet with you as you run these commands.

**Preprocess and QA: Airpuff Data (Repeat Dose)**

Navigate to the airpuff analysis directory

[username@helix]$ cd /data/MoodGroup/17M0060\_MEG\_analysis/airpuff

Now, you’ll need to apply some basic filtering, using the newDs command to create a new, filtered dataset. The syntax is:

newDs -filter filterfilename.cfg input\_dataset.ds output\_dataset.ds

We’ll also want to distinguish between the baseline airpuff recording and the infusion airpuff recording, Use the runsheet to verify this.

[username@helix airpuff]$ newDs -filter processing\_airpuff.cfg /data/MoodGroup/17M0060/sub-1111\_RD401/meg/20171108/ABC\_airpuff\_20171108\_01.ds 401\_20171108\_BL\_airpuff-f.ds

[username@helix airpuff]$ newDs -filter processing\_airpuff.cfg /data/MoodGroup/17M0060/sub-1111\_RD401/meg/20171108/ABC\_airpuff\_20171108\_02.ds 401\_20171108\_INF\_airpuff-f.ds

Next, you need to run a script to place the markers. (Note that there is no “.df” at the end)

[username@helix airpuff]$ marker\_script 401\_20171108\_BL\_airpuff-f

[username@helix airpuff]$ marker\_script 401\_20171108\_INF\_airpuff-f

Now you are ready to QA! Open the dataset using DataEditor

[username@helix airpuff]$ DataEditor -data 401\_BL\_airpuff-f.ds

Set up the CTF viewer window so that you can visually inspect a quadrant of sensors at a time, or a hemisphere at a time. You are looking for bad channels and artifacts.

If you want black and white lines instead of color:

Display > Show > make sure that Color Traces is deselected

To inspect a hemisphere of channels at once:

Click on the Sensors icon [] and select Base > MEG Left; also scroll down on the Unselected side until you see UADC001; highlight that and move it to the Selected column using the right arrow. This will allow you to see all the left-hand sensors and the airpuff trigger. It’s also a good idea to include a couple of sensors from the opposite hemisphere, right by the eyes, to aid in the identification of eye blinks – these deflections will be positive in one hemisphere and negative in the other.

To inspect a quadrant of channels, or a selection of channels:

After clicking the Sensors icon [] press the “Sensors” button, and you’ll see a map of the sensors. You can outline with your mouse the sensors you want to see, or click on sensors individually.

At the bottom left-hand side of the viewer window, make sure your Time Scale is either 0.125 sec/cm or 80 mm/sec

Below the Time Scale, make sure that npuff (e) is selected as your Marker

Make sure Edit Markers is selected on the upper right-hand side of the CTF viewer window

Set the Gain to the right of the Time Scale so that the MEG-SENS gain is 2p or 5p, depending on how many channels you are looking at.

Now do the QA!

You should now be able to scroll through the data and remove trials that occur during an eyeblink or jaw clenching by selecting it in the viewer window, then dragging and dropping it below the viewer window. Remember that we are interested in the time period 100ms before and 300ms after each stim mark, so make sure to remove marks near artifacts as well.

Once you’ve scrolled through the entire run using the left sensors, go back to the beginning of the run and then select the MEG Right sensors and scroll through once more to ensure that all artefacts are removed from the run

If you encounter a noisy/bad channel, you should highlight it by clicking on the sensor name in the left-hand side of the viewer window. Once the sensor is highlighted in red, right-click and select Set Good/Bad. This will mark the sensor as bad and remove it from subsequent analyses.

Once you’ve reviewed both hemispheres and removed artefacts, you’ll save the data by selecting File > Save Dataset (Edit Info)

Select Edit > Marker Sets and record the number of remaining npuff trials. Also note if you removed any channels based on noise.

**Preprocess and QA: Hariri Hammer Data**

Navigate to the haririhammer analysis directory

[username@helix]$ cd /data/MoodGroup/17M0060\_MEG\_analysis/haririhammer

Similar to the airpuff data, we’ll need to filter the datasets, and rename (according to the runsheet) baseline and infusion runs.

[username@helix airpuff]$ newDs -filter processing\_hh.cfg /data/MoodGroup/17M0060/sub-1111\_RD401/meg/20171108/ABC\_haririhammer\_20171108\_01.ds 401\_20171108\_BL\_hh-f.ds

[username@helix airpuff]$ newDs -filter processing\_hh.cfg /data/MoodGroup/17M0060/sub-1111\_RD401/meg/20171108/ABC\_airpuff\_20171108\_02.ds 401\_20171108\_INF\_airpuff-f.ds

Now, you need to place the markers.

[Under Construction - - script to place and name markers in progress]

Finally, you will do the QA.

The QA process is the same for airpuff data, flagging bad channels and removing marks that occur during artifacts. This time, however, each marker will have a different name depending upon the stimulus type. You’ll need to remove all the marks that occurring during an artifact, which may require you to switch back and forth between the markers.

**Preprocess and QA: Rest Data**

Navigate to the rest analysis directory (change the protocol number as necessary)

[username@helix]$ cd /data/MoodGroup/17M0060\_MEG\_analysis/rest

Similar to the airpuff data, we’ll need to filter the datasets, and rename (according to the runsheet) the runs. Notice that the 20 minute resting scan is the INFBEG scan, because the infusion begins in the middle of the run.

