Documentation for Publication Venue Gap Analysis

This document summarises the various code involved in the data preparation and analysis for the Open Access Publication Venue Gap analysis. The code was written by Tom Kenny.

# Data sources

## Dimensions

The base of articles used in this analysis was taken from Dimensions, filtering for articles and proceedings published between 2017-2020 and associated with UKRI funding (a total of 162,184 unique articles as of 1st April 2021). Dimensions is an inter-linked research information system provided by Digital Science (https://www.dimensions.ai). Dimensions is a comprehensive citation database (comparable to Web of Science or Scopus) that links around 128 million academic publications such as journal articles, books and chapters to publication metadata and metrics as well as grants and other sources of funding article metadata. It harvests its data from a variety of sources including Crossref, PubMed, Europe PubMed Central, arXiv and direct contacts with more than 130 publishers to create a baseline metadata model for publications data. It then uses deep-indexing of the full text of an article, to categorize the record, extract references, affiliations and funding sources etc. These are then used to link to other content types in Dimensions where possible and applicable[[1]](#footnote-1)

The key Dimensions variables used included: Title, Digital Object Identifier (DOI), Publication Date, Publication Year, Source (journal) title, Publisher, ISSN, e.ISSN, Publication Type, Country of Research Organisation, Fields of Research Categories, Open Access Categories, and Funders. Metadata for each of these [is available here](https://docs.dimensions.ai/dsl/datasource-publications.html).

## Sherpa

[Sherpa](https://v2.sherpa.ac.uk/) is a service provided by JISC which aims to help authors and institutions make informed and confident decisions in open access publication and compliance. We used [Sherpa Romeo](https://v2.sherpa.ac.uk/romeo/), which provides data on publishers' conditions for open access archiving on a journal-by-journal basis. Journals are only included in Sherpa Romeo if they have a valid ISSN and all policies are readily available online. For each journal on Sherpa Romeo, we identified the most permissive policy for Green and Gold OA and then merged this with the articles from Dimensions by matching on journal title and/or ISSN.

Data taken from Sherpa included: whether the journal is listed in DOAJ, article version, license, embargo, permitted respositories, and copyright owner.

## Unpaywall

Unpaywall is an open database of almost 30m free scholarly articles. It harvests Open Access content from over 50,000 publishers and repositories, and make it easy to find, track, and use. Essentially it aims to capture every single open access version of an article as well as key metadata (e.g. the version and the licence). There are two reasons we needed it for the gap analysis. First, because Dimensions’ Open Access Categories are mutually exclusive and rank Bronze OA over green OA. Since UKRI considers Bronze equivalent to closed, this would mean underestimating the proportion of articles compliant through green OA (since some Bronze articles also have green versions). To address this we rederived the Open.Access variable using data on green OA from Unpaywall. Second Unpaywall has licence information for OA versions which is not on Dimensions. We do not know how accurate the licence information is, and there is lots of missing data for green OA versions, however it is the best source we have.

Data from Unpaywall in the final data is the version of the most permissive repository copy (prioritizing licence then version), the licence of that copy, and the licence of the gold version if it exists.

## ESAC

ESAC aims to provide context and share information with the broader community on the progress of transitional license models such as “Offsetting” and “Read & Publish,” to increase the understanding of their scope and transformative power.[[2]](#footnote-2) For the gap analysis we used the [register of Transformative Agreements](https://esac-initiative.org/about/transformative-agreements/agreement-registry/), which includes details of current Transformative Agreements negotiated by JISC. We consulted with JISC on this list, and while there are no publishers erroneously associated with TAs, there were some publishers whose TAs were not recorded on ESAC at time of writing (April 2021). This meant we had to supplement the ESAC data by adding these publishers in manually (we confirmed the final list of with JISC).

The final list of publishers included (TAs negotiated by April 2021) was: American Physiological Society, Association for Computing Machinery (ACM), Bioscientifica, BMJ, Brill Academic Publishers, Cambridge University Press (CUP), Cold Spring Harbor Laboratory, The Company of Biologists, European Respiratory Society (ERS), Future Science Group, IOP Publishing, IWA Publishing, Karger Publishers, Microbiology Society, Oxford University Press (OUP), Portland Press, Rockefeller University Press, Royal College of General Practitioners, Royal Irish Academy, SAGE Publications, Springer (*not including Nature*), Taylor & Francis, Geological Society of London, The Royal Society, Thieme, De Gruyter, Wiley, Royal Society of Chemistry (RSC)

