July 2009

Pilot study of bibliometric indicators of research quality: Development of a bibliographic database

Report to UK HE funding bodies by Evidence

Contact details

Evidence Ltd	103 Clarendon Road, Leeds LS2 9DF
t/	0113 384 5680
f/	0113 384 5874
e/	enquiries@evidence.co.uk

Evidence Ltd is registered in England, Company no 4036650, VAT registration 758 4671 85
http://www.evidence.co.uk

Contents

Executive summary

This report covers the work required to address the development of an initial bibliographic database to evaluate the feasibility of a Research Excellence Framework (REF) methodology. Other reports[1] assess the workload and challenges faced by the contributing universities and colleges, but acknowledgment is made here of the extensive support and enthusiasm extended by the staff in those higher education institutions (HEIs).

The REF is intended to make more extensive use of quantitative research performance indicators than the Research Assessment Exercise (RAE). The metrics discussed in reference to the REF are restricted to 'bibliometrics', which are the indicators created by an analysis of research journal articles and their subsequent citations. The collation and normalisation of citation data for the bibliographic database and the evaluation of variant bibliometric analyses will be described in later reports.

The census period of the exercise is 2001-2007. Data were supplied by a group of 22 pilot HEIs. These were selected by the Higher Education Funding Council for England (HEFCE) to cover a wide range of research management systems and processes. Subject areas were captured within 35 Units of Assessment (UoAs) selected by HEFCE because they had 40% or greater coverage of RAE-submitted outputs in principal commercial data sources (either Thomson Reuters' Web of Science® or Elsevier's Scopus). Not all pilot HEIs elected to supply data for all UoAs.

Preparation and specification

The project was launched in June 2008. It was expected that the development of the bibliographic database would take up to six months. This was a challenging timetable both for the pilot HEIs involved and for the contractors. Data collection took place over the summer, when many HEI staff were on leave. It was therefore agreed that REF pilot data specification should match the RAE2008 data collection as closely as possible. This would reveal the challenge of implementing a national exercise and provide important information about the current readiness of data management systems in the higher education research base.

The pilot work was designed to compare two variant approaches: a low-burden *address-model*, with data collated by address and linked to subjects via journal categories; and a more onerous *author-model* in which outputs are linked to subjects via author-staff disambiguation. To ensure that sufficient data would be available for each pilot HEI, the project made use of a *presumptive* dataset for each institution, supplied by *Evidence* from prior work to collate institutional article records. The presumptive data would form the entire database required for the address-model variant. *Actual* data are those article records already collected by institutions from their staff and therefore explicitly validated as part of the publication record submitted for the REF pilot exercise.

An outline specification for pilot HEI data was circulated in July 2008. For staff data, the RAE specification and definitions were used as a starting point. Additional (non-RAE) data were requested to enable a determination of the effects of varying the staff selection (and hence the collated output data). For output data, the RAE specifications were again used. Some additional fields were requested to help in matching outputs to citation databases. A comprehensive list of outputs (in addition to journal articles) was sought, to provide a context for benchmarking indicators and tracking publication behaviour.

A third necessary and central part of the data requirement for the REF pilot project was the association of output data with named staff for the author-model. Institutions were asked to provide a pair-wise association between staff and publication IDs. To ensure that sufficient links would be available for each pilot HEI, the project made use of the Symplectic Publications system to enable a comprehensive search for additional links.

---

[1] The ICT Implications Arising From the Research Excellence Framework Bibliometrics Pilot. Project Report. Stuart Bolton http://ie-repository.jisc.ac.uk/338/ and Identification and dissemination of lessons learned by institutions participating in the Research Excellence Framework (REF) bibliometrics pilot: Results of the Round One consultation, Technopolis http://hefce/pubs/rdreports/2009/rd09_09/

Six pilot HEIs indicated that their total REF submission would not be more than they submitted to the RAE, even if time were available. Several pilot HEIs indicated that they expected a roughly four to five-fold additional data submission compared to RAE2008. In every case, pilot HEI estimates were less than *Evidence*'s 'presumptive' estimate, on average by about 30%. In the outcome, most HEIs were able to extend their submission beyond solely RAE data.

Receipt and processing

Data development and collection was supported by regular contact between the pilot HEIs, the contractors and HEFCE. Because of the very compressed timetable set for HEFCE, the contractors agreed to accept pilot HEI data that fell outside the published specification and to clean this centrally. Additionally, some HEIs could not have submitted data to the specification required. This subsequently had serious consequences for resource capacity in later stages, but also provided valuable insight into the quality of institutional data systems.

Data were submitted via the HEFCE extranet in agreed formats (Excel, Access or xml). Some pilot HEIs made multiple, serial and overlapping data submissions, rather than single data submissions. The multiple data submissions included the following total records:

- 87,641 staff and other researcher records in versions of Table 1, of which 44,136 cleaned and deduplicated records were passed to the Symplectic Publications system;
- 678,077 article and other output records in versions of Table 2, of which 328,136 cleaned and deduplicated article records were passed to the Symplectic Publications system;
- 872,132 links between staff and authors in versions of Table 3, of which 433,447 properly indexed and deduplicated links were passed to the Symplectic Publications system.

