

Annex B: Compilation of the Scopus databases

1. This section explains the processes undertaken to load the raw Scopus data files, match the pilot institution's data to it, and construct bibliometric indicators for the purpose of the REF pilot. This process was undertaken in-house at HEFCE.
2. The key steps in the generation of the database used in the pilot are as follows:
 - a. Loading the raw Scopus data to a database server, so that we can perform the analyses described.
 - b. Making the citation links and citation counts for each document in the database.
 - c. Generating normalisation factors for each item in the database, against which we benchmark the papers under assessment.
 - d. Receipt of pilot data set from Symplectic via Evidence and matching to Scopus.
 - e. Extracting the appropriate records from the database, and calculating bibliometric indicators for these items:
 - i. For the author-based models, we select records by matching the publications data provided to us by Symplectic via Evidence to the database.
 - ii. For the address-based models we extract records based on institutional address.
 - f. Aggregating these paper-level indicators to the submission (that is, UOA) level and generating submission level indicators, such as median citation score, from them.
3. The focus of this section is on steps a, to c . Step d is discussed in paragraphs 11 to 22 of this annex. For the author-based models, step e is identical for both the Scopus and Web of Science databases, and is discussed in paragraphs 68 to 79 of the main report. The process for the address-based model for Scopus is briefly discussed in this section; the corresponding process for Web of Science is discussed in Annex C.

Loading the raw Scopus data

4. Elsevier provided the Scopus data in a set of GNU PGP encrypted zip files, with each zip file containing the data for 10,000 articles, with each article's data stored in a separate XML file. As part of the data transfer process Elsevier also provided the XML Schema files that the article XML files used.
5. From the XML Schema files a relational database structure was built that followed the structure of the XML Schema and allowed for the XML files to be loaded with minimal processing. Once the database structure was created a C# program was written which decrypted and unzipped the files provided by Elsevier and loaded each XML file to the database. The loading process took 10 working days to load the 11.8 million articles and the final database is 366GB in size.
6. In performing the loading, we assigned each record (corresponding to a single output, for example an article in a journal) a unique identifier, to which we refer as the

'keyid'. The keyid is not a part of the Scopus data, as supplied, but provides a key which we can use when manipulating the data. It serves an equivalent role to the 'UT' identifier used in Web of Science.

Constructing citation links

7. We constructed a table of citation links by matching on the Scopus supplied identifier 'SGR', to produce a table of citing keyids and cited keyids. In other words, this table indicates the links between items in the Scopus database.
8. There were multiple citation links between some pairs of source and target documents. These were assumed to be in error, and were reduced to a single instance of the link. There were 148,885 document pairs that had multiple links in this way. There were also some links where the source and target document were the same. There were 1,238 such records. These too were removed from the links table. There are nearly 64 million unique and non-self links in the database.
9. We computed the number of citations that each document has acquired by summing the number of times each was a 'target' document, in other words it was cited by another document within the database. Papers with no 'incoming' links are uncited.
10. Even though we only include articles and reviews in our bibliometric analysis, we include citation links to these items from anywhere in Scopus, including conference proceedings, letters, and so on. This contrasts to Evidence's analysis, where only citations from articles and reviews are counted.

Constructing normalisation factors

11. As we discuss in paragraphs 57 to 60 of the main report, we need to benchmark the citation count of each output that is being assessed against a normalisation factor. This allows us to take account of the accumulation of citations with time, the document type and the subject area of the work. In this section, we discuss the process of constructing these normalisation factors in Scopus.
12. Scopus features a classification of journals into subject areas; the all science journal classification (ASJC). These 334 categories form the basis of our normalisation groups. We also take account of document type (whether the item has been classified as an article or a review), and publication year.
13. The subject categorisations used in Scopus assign some journals to several subject categories. When computing normalisation factors (as discussed in paragraph 64) we include each item wholly in all of the subject categories to which it has been assigned; we don't fractionally apportion papers between the categories.
14. The ASJC code(s) are applied at the level of the journal; every item in a journal will, therefore, be assigned to the same category or categories. This is with the exception of one subject category, 'multidisciplinary science', which we discuss in paragraph 17 below.

15. We compute the average number of citations per paper in each normalisation ‘family’ of papers of the same document type, publication year and subject classification. Each factor is computed as¹:

$$\frac{\text{(number of citations to items in document type/year/classification group)}}{\text{(number of items in document type/year/classification group)}}$$

16. A small proportion of journals in Scopus have not been given ASJC categories. These are the journals that come directly from Medline and are not otherwise covered in Scopus. We cannot compute normalisation factors for these items, so they take no part in the analysis.

