# Preparing data

The following are steps I took in Excel to prepare the data after exporting it into csv (starting with sherpa\_all\_policies.csv). NB. This data was missing about 95 journals from the full Sherpa download (ones that were skipped because they were breaking the loop).

Removed all rows which do not have a corresponding ISSN and/ or journal title in the Dimensions file using a series of join\_left commands in R. See R file “dim issns into sherpa”

* This cut the Sherpa file down to 31417 rows/ 6159 journals (from 117,294 rows/ 30,874 journals).
* Since our Dimensions data has 6382 unique journal names this suggests we have Sherpa data about 97% of the journals.
* Neither journal title nor issn are sufficient to merge on their own – a fairly substantial number either match on title but not issn or issn but not title (possibly because I tested on only the first issn number stored in dimensions)

Edited title column to change name from “ï..title” to “title”

Renamed NA to 0 in embargo.units (I checked with JISC and if embargo = NA this means no embargo/ 0 months).

Converted all embargos recorded in days or years into months then deleted embargo.units

Deleted columns issn\_NA and issn\_legacy as both were all blank

Created a new column called live\_num\_pathways by using countif to count the number of policies for each journal id. Copied the folder (values only) to another columns called orig\_num\_pathways

**Set up filters**

Set up a series of filters to help choose policy options:

* One for each license and one for license missing
* Repository options
  + Rep any = either any\_repository or any\_website

**Started to separate out journals with only one pathway**

First I filtered orig\_num\_pathways = 1 and copied out all the rows into a new sheet (1 path) then deleted them from the main sheet

* 364 journals had only one pathway
* 5794 journals had multiple pathways (31053 rows)

I deleted all rows from multiple pathways file which refer to submitted articles only. Before doing this, I ran a check to see if there were any journals which had more than one pathway but all for submitted only and found one (id: 24519), which I removed and added to the one pathway file as one row.

I then generated sorted by live\_num\_pathways and moved all the policies which now had only one pathway into the one pathway file. This left us with 24746 rows covering 5523 journals.

I then moved all rows which mentioned additional\_oa\_fee=yes to another worksheet. I checked which of the id’s had duplicates in the no fee worksheet and for the ones with no duplicate I cut and pasted them into the one pathway file (3 in total). I then deleted the remainder and checked for and removed journals which were down to one pathway. This left the full sheet with 12755 rows covering 4590 journals in the multiple pathways file.

I then moved all rows with open\_access\_prohibited = no to the one pathway, selecting the first of each pair (each had 2 policies/ publishers).

I removed duplicates where the policies are identical other than ‘name’ – i.e. they have the exact same policy attributed twice to two different publishers. This left us with 10050 rows and 4587 journals

**Final filtering**

I then ranked them all by ID and began a stage by stage filtering starting from ideal and working downwards. Each time I removed a set I looked for their corresponding ID’s in the remaining data and deleted them by finding ones with only one remaining pathway!

1. Authors, cc-by, 0 embargo, any repository, published

COULD I DO ONE FOR NO DATA? I.E. IF THERE ISN’T ANY USEFUL DATA IN ANY OF THE POLICIES WE’RE GOING TO RETURN NOTHING WHATEVER WE CHOOSE

Decisions:

* How to deal with/ prioritise missing data
* Single policies or pick and choose
  + Issue with former is it means you have to have very clear ranking of policy importance. Methodology for this is to run through every possible combination of policies in order of permissiveness.
  + Issue with latter is it means you might be selecting policies which aren’t actually available. Methodology for this would be to pull out just id and each value then do the filtering process individually for each one. Probably would be quite easy since it doesn’t involve conflicts.

**Create new worksheet/ dataframe with the following values for each journal**

I copied the multiple pathways csv and removed all duplicated values (using ID). I then deleted all the values for additional\_oa\_fee, article\_version, license, copyright\_owner, conditions, location.location, and embargo.amount. This leaves all the columns which should have only one value per journal = title, id, issn\_print, issn\_electronic, listed\_in\_doaj, name, sherpa\_web, open\_access\_prohibited.

# Creating rules for filling the blank cells

**Additional\_oa\_fee**

If any additional\_oa\_fee = yes, return has\_oa\_fee\_option = yes

Remove all article\_version = submitted

**For assessing compliance of full gold**

If additional\_oa\_fee = yes,

* return new column license\_fee (according to hierarchy below)
* return copyright\_owner\_fee - if any author, return author,
  + if none author, and any publisher, return publisher
  + if none author or publisher return ?

**For assessing Green OA**

Filter out where additional\_oa\_fee = yes

Filter out where listed\_in\_doaj = yes

Create new row for ‘mp\_green\_license’:

* If any contains “cc\_by,” or “CC\_BY,” return cc\_by
* If no “cc\_by,” or “CC\_BY,”
  + If contains “cc\_by\_sa” or “CC\_BY\_SA” return cc\_by\_sa
    - If none of above, if contains “cc\_by\_nd” or “CC\_BY\_ND”, return cc\_by\_nd
      * If none of above, if contains “cc\_by\_nc” or “CC\_BY\_NC”, return cc\_by\_nc
        + If none of above, return 1st policy with “XX” in front of it

Create new row for “mp\_green\_embargo”

* Need to figure out how to distinguish between blank and 0 embargo
* If any contains 0 embargo return 0 embargo
  + If none return 0 embargo return lowest value

Create new row for “mp\_green\_copyright\_owner”. if any author, return author,

* if none author, and any publisher, return publisher
* if none author or publisher return ?

Option 1 – go back to work with publications

Option 2 – work with new data frame