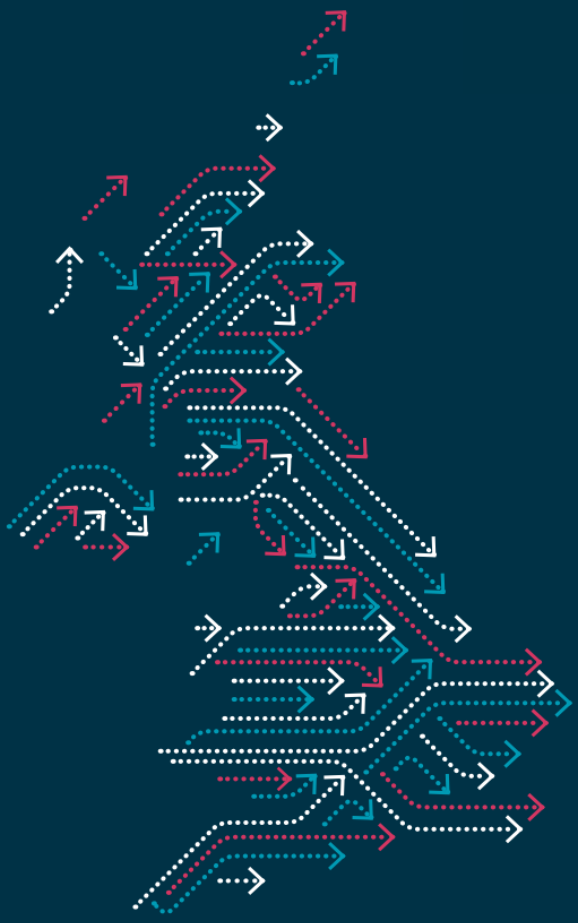
MONITORING THE TRANSITION TO OPEN ACCESS DECEMBER 2017

ANNEXE 1

NOTES ON METHODOLOGY



Universities UK
Research Consulting
Jubb Consulting
The University Of Sheffield.
ELSEVIER

MONITORING THE TRANSITION TO OPEN ACCESS: NOTES ON METHODOLOGY

AVAILABILITY OF OPEN ACCESS (OA) OPTIONS

The analysis is based on three sources:

- an analysis of the Scopus database, which records the business models of the journals it indexes. The global sample includes all journals; the UK sample includes only those journals in which UK authors have published in the years 2013–16
- evidence from a sample of 40 publishers, including all those responsible for the journals most popular with UK authors (see below). For the larger publishers, information on categories of journals, article processing charge (APC) levels, licensing and posting policies was sought direct, and checked where possible against information provided on their websites. For smaller publishers, information was gathered from websites
- a detailed analysis of the policies of 30 journals in four subject areas: medical and life sciences, physical sciences and engineering, social sciences, and arts and humanities. The selected journals were those in each of the four areas that had published the highest numbers of UK-authored articles in the years 2013–15. Information was gathered from journal websites, and in particular from the information for authors. Diligent search was made for information on APCs, licensing, and posting policies, including embargo periods. In some cases, in particular relating to posting policies, the information was unclear, difficult to interpret, or not available at all after diligent search

Analysis of APCs does not take account of discounts provided for members or for other reasons, but it does take account of a few cases where additional amounts are charged for the use of specific licences. Analysis of licensing and posting policies takes account of journals where those policies relate to whether or not an author is funded by an agency that requires the use of a Creative Commons Attribution or other licences, or the deposit of articles in a specific repository.

ACCESSIBILITY

Two main approaches were used for this assessment:

- a census of all sources covered in Scopus to determine the publishing models used. This enabled counts of journals and articles by the main publishing options: Gold – APC; Gold – no APC; Hybrid; Delayed OA; Subscription. A direct-counting approach determined the level of Gold – Hybrid (where an APC has been paid for immediate OA within a hybrid journal)
- a sample-based approach to determine the level and type of publicly accessible online postings to various repositories (Green OA)

Census-based approach

- Scopus publication data were extracted from the SciVal Analytical Services Scopus database, a database snapshot of Scopus data created in May 2017.
- Only publications corresponding to the four main peer-reviewed document types in Scopus were included: 'Article', 'Review', 'Conference Paper' and 'Short Survey'.
- For each source covered in the dataset, aggregate counts of publications and citations in appropriate time windows were compiled, and advanced citation indicators (such as field-weighted citation impact or FWCI – see box below) were calculated. This was repeated for publications where at least one of the authors has an affiliation to a UK institution.
- Counts were made for 2012 and 2014. As the snapshot was taken before full 2016 data were processed in Scopus, the full-year 2016 values were extrapolated based on Scopus coverage profiles to the end of 2015.
- While journals are the predominant source for peer-reviewed articles, some articles are also found in edited collections or conference proceedings. These were also included where covered by Scopus.