[username@helix rest]$ newDs -filter processing\_rest.cfg /data/MoodGroup/17M0060/sub-1111\_RD401/meg/20171108/ABCDEFGH\_rest10\_20171108\_01.ds 401\_20171108\_BL\_rest10-f.ds

[username@helix rest]$ newDs -filter processing\_rest.cfg /data/MoodGroup/17M0060/sub-1111\_RD401/meg/20171108/ABCDEFGH\_rest20\_20171108\_01.ds 401\_20171108\_INFBEG\_rest20-f.ds

[username@helix rest]$ newDs -filter processing\_rest.cfg /data/MoodGroup/17M0060/sub-1111\_RD401/meg/20171108/ABCDEFGH\_rest10\_20171108\_02.ds 401\_20171108\_INFEND\_rest10-f.ds

[username@helix rest]$ newDs -filter processing\_rest.cfg /data/MoodGroup/17M0060/sub-1111\_RD401/meg/20171108/ABCDEFGH\_rest10\_20171108\_03.ds 401\_20171108\_POSTINF\_rest10-f.ds

Now you are ready to QA! Open the dataset using DataEditor

[username@helix rest]$ DataEditor -data 401\_rest10\_20171108\_01-f.ds

Set up the CTF viewer window so that you can visually inspect a quadrant or a hemisphere of sensors for artefacts. For the resting state scans, we are not worried about eye movements/blinks. We will only mark muscle artifacts, and any other large spikes. You’ll also be looking for bad channels.

If you want black and white lines instead of color:

Display > Show > make sure that Color Traces is deselected

Click on the Sensors icon [] and select Base > MEG Left, or press the “sensors” button to select a quadrant or selection of sensors. Also scroll down on the Unselected side until you see the EEG channels; highlight that and move it to the Selected column using the right arrow. This will allow you to see the EOG eyeblink sensor.

At the bottom left-hand side of the viewer window, make sure your Time Scale is either 0.125 sec/cm or 80 mm/sec

Set the Gain to the right of the Time Scale so that the MEG-SENS gain is 2p or 5p depending on how many sensors you are viewing at once.

[insert more language here on QA]

Now do the QA for bad channels.

If you encounter a noisy/bad channel, you should highlight it by clicking on the sensor name in the left-hand side of the viewer window. Once the sensor is highlighted in red, right-click and select Set Good/Bad. This will mark the sensor as bad and remove it from subsequent analyses.

Once you’ve reviewed both hemispheres and highlighted any bad channels, you’ll save the data by selecting File > Save Dataset (Edit Info)

Now you are ready to run the preprocessing python script.

[username@helix rest]$ python RD\_MEGrest\_PreProcess.py 401\_rest10\_20170705\_01-f.ds 401\_rest10\_20170705\_01

This will take some time, as it is performing an ICA on the data.

QA the ICA:

\*\*Under Construction: To be continued\*\*

**b. Access Data Recording**

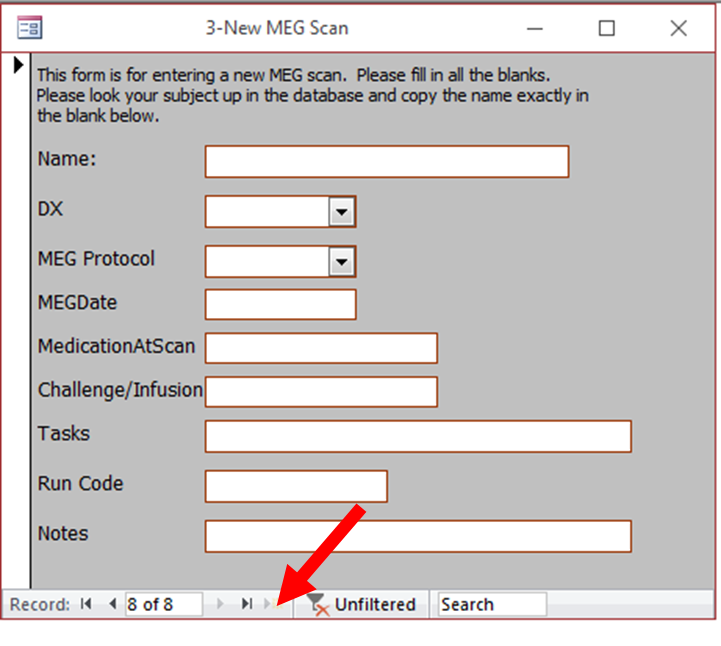
**The Imaging record should be entered in the database within one business day of the scan by the IRTA attending the scan session**

Open the 1. New Subject Entry form. Locate your subject using the “Find Subject” button.

Highlight the Name of the subject, so it is spelled and formatted exactly the same, and right click or press Ctrl-C to copy.

Open the 2. New MEG Scan form

On the New MEG Scan form press the yellow star to enter a new record:



Right click in the name field or press Ctrl-V to paste the name you copied from the subject form.

Medication at Scan refers to any psychiatric meds the subject is taking daily at the time of the scan, this will be “None” for most baseline studies. This does not include any recent infusion.

Please enter all the information.

Run Code refers to the eight letter anonymization code generated by the MEG acquisition software.

1. **MEG Task Data**

[under construction]

1. **Headshape Data**

[under construction]