# Setting up the project folder

* The first step is to create an RStudio project file (see [blog on this here](https://martinctc.github.io/blog/rstudio-projects-and-working-directories-a-beginner's-guide/)), along with subfolders for
  + ‘Data’ = this is where all the data will be stored, it should also contain a subfolder called ‘Raw data’
  + ‘Outputs’ = this is where all the tables and charts will go
  + ‘Scripts’ = this is where all the code goes

# Data scripts

The first three scripts download, filter and clean the data ready for merging, further filtering and cleaning, and deriving analysis variables in the fourth script. Then analysis used in the impact assessment in the fifth script. These scripts were last run on the week commencing 5th July 2021, with the final merged dataset then uploaded to the DataHub

## SHERPA RoMEO API

* This code downloads data on journal OA policies from the SHERPA API in .json and then filters it to return only the variables we are interested in, and converts it into a data frame with one row for each journal policy (i.e. if a journal has four policies it will have four rows). It also does a bit of basic cleaning of the data. It then saves the file out as .Rda and .xlsx to be used in subsequent parts of the code.

## Dimensions import and cleaning

* This code imports Dimensions data and cleans it. The data was downloaded for us by Katie Shamash at Dimensions as this is the easiest option, but it could also be downloaded using the Dimensions API (see code for doing this in superceded folder).
* Once the data is imported, we:
  + removed duplicates (on both DOI and title),
  + simplified variables (e.g. reduce number of categories),
  + derived new variables (see below),
  + split up Springer Nature into ‘Springer’ and ‘Nature’, by recoding all Springer Nature journals with Nature in the title into ‘Nature’. This was because Springer and Nature journals have different deals with JISC.
  + Filtered the data to focus on 2017-2020 and articles only
  + saved out as .xlsx and .Rda for use by future code.

New derived variables

* ukri\_funders – this takes dimensions ‘Funder’ variable and pulls out UKRI funders only, concatenated and separated by “, “.
* for\_division – this is derived from Fields of Research categories in Dimensions. The FOR division is identified by removing all characters in the string before the first FOR group (“; [0-9]{4}.\*”), this works because FOR divisions come before FOR groups.
* for\_group – this is derived from Fields of Research categories in Dimensions. The FOR group is identified by selecting all strings after the pattern “[0-9]{4}[:space:]{1}.\*”. This works because FOR groups come after FOR divisions.
* discipline – this is derived from for\_division above. FOR does not have a less granular grouping than division, so I manually coded FOR divisions into ASJC journal classifications (since this was the grouping which had been used in another report for the OA analysis).[[3]](#footnote-3) One limitation is that the splitting up into divisions isn’t perfect since some divisions might fit into two disciplines. In these cases I chose what seemed the closest match.

## Unpaywall API and cleaning

* This code uses the Unpaywall API to import all OA versions relating to the DOIs in the Dimensions sample.
* It then filters for gold and VoR/AAM green versions and picks the most permissive for each according to the UKRI policy.
  + For green OA (hosted in a repository and is either published or accepted version) that is ranking according to (1) does it have a compliant licence according to the RCUK policy, (2) is it the VoR, (3) does Unpaywall identify it as the best OA version, and (4) which particular licence it has (ranked according to creative commons ranking).
  + For gold OA (hosted on publishers website, is the published version, and has a licence (i.e. not bronze OA)).
* New derived variables
  + Upw\_green\_licence: Licence of most permissive version e.g. CC BY
  + Upw\_green\_version: Version of most permissive version (AAM/ VoR)
  + Upw\_gold\_licence: Licence of publisher version (where gold version exists)

## Preparing and merging PVGA data

This script imports previously produced data from Sherpa and Dimensions, and imports new data on transformative agreements from ESAC. It then merges them all together and creates new variables from the merged data, in particular focusing on compliance with various policy scenarios.