The **field-weighted citation impact (FWCI)** is the ratio of actual citation count to expected citation count for any grouping of articles, such as those published in a single source or under a given business model. It takes into account the differences in publication and citation behaviour across disciplines, and for the prevalence and citation rates of different document types. A value of exactly 1.00 means that the source is cited at the expected rate, while a value greater than 1.00 means that the output is cited more than expected. FWCI uses a single publication year and up to five years of citation thereafter (or as data currency allows); for example, the FWCI reported for 2012 includes publications in 2012 and citations received in 2012–16 inclusive, while the value reported for 2014 includes publications in 2014 and citations received in 2014–16; the relative nature of the FWCI means that such shifting windows necessitated by the currency of data and the lagging nature of citation accrual do not alter the validity of the measure.

Journal and article publishing models

Publishing models were assigned to each source in Scopus using a combination of the Directory of Open Access Journals (DOAJ), which mostly comprises Gold – APC and Gold – no APC journals, and desk research on publisher price lists and catalogues, and individual journal website information on publishing models (Gold – APC, Gold – no APC, Hybrid or Subscription). For the latter two classes, if there is a journal-specific delayed access policy, these were classified as Delayed OA. Further manual checks were done to improve the classification of publishing models as follows:

- All sources without an assigned publishing model after the first round described above and with article counts greater than 500 in any recent publication year were assigned using individual journal website information.

- The 50 largest journals assigned as Gold – APC were also spot-checked to ensure the assignment was correct.
- All remaining titles were assumed to be Subscription.

Estimating Gold – Hybrid

For several of the major publishers of hybrid journals (including Elsevier, Springer Nature, Wiley, Taylor & Francis, Royal Society of Chemistry, American Chemical Society and Cambridge University Press), advanced searching on each publisher's database platforms was used to determine the overall and UK-authored uptake of the Gold – Hybrid option in 2016 in hybrid journals for the main peer-reviewed document types.

SAMPLE-BASED APPROACH

Scopus publication data was extracted from the SciVal Analytical Services Scopus data feed, a syndicated version of Scopus data that permits extraction of article-level metadata with a weekly refresh.
Only publications corresponding to the four main peer-reviewed document types in Scopus were included: 'Article', 'Review', 'Conference Paper' and 'Short Survey'.

Sampling plan

For each period to be analysed (May–June 2017, November–December 2016, May–June 2016 and May–June 2015), all documents with publication dates within the period were extracted from the data feed and were randomised as follows: each document was assigned a random number and then sorted from smallest to largest, then assigned a fresh random number and so on until the process had been repeated three times. The required number of documents were then taken from the top of the final sorted list (4,464 documents, split equally between the four periods). This was repeated for publications where at least one of the authors has an affiliation to a UK institution (2,762 documents, split equally between the four periods). For reporting, publications where at least one of the authors has an affiliation to a UK institution in the global sample were grouped with this UK sample to increase robustness.
The publishing model assigned to each source in the Scopus database for the census-based approach (see above) was applied to the sample data to ensure consistency and comparability across all analyses.

Search and coding

An algorithm was created to derive from each sampled document key metadata elements (including the document title) and an automated Google query was designed to replicate human search behaviour: from only the top 10 links returned by the Google search, unique links were stored. Of these links, those indicating HTTP response status codes (eg 404 Not

Found and 502 Bad Gateway) and domains shown through manual verification to never contain full-text copies of sample documents were removed.

The remaining links were stored in a database and marked by a temporary workforce of two individuals trained to code each link as representing a full-text version of the document in question (marked 'TARGET') or not (marked 'NOT TARGET'). A link was marked as 'TARGET' if two criteria were fulfilled: the document was (a) recognisably the same as the published article being searched for, typically indicated by the article title (with caution exercised for very generic titles or articles with very similar titles), and (b) a full-text copy of the document was available (not an abstract or just the first page).
Those links identified as 'TARGET' were then coded manually to differentiate between:

- preprint (PP)
- author's version prior to submission for publication
- accepted author manuscript (AAM)
- author's version accepted for publication after peer review and that incorporates any revisions required
- version of record (VoR)
- published version, complete with volume/issue/pagination and the imprimatur of the journal and its publisher

However, the differentiation between PP and AAM versions is notoriously difficult and depends on often subtle markers in the text of a document; the guiding principle used was that versions lacking any indication that they have been accepted for publication in a journal were classed as PP (this includes working papers in fields where these are used), while those showing some indication that they have been accepted for publication in a journal were classed as AAM. Often, the latter have watermarks or text on the title page making their status obvious, but if the acknowledgements section (where present) mentioned the contributions of anonymous peer reviewers to the improvement of the manuscript, this was deemed to constitute evidence that the paper had passed peer review and so should be considered as an accepted author manuscript.