Multidisciplinary science

17. The subject category “multidisciplinary science” contains journals whose remit is very broad. It includes journals such as Nature, Science and the Proceedings of the National Academy of Sciences (PNAS). So that we can more easily compare like with like, we reassign papers in the multidisciplinary science category to more appropriate categories. This is achieved by using the bibliographies of the multidisciplinary items to infer the subject area they are a part of.

18. To do this mapping, we take each item in a multidisciplinary journal² and look at the subject categories of the journals it refers to in its bibliography. We assign unit weight to each paper in the bibliography, and split this weight equally over all the subject categories to which the paper cited is assigned. We then sum up the weights attributable to each subject category, and assign the item we are reclassifying to the subject category with the largest weight, provided this is at least two. If the item has equal weight attributed to several categories, it is assigned to them all.

19. We perform this process for all items in multidisciplinary science journals (not just those submitted to the pilot). We do this so that the normalisation factors we compute take account of the reclassification process. This does mean that the multidisciplinary science category has a rather low normalisation factor, because most of the work in it has been removed by the reclassification process.

20. Many of the items in the multidisciplinary items’ bibliographies will refer to work produced before the coverage of our extract of Scopus (that is, pre-2001). We handle these cases by constructing a ‘look-up’ table of journal name variants to ASJC codes, using the journal names and ASJC codes in the data that we hold.

¹ Some of our normalisation factors may be zero, if none of the items in the normalisation group have ever been cited. We set these values to ‘NULL’, to avoid ‘divide by zero’ errors when computing normalised citation scores. Note that an item in a normalisation group with a normalisation factor of zero can never have been cited (otherwise the group’s normalisation factor would be greater than zero). As such, we give these items a normalised citation score of zero at the end of the analysis.

² Note that some journals are assigned to ‘multidisciplinary science’ and additional All Science Journal Classifications (ASJC). There are treated as purely multidisciplinary journals in this analysis.

Matching pilot data to Scopus

21. We received three tables of data from Evidence's contractor, Symplectic. These are described in greater detail in Evidence's report on the development of the pilot databases³. Symplectic supplied us with a table of staff, a table of outputs, and a table of links. The links table matches staff included in the pilot to their publications. We only used the data contained in the table of outputs in our matching to Scopus. We discuss how we used the Symplectic author data to generate sets of outputs representing submissions in the section on staff-based models in the main report.

22. The Symplectic outputs table contained two types of output record; those supplied by the pilot HEIs and the additional outputs found with Evidence's presumptive database, based on institutional address. Where an output had been supplied by both an institution and matched in Evidence's presumptive database, it shared the same Symplectic identifier. We treated such records independently (that is to say, we did not attempt to construct a 'composite' record for outputs supplied by both an HEI and Evidence, using the data from both). At the end of our matching process, we used these records to obtain a crude bound on the level of accuracy in our matching process, as described in paragraph 32. The records that were found by Evidence had the Thompson 'UT' field corresponding to their record in Web of Science included in the outputs table. As described in paragraph 29 we were able to use this to augment the data in the Symplectic outputs table, in order to allow a stronger match to be made.

23. The data we received from Symplectic had been de-duplicated: where a paper had been submitted by several institutions, these were linked (via the table of links) to a single record⁴ in the table of outputs. This means that we cannot trace a record back directly to a submitting HEI, because we do not know which of the submitting HEIs returned the record that is retained in our outputs table. This does not present any problems with our analysis, but it does have implications when returning information to HEIs, and in auditing the data.

24. We adopted a sequential approach to matching the data to Scopus, because this provided a relatively straightforward method of matching the outputs table to the bibliographic database. In a real implementation of the REF, it is likely that we would adopt a more complex methodology, and would iterate the results of our preliminary matching with institutions (probably interactively) in order to maximise the number of and quality of the matches used.

25. Evidence matched the Symplectic data to Web of Science independently. The approach they took is described in paragraphs 130 to 134 of their data collection report⁵.

³ 'Pilot study of bibliometric indicators of research quality: Development of a bibliographic database. A report to UK HE funding bodies by Evidence Ltd' (July 2009). Available at www.hefce.ac.uk under Publications/Research & evaluation.