* First Dimensions is imported and filtered to focus on articles and proceedings only, and publication years 2017:2020. We also removed articles not linked to an ISSN or journal title (these were almost all proceedings and it is not clear if they are in peer reviewed journals and thus within scope).
* Second, ESAC is imported using xml2 and rvest to read the table in from the ESAC website. We then filter for only JISC-negotiated (UK) deals. We then rename publishers so the names match with those in Dimensions (to facilitate matching). Finally we add in additional TAs which had been reported to us by JISC but not recorded in ESAC at the point of download.
  + Assumption: we assume that every article for each publisher in ESAC is covered by their TA with JISC. In practice this is not likely to be true (e.g. some articles do not involved UK institutions), however policy agreed it was the most reasonable assumption.
* Third Sherpa RoMEO data downloaded previously is imported, variables are simplified, and both green and gold policies are ranked to identify the most permissive for each journal. This process involves using expand.grid to create tables containing every possible combination of license, copyright, repository location, and embargo, ranking that table in order of permissiveness, assigning a rank to each combination, and then matching that rank with journal policies with that combination. Once the policies all have a rank, we then select only the most permissive green policy (additional\_oa\_fee == “no”) and the most permissive gold policy (additional\_oa\_fee = “yes). Finally, we merge the most permissive green and gold policies for each journal, so that each journal is only one row in the dataframe.
  + Assumption: where no information is recorded for embargo we assume no embargo is required (based on advice from JISC)
  + Assumption: where no information is recorded for license we assume the strictest license (based on advice from JISC). The main impact of this assumption is to likely lead to an underestimate of zero embargo, CC BY green OA.
  + Assumption: all policies which record additional\_oa\_fee == “no”, and do not apply to the submitted version only, and do not have their only location recorded as “this journal” are green. NB. in the analysis we need to exclude fully OA journal policies from analysis of green OA – this is because green OA for fully OA journals doesn’t really make sense
  + Assumption: All policies which have an additional OA fee are gold, as are all policies referring to the published version for journals listed in DOAJ, as are all articles which involve a VoR which is only available in ‘this\_journal’.
* Fourth, we import the Unpaywall data
* Fifth, ESAC is left joined into Dimensions, matching on the ‘publisher variable’ – this essentially adds the has\_ta column to the Dimensions data (see below)
* Sixth, Sherpa is inner joined into the Dimensions/Esac data – for each article where a match can be identified (on issn, e-issn, or journal title), this adds information on the journal policies for the journal the article was published in. NB this is not equivalent to the journal policies at the time of publication.
* Seventh, we add the additional variables from Unpaywall with a left\_join on DOI
* Eighth, we derive new variables based on the merged data (see below) and remove unnecessary variables from the merged data
* Ninth, we derive a set of new variables to understand potential compliance with various policy scenarios. For each policy scenario we again create tables ranking each combination of policies, rank them, and then apply this rank to the data as a new variable. We then use the results of those variables to assign observations to different categories in new compliance variables (see below). Finally, we simplify those variables.
  + Assumption – where we do not have license information for a gold OA policy we assume CC BY is allowed. This is because CC BY is allowed for the vast majority of gold publications (98%) so it is unlikely that deviation from this would not be specified in the policy. This applies to about 1200 articles in the sample (0.9% of those with a gold OA policy recorded in Sherpa and matched with dimensions).
  + Assumption – where we do not have license information for a green OA policy we assume CC BY is not allowed. This is because CC BY is not allowed for the vast majority (90%) of green publications where a licence is specified.
* Finally, the rows of merged\_pvga are randomized. This is because the analysis script uses !duplicated() and distinct() to reduce the data down to a single row per journal. The issue with this is that it is possible for individual journals to have different data e.g. on OA status because of inconsistencies in Dimensions (e.g. Dimensions may show some articles as closed and some as hybrid in the same journal, or policies may have changed since article was published). Randomising the order removes the risk of systematic bias in which article is selected for each journal (though this is still at best a proxy for analysis at a journal level).