Each link's root domain was assigned to a website class (eg, social sharing network, institutional repository, etc). For each 'TARGET' document, all versions and locations in which it appears were recorded: for example, a publication for which a PP version appears on an author's homepage and an AAM is deposited at a subject repository will both be recorded, but of course de-duplicated in aggregated counts where necessary in subsequent analysis.

Adherence to journal/publisher posting policy was assessed on the basis of information at howcanishareit.com
In the final analysis, only those documents published under the Subscription model and found in one or more versions at non-publisher websites were included for analysis (4,182

documents for the global sample and 2,471 for the UK sample, split between the four periods).

USE OF OA ARTICLES

The analysis is based on data from a range of sources:

- data from the Journal Usage Statistics Portal (JUSP), which aggregates data on downloads of articles via publishers' and intermediaries' sites from 180 university libraries in the UK. The data includes a separate category for articles marked as OA
- data from a number of publishers who responded to requests for information
- data from IRUS-UK, which aggregates data from 110 UK institutional repositories. Repositories hold a range of different types of content, and the mixed quality of the metadata means that it is not always possible to identify which items relate to published articles. Hence, the figures have to be treated with some caution.
- data from PubMed Central

We were unable to secure data on downloads from any of the sharing sites such as Research Gate or Academia.edu

FINANCIAL IMPLICATIONS FOR UNIVERSITIES AND RESEARCH FUNDERS

Data used in the study came largely from published datasets rather than having to be gathered for the purposes of the research. This is a very positive step forward compared with earlier work.

UK data for article processing charges (APCs) from 2013–16 were derived from datasets made available by Jisc on GitHub (Shamash, 2017). These data were processed before their publication by Jisc in various ways, including data cleaning, normalisation and de-duplication as described by Shamash (2017). The data was generally used in its published form, with the exception of work carried out to supply missing journal 'type' information ('full OA' or 'hybrid'). In the original 2016 dataset for the full sample of 38 institutions, 2,675 of the 11,914 total records were blank in the journal 'type' field. Of the sample of institutions used for the longitudinal analysis (2013–16, 10 institutions), 1,073 of the 4,200 total from 2016 were unknown, 165 of the 1,774 from 2015, and 71 of the 1,234 from 2014. Journal types were supplied for these records in two stages. First, the journals in the APC dataset were matched with the journal categories ('full OA' or 'hybrid') assigned for the analysis of Scopus data for chapter 2 of the main report. The automatic matching eliminated most of the 'unknown' journal types but did not eliminate them entirely, and 780 remained unallocated. A random sample of 20 records with 'unknown' journal type were checked manually and were confirmed to follow a similar proportion to that of the dataset overall. It was therefore decided to allocate these remaining 'unknowns' formulaically (the second stage of the process), assigning them using the same proportions as the known data in terms of numbers of APCs to the different categories of 'full' or 'hybrid' journals. This was done separately for each year, 2014 to 2016, according to the numbers for the relevant year. The dataset published with this study, therefore, shows these matches.

The fact that this automatic matching was necessary illustrates the point that the data available is still imperfect, with somewhat different approaches taken by institutions contributing to the Jisc dataset to data accuracy and completeness. One key recommendation of this study is that institutions should agree greater consistency in their approach to recording data in the agreed template. At present, institutions are making very different assumptions about, for example, how they account for offsetting arrangements, how they deal with split payments, and the priority they give to noting funder etc. Greater consistency across all of these and other areas would enhance the dataset considerably. Additional datasets consulted for comparison purposes for APCs included UK data made available by RCUK (RCUK, 2017) and the Wellcome Trust (Wellcome Trust, 2017), and international data made available by the OpenAPC service (OpenAPC, 2017). Subscription data for UK institutions used is available on Figshare (Lawson, 2017a; Lawson & Meghreblian, 2014; Lawson, Meghreblian & Brook, 2015).

