Guide to article distillation for the White Matter Anatomy Database (WMAD) project

# Overview

In this document you will find an overview of the process for contributing to the WMAD project as it relates to the curation and distillation of journal articles. This work is part of an effort to help digitally synthesize the disparate accounts of white matter anatomy currently presented in our field

# Background

# Overview of Article Extraction

For our purposes, the process of extracting the relevant information from articles entails two general processes:

1. Cataloging the relevant information about the article itself.
2. Cataloging the white matter structures they discuss, and those constituent discussions.

## Directory and file structure

Within the overarching directory for completed articles (**completedArticles** as of 8/18/2021) you should find some number of directories, each of which corresponds to an article which has been subjected to the extraction process.

As you begin the extraction process create a new directory with a name of your choosing., The name for this directory should:

1. Be useful to curators such as yourself
2. Delineate between articles (e.g. in cases where the first author had two papers published in the same year).

### Directory structure for an individual article

Within this directory there are **two** general Excel files:

1. A file with data about the article itself, entitled **doc.xlsx**
2. Some number of excel files, entitled **tract1.xlsx, tract2.xlsx, tract3.xlsx**, etc. for each white matter structure discussed.

Templates for both of these files can be found within the [**ExcelFiles**](https://drive.google.com/drive/folders/16PBwj-FeOBcCIvaFqkKElihmSBstXKV-?usp=sharing)directory.

#### The document excel file (“doc.xlsx”)

The first of these two file types, the “doc” excel file, contains information about the document itself.

NOTE: the template file contains the following fields:

| curator |
| --- |
| title |
| journal |
| dateIssued |
| authors |
| doi |
| species |
| methods |

While you are welcome to record information for the **title**, **journal**, **dateIssued**, and **authors** fields, this is not necessary (though you may wish to do so to make the sheet more readable and informative). Why is this? When we convert these excel records to the format used in our proto-database ([JSON](https://www.json.org/json-en.html)), we are able to use the doi to query those records in a standardized format, which automates that part of the recordkeeping process.

As such we really only need to record:

1. The curator’s name
2. The doi
3. The pertinent species
4. The pertinent methods

In the case of the **species** and **methods** fields, curators should enter one entry (e.g. “human”, “macaque”, etc. for species and “tractography”, “dissection”, etc.) per column (starting with column b).

#### The tract excel file (“tractX.xlsx”)

Along with the **doc.xlsx** file, an article extraction directory will also include some number of excel files (entitled **tract1.xlsx, tract2.xlsx, tract3.xlsx**, etc.) each of which corresponds to a distinct white matter structure discussed in the article.

Overall, these tract records are intended to be structured around the “list” of white matter structures discussed in the article. Conceptually, for each white matter structure, we are developing a “filter” of sorts, that removes all text and figures NOT related to a tract of interest. As such, the extraction of records should be approached from the perspective of creating a list of the white matter structures discussed/depicted in the article, and then cataloguing those descriptions and depictions. The *actual* process you engage in to perform an article extraction need not reflect this in a literal sense--you don’t need to start by creating a master list for each article.

Regardless, the creation of excel spreadsheets for tracts (i.e. white matter structures) is somewhat more laborious than the doc.xlsx file. It entails (for each tract) filling out the following 3 fields.

1. **tractTermsUsed**: In each cell, one of the terms used to refer to the white matter structure in question, either the full name, abbreviation, or alternate names provided. Given that *some* structure is being referred to by the descriptions or depictions (either implicitly or explicitly), it is presumed this is a *necessary* field.
2. **descriptions**: using the new Google Chrome “link to highlighted text” capability (described [here](https://support.google.com/chrome/answer/10256233)) create a link to each individual section that discusses the tract in question. In the Chrome browser, highlight the relevant text, and then right click the highlight and select the “**copy link to highlight**” option. NOTE: To some degree, there is an element of subjectivity as to what is pertinent. Remember our main interest here are descriptions of anatomy, and that to some extent, more granular selections are preferred to selection with a lot of “chaff” (i.e. sections discussing other tracts or unrelated matters). That being said, when in doubt, lean towards overinclusion as we can perform post-extraction curation if necessary.
3. **figures**: links to *the images themselves* for any depictions of the tract. Typically these should be a web link that has a .jpg, .png or some other image file type at the end.

For an example of this, done on Bullock et al. 2019, check the [CompleteExample](https://drive.google.com/drive/folders/1kseHVV3ZGXu6CkrwoWUbraZGeV8q3CMf?usp=sharing) or look at the [completedArticles](https://drive.google.com/drive/folders/10Z_O6ybJ6W5A12WhZbcWPsszLoq52JyY?usp=sharing) directory.

NOTE: the .json file that you may or may not find in completed article directories is the product of a code pipeline that extracts information from the excel files. Curators are not expected to produce these.