New derived variables in final data

* has\_ta – every article by a publisher with a TA recorded in the ESAC data is ‘yes’ and every other article is ‘no’. This allows us to identify all articles covered by TAs.
* embargo\_grouped – simplifies embargo variable from exact number of months to either 0, 6, 12, or 12+ (12+ is recorded as 13 in this variable)
* embargo\_zero – simplifies further still to either zero embargo or has embargo
* license\_single = simplifies license data to make it easier to read and to merge categories which are essentially the same for our purposes (cc\_by & public domain & gnu\_gpl).
* license\_compliant - this pulls out only licenses which are of interest to our analysis (cc by, cc by-nd, cc\_by\_nc) and makes everything else ‘no compliant license’.
* license\_cc\_by – this further simplifies to only leave cc by and not cc by (this is useful for creating tables/ charts etc. in the analysis and for generating green rankings)
* compliant\_repository (later g\_compliant\_repository) – this identifies whether the policies specification of repository location is compliant with new policy.[[4]](#footnote-4)
* copyright (later g\_copyright and fee\_copyright) – this identifies whether the author retains copyright ownership of the article[[5]](#footnote-5)
* journal\_type – divides journals into Pure Gold (listed in DOAJ or article recorded as Pure Gold in Dimensions), Hybrid (not Pure Gold and has a policy associated with a fee or is article is recorded as Hybrid in Dimensions) and Closed (not Pure Gold or Hybrid).
  + Assumption: Journals listed in DOAJ are what we would consider to be Pure Gold. This is not necessarily the case as it appears a small minority may include Bronze OA, however, this number is considered too small to require adjustment.
* Open.Access\_ukri – this takes Dimensions variable Open.Access (open access categories) and edits it to match UKRI policy perspective on the categories. It removes the All\_OA value, merges green published and accepted into green, and merges green submitted into closed into closed. Finally, it splits Bronze OA into ‘Green’ or ‘Closed’ depending on whether a suitable green version was identified in Unpaywall for that article.
* articles\_per\_journal – the number of articles in our sample in each journal (i.e. this number will be repeated for every article in that journal in the sample)
* num\_fee – the rank allocated to the most permissive gold OA policy for each journal (from most permissive = journal type is Pure Gold, license is CC BY, to least permissive (journal type is closed, license is not cc by). This is used in the compliance variables.
* num\_current\_green – the rank allocated to the most permissive green policy under the old RCUK policy (allows cc by nd and nc in addition to cc by, allows 6 month embargo for all articles, and 12 month for AHSS)
* num\_new\_green – the rank allocated to the most permissive green policy under the new UKRI policy (stricter embargo and license requirements)
* compliance\_current – through which routes is the journal compliant with current (old) policy
* compliance\_new\_hybrid - through which routes is the journal compliant with a new policy scenario where all hybrid gold is allowed
* compliance\_new - through which routes is the journal compliant with the favoured new policy scenario (gold OA in pure gold or hybrids with TAs)
* compliance\_new2 – simplified version of compliance new (either compliant or not compliant)
* compliance\_new\_pure - through which routes is the journal compliant with a new policy scenario where gold OA is only allowed in pure gold journals
* ta\_split – for each article this variable shows in which year the TA was first introduced with the publisher of that article (if at all), or whether JISC are targeting a TA with that publisher

## Analysis for OA impact assessment

This script runs through the analysis used in the OA impact assessment, which is the main venue in which these findings are being published.

## Uploading final merged\_pvga to Datahub

* This code was used to upload the data created above to the datahub to allow reproducibility if anyone wants to approach this data again. Sherpa, Dimensions and ESAC all regularly update data (even of historic records), so findings do change over time. The data was updated w/c 5thth July 2021.

# Analysis scripts

The other scripts cover the main analysis and various different smaller bits of analysis. These scripts were last run on the week commencing 29th March 2021, with the final merged dataset then uploaded to the DataHub

1. <https://support-funder.dimensions.ai/support/solutions/articles/13000042666-how-is-the-publications-data-harvested-> [↑](#footnote-ref-1)
2. ESAC (2021), [esac-initiative.org](https://esac-initiative.org/) [↑](#footnote-ref-2)
3. Adapted from ASJC (Scopus) <https://service.elsevier.com/app/answers/detail/a_id/15181/supporthub/scopus/>

   physical\_sciences <- c("01 Mathematical Sciences", "02 Physical Sciences", "03 Chemical Sciences", "04 Earth Sciences", "05 Environmental Sciences", "08 Information and Computing Sciences", "09 Engineering", "10 Technology", "12 Built Environment and Design")

   health\_sciences <- c("11 Medical and Health Sciences")

   social\_sciences <- c("13 Education", "14 Economics", "15 Commerce, Management, Tourism and Services", "16 Studies in Human Society", "17 Psychology and Cognitive Sciences", "18 Law and Legal Studies")

   life\_sciences <- c("06 Biological Sciences", "07 Agricultural and Veterinary Sciences")

   arts\_humanities <- c("19 Studies in Creative Arts and Writing", "20 Language, Communication and Culture", "21 History and Archaeology", "22 Philosophy and Religious Studies") [↑](#footnote-ref-3)
4. compliant\_repository <- c("any\_website", "any\_repository", "institutional\_repository", "non\_commericial\_institutional\_repository", "non\_commercial\_repository", "non\_commercial\_subject\_repository", "subject\_repository", "non\_commerical\_website", "funder\_designated\_location", "preprint\_repository", "no repository requirement") [↑](#footnote-ref-4)
5. Author copyright = "authors", "authors\_institution", or "no copyright requirement" [↑](#footnote-ref-5)