Other publications to which reference was made for the purposes of interpreting our analysis include Lawson's (2017b) report on UK offset agreements and Pinfield, Salter & Bath's (2017) study comparing APC and subscription data from the UK.

We are grateful for the contributions of the sample institutions for their data, for the work of Stuart Lawson (and colleagues) in assembling subscription data and in his work on offset agreements, and in particular for the work of Katie Shamash of Jisc in processing APC data and for her help in interpreting the data during our study.

REFERENCES

Lawson S (2017a) *Journal subscription expenditure in the UK 2015–16*. Retrieved 2 December 2017, from https://figshare.com/articles/Journal_subscription_expenditure_in_the_UK_2015-16/4542433

Lawson S (2017b) *Report on offset agreements: Evaluating current Jisc Collections deals. Year 2 – Evaluating 2016 deals*. Retrieved from https://doi.org/10.6084/m9.figshare.5383861.v1

Lawson S & Meghreblian B. (2014) Journal subscription expenditure of UK higher education institutions. *F1000Research 3*. http://doi.org/10.12688/f1000research.5706.1

Lawson S, Meghreblian B & Brook M (2015) Journal subscription costs – FoIs to UK universities. Retrieved 2 December 2017, from https://figshare.com/articles/Journal_subscription_costs_FOIs_to_UK_universities/1186832

OpenAPC (2017) OpenAPC. Retrieved from https://github.com/OpenAPC/openapc-de

Pinfield S, Salter J & Bath P A (2017) A 'gold-centric' implementation of open access: Hybrid journals, the 'total cost of publication', and policy development in the UK and beyond. *Journal of the Association for Information Science and Technology* 68(9), 2248–2263 http://doi.org/10.1002/asi.23742

RCUK (2017) *RCUK open access block grant analysis 2013/14, 2014/2015 and 2015/2016*. Research Councils UK. Retrieved from http://www.rcuk.ac.uk/documents/oadocs/rcukapcreturnsanalysis2014-16-pdf/

Shamash K (2017) Article processing charges. Retrieved from https://github.com/kshamash/Article-processing-charges

Wellcome Trust (2017) Wellcome and COAF open access spend 2015-16. Retrieved from https://wellcome.ac.uk/funding/managing-grant/wellcome-and-coaf-open-access-spend-2015-16

FINANCIAL SUSTAINABILITY: LEARNED SOCIETIES

The methodology followed in our work on learned societies can be summarised as follows:

Step 1: For the purposes of our original 2015 study, we developed a comprehensive list of potential organisations for inclusion from the following sources:

- UK learned societies listed by Europa World of Learning
- a list of learned society members supplied by the Association of Learned and Professional Society Publishers (ALPSP)
- a list of UK learned societies found on Wikipedia
- The British Academy Directory of Subject Associations and Learned Societies in the Humanities and Social Sciences
- the approved list of professional organisations and learned societies identified by HM Revenue & Customs (specifically those in the list that were identified as allowing reclaimed tax on journal subscriptions and other publications)
- the list of members of the Academy of Social Sciences

Step 2: From a consolidated list of nearly 600 societies, we identified those societies with their primary, registered address in the UK.

Step 3: Among the identified UK learned societies, we selected only those that publish academic journals or conference proceedings (ie, peer-reviewed publications with an ISSN).

Step 4: We then identified the number of journals/proceedings published by each society (societies publishing only one journal, those publishing two journals and those publishing three or more journals), and the value of their incoming resources/turnover for the most recent available financial year. For those with a turnover exceeding £10 million, we also recorded the value of their income from publishing.

Step 5: We established how many self-publish journals and how many are contracted out, and recorded the identity of any third-party publishing partner.

Step 6: We categorised the societies by discipline using the classifications adopted by the UK's Research Excellence Framework (REF)[1] (an indicative classification only, given that several societies have a multi-disciplinary focus).

Step 7: From the list of UK learned societies producing academic publications, we selected a stratified sample of 25 organisations reflecting the characteristics of the broader population

[1] REF classification: disciplines falling under panel A (medicine and biological sciences), panel B (maths, physics, natural sciences and engineering), panel C (social sciences) and panel D (arts and humanities)

of learned societies[2], supplemented by a further judgemental sample of five UK societies with high levels of publishing activity.
Step 8: We analysed the financial statements of the selected 30 learned societies, based on the published financial statements from the 2011 to the 2015 calendar years (which is the most recent year for which data is consistently available).