⁴ Or pair of records, if the item had been found both by an HEI and in Evidence's presumptive database.

⁵ See footnote 3

26. In our sequential matching process, we adopted the following, iterative procedure:
 - a. We took all items in the Symplectic table of outputs and attempted to match them to Scopus using a given match key (as described below).
 - b. We kept all matches that mapped to a single record in Scopus, and removed these from further matching attempts.
 - c. If a Symplectic record matched to more than one record in Scopus, then none of these were kept, and the Symplectic record was used on successive matching iterations.
27. At each stage of the above procedure, we monitored the number of 'one to many' matches that were present, because these were indicative of unduly lenient match keys. The match keys described below are the result of the monitoring of this.
28. We initially matched items using their Digital Object Identifier (DOI). In principle, this provides a unique match to a single article. In cases where the DOI existed, this was generally the case, though we found some examples of items in Scopus sharing the same DOI. This may be legitimate (for example, some journals assign the same DOI to all of their 'letters' page, rather than assigning a DOI to each letter on it), in other cases it appeared to be due to data error in Scopus or from the pilot HEIs. We have since found a few instances where, although the match on DOI was unique, it has linked the output as supplied to an incorrect output in the database.
29. Unfortunately the Symplectic outputs table did not include journal ISSNs. We made use of the Thompson 'UT' included with the records supplied by Evidence to add ISSNs to these records, using the raw Web of Science data that we hold. We then used these augmented records to match to Scopus, using the ISSN, volume and first page of the item.
30. In the next step, we 'cleaned' the journal titles by removing all non-letter characters from them. Scopus captures both the full source title, and an abbreviated source title. We used either of these, the volume and start page as our next match key. We removed non letter characters from the titles to allow cases like 'Phys. Rev. E.' and 'Phys Rev E' to match. The matching in this step was case insensitive.
31. Finally, we matched on publication year, and the first sixty characters of the cleaned item title (i.e. we stripped out everything besides letters). Although this may seem a rather 'loose' match key, recall that, at every stage of this procedure we kept only matches that linked to a unique record in Scopus. The matching in this step was case insensitive.
32. In addition to visually inspecting a sample of the matched records to check that the process had worked correctly, we can run a simple cross check by looking for cases where the university and Evidence supplied record have each been matched to different items in Scopus. We find 81 cases of this, out of approximately 250,000 matched records (both institution and Evidence supplied), which suggests that the matching procedure we followed is fit for purpose.

Address-based models

33. In this section we briefly describe the method we used to map Scopus ASJC categories to RAE 2008 UOAs. Our use of this model is discussed in paragraphs 61 to 67 of the main report.

34. We used the RAE 2008 database to look at the frequency with which work from each journal was submitted to each UOA. We used the International Standard Serial Number (ISSN) as a proxy for journal. We used these values to construct a mapping that compared the similarity of the journals assigned to each ASJC with the frequency with which the journals were submitted to each UOA⁶.

35. We used this to assign each ASJC code to the UOA to which it was most similar. This process maps each ASJC code to one UOA. It does not, however, ensure that all UOAs have any subject categories mapped to them. For example, in this methodology UOA 05 has no subject categories mapped to it; all the ASJCs map more strongly to other UOAs.

36. There were some cases where an ASJC did not map that strongly to any of the UOAs, and some cases where an ASJC mapped strongly to several UOAs. In these cases we used the mapping that was strongest, even if it was not particularly strong, otherwise the ASJC could have been mapped to several UOAs.

37. Although this process forces each ASJC to be mapped to precisely one UOA, it does not ensure that each item of work appears in at most one UOA in the analysis of the address-based model. This is because each journal (and therefore item) may be assigned to more than one subject category. In common with the author-based models, we take the approach of allowing each item in the assessment to appear at most once per submission. For example, if a journal is assigned to three subject categories, two of which have mapped to the physics UOA and the other to the chemistry UOA, the output will appear once in the institution's physics submission and once in its chemistry submission.

38. Given more time, we would have explored the effect of allowing ASJCs to be mapped to several UOAs (where they were almost as strongly mapped to several areas), and also, perhaps, the effect of pro-rating subject categories to UOAs. For example, if an ASJC maps twice as strongly to the Physics UOA as the Chemistry UOA, we could map the item to both UOAs, with a two-thirds: one-third weighting. This would, however, have further increased the 'fuzziness' of the mapping to UOAs.

⁶ Further details of the methodology we followed are available on request.