Limitations in the availability and reliability of financial data

We chose to draw on published financial information to complete our work since virtually all societies, whether registered charities or companies, are required to provide this information annually, and make it publicly available. Statutory financial statements must be prepared in accordance with generally accepted accounting practice (GAAP) and, in the case of those societies that are charities, the appropriate Charity Commission Statement of Recommended Practice (SORP). Nevertheless, it is important to acknowledge that the information disclosed on learned societies' publishing revenues varies in scope and quality, and is often not directly comparable between societies. The level of publishing revenues disclosed depends not only on the total income generated by a journal, but also on the precise terms of the agreement between the society and any third-party publisher. For example, in some cases, a third-party publisher may only pass the net revenues generated by a journal on to a society, meaning the total value of subscriptions revenue is not reflected in the society's accounts. Many publishers also operate websites and provide other services to learned societies, which may be invisible from an accounting perspective, but can be of vital importance in practice. Finally, practices in cost and overhead allocation are also highly variable, and these could have a significant bearing on the figures reported for the surpluses generated from publishing.

In a small number of cases, some relevant information on the sampled societies was unavailable, particularly in the case of measures such as expenditure on publishing, and income/expenditure on peer-reviewed journals. In such cases, we either used the best available data, or excluded the society in question from some elements of the analysis.

Financial values and metrics adopted

A set of key financial values and metrics were identified to allow the large amount of data gathered to be analysed effectively, and have been categorised as either 'income and expenditure metrics' or 'financial health metrics'. The chosen financial values and metrics for income and expenditure are as follows:

- **total income:** defined as total incoming resources or total revenue
- **net income:** defined as net incoming resources, equal to total incoming resources less total resources expended; or operating deficit/surplus, equal to total income less

[2] Characteristics reflected in the sample: (a) the different levels of publishing activity by learned society (LS); (b) the overall number of active LS publishers across the four groups of academic discipline used in the REF; (c) the proportion of large and small LSs; (d) representative proportion of LSs operating with (22) and without (8) a publishing partner.

total expenditure. Accordingly, net income does not include other recognised gains/losses such as gains/losses on investment assets, finance income, taxation and staff pension scheme, etc

- **total income from publishing:** defined as total income from sales of peer-reviewed journals, monographs and other publications; journal royalties and online journal subscriptions. Income from member subscriptions is only included in this amount in a small number of cases where access to the peer-reviewed journals is deemed to be the primary benefit of membership
- **total charitable expenditure (excl. publishing):** defined as resources expended on charitable activities, less total publishing expenditure
- **publishing income as a percentage of total income:** defined as the ratio (expressed as a percentage) of total income from publishing to total income
- **net income from publishing as a percentage of total publishing income:** defined as the ratio (expressed as a percentage) of net income from publishing to total income from publishing, where net income from publishing is defined as total income from publishing less total expenditure on publishing. Total expenditure on publishing is calculated as the sum of journal expenditure, other publication costs, and costs associated with online journal subscriptions
- **net income from publishing as a percentage of charitable expenditure (excl. publishing costs):** defined as the ratio (expressed as a percentage) of net income from publishing to total charitable expenditure (excluding any publishing costs included under this heading)

Those for financial health are:

- **net assets:** defined as total assets (fixed assets and current assets) less total liabilities/creditors
- **discretionary funds/reserves:** defined as total unrestricted funds (excluding any designated funds at the financial year-end)
- **cash at bank and in hand:** defined as sum of cash at bank and in hand, cash held in liquidity funds and short-term deposits
- **discretionary funds/reserves as a percentage of total income:** defined as the ratio (expressed as a percentage) of discretionary funds/reserves to total income
- **current ratio:** defined as the ratio (expressed as a number) of total current assets to total current liabilities/creditors (amounts falling due within one year) at the financial year-end
- **liquidity:** defined as net current assets expressed as number of days' expenditure

Additional qualitative research

In order to get a better understanding of the strategic thinking underlying the published reports, all 30 sampled societies were invited to participate in a qualitative interview: 15 accepted and constitute a broadly representative sample. The interviews were conducted by

Robert Dingwall of Dingwall Enterprises Limited, and followed the question set provided at Appendix A. Notes were taken, and these transcripts were reviewed and coded to identify key themes arising from the interviews.

Finally, an open discussion event was held at the Royal Society of Biology on 27 September 2017, at which preliminary results were made available for review by a wider group of society and publisher representatives. Details of the event were circulated through a number of representative bodies and publisher trade associations, and promoted via social media. Approximately 50 individuals attended the event, representing some 40 societies, across the full range of disciplinary areas.

Feedback derived from the interviews and the event was used to inform the final conclusions of our work, as reflected in the main body of the report.