This compares with the roughly 50,000 staff records and 200,000 output records handled within the RAE system (based on estimates from 2001 data; the indication is that 2008 was a somewhat but not significantly larger submission).

It became evident that the limit to staff data that could be provided by many pilot HEIs corresponded to the staff list submitted for RAE2008. Output data were also limited. Few institutions have in place a system for the regular submission of standard and comprehensive publication data or content by academic departments to any central database or repository.

Because there was a greater level of central data processing, cleaning and management than had originally been planned, the project became increasingly engaged with data management. The greatest impact on the project was in the speed of development of the core database. A second area of delay was in the processing of additional records from the presumptive data. Because of the underlying deficits in data quality, the task of linking authors with staff was also more onerous and complex than intended. The combined effect of delayed data handover between *Evidence* and Symplectic and the poor relative quality of the data at that point exacerbated the delay in offering enhanced data to pilot HEIs for verification.

Whereas it was originally intended that pilot HEIs should be offered supplementary data records in September 2008 and the opportunity to verify additional staff-author links through October and November, the outcome of resolving data issues delayed this into a compressed period during December and January. Some further data development continued into February 2009.

Important lessons have, hopefully, emerged. The central one is that most institutions will require a very clear and extended implementation pathway before the REF could be introduced on a national scale.

Management of staff data

The process of building the staff table was an iterative one of importing, reviewing, returning to pilot HEIs and amending records.

The quality of the data submissions was affected by two things. First, there was an enforced haste to supply information, which was then submitted in a form that, had more time been available, the pilot HEIs themselves would have corrected (for example, fields were confused or mislabelled). Second, fundamental deficits in pilot HEIs' systems meant that data could not be readily retrieved in a

particular format. Beyond data quality there was an issue of data deficits. There was also a variable outcome because of differences in the approach that pilot HEIs took to supplying data.

The most critical piece of data that was widely absent was any information about the prior employment record of staff currently employed at a pilot HEI. Although institutions hold such information, it has not previously been a part of normal electronic database records. For the REF, the significance of this is in identifying and examining output data prior to current employment. If this information is to be a standard part of analysis, then there will need to be a systematic and systemic change in the way it is captured.

Initially, all staff data were included. It was later decided to restrict subsequent analysis to RAE-eligible staff only. After initial data review, it was determined that staff who were ineligible for the RAE appeared to have relatively few publications that were not co-authored with an RAE-eligible member of staff. *Evidence* transferred 44,136 cleaned staff records to the Symplectic Publications system.

Management of output data

The integration process to create a single bibliographic database for matching and processing took longer than anticipated. This was because of data quality issues (both missing and erroneous data) and because more updating was required than had been expected. Twelve pilot HEIs serially submitted output data as many as six times. This was complicated by supplementary datasets, complete updates and partial replacements, not always in the same format as original data from the same institution.

The data request to pilot HEIs asked them to submit not just journal articles but also non-journal outputs, to throw light on the broader publication context.

- About 250,000 of 328,136 output records supplied by pilot HEIs appeared to be from research journals.
- On balance, the data suggest that articles and reviews probably account for 65-70% of significant outputs in the subject areas under examination.
- Of the residual records, about 25,000 appeared to be books or chapters and 40,000 to be conference contributions.
- Conference proceedings, which will soon be subject to much improved evaluation, account for 10-15% of significant outputs in the subject areas under examination.

This balance would allow the REF to explore academic impact for upwards of 80% of the potentially available material in these subject areas.

Cleaning regimes were applied to selected fields in the outputs database, concentrating on those essential for a satisfactory match to be made to commercial citation databases. This prioritised fields such as journal titles, volume and page numbers and unique identifiers including DOIs (digital object identifiers) and Thomson UTs (unique tags). The data were combined with the staff data and used as the subsequent outputs' dataset for verification in the Symplectic Publications system.

Reconciliation of journal outputs to the Web of Science

Evidence was responsible for reconciling the article records supplied by pilot HEIs to the article records in the Thomson Reuters Web of Science commercial database. This was verified by checks through the Symplectic Publications system. Reconciliation of pilot HEI data to Elsevier's Scopus database, an alternative source of commercial supply, was carried out by HEFCE. This, and comparison between the two, will be reported elsewhere.

For apparent article and review records, DOI data were available for 49% of outputs and gave an overall matching success of 28% or 70,147 outputs. Journal and article title data were available for 85% of outputs and gave an overall matching success of 62% or 155,986 outputs. Journal title, volume and pagination data were available for 78% of outputs and gave an overall matching success of 61% or 152,440 outputs.

Issues arising from data gathering and processing

Many issues that arose during the REF pilot exercise are much less likely to arise during a full-scale national implementation. Nonetheless, the problems that did occur will have to be taken into account. Many of them reflect the fact that, currently, most institutions are not readily able to supply the data that would be required. This constraint is not limited to any particular group or to any particular type of data.

Owing to the constrained timetable of the pilot project, three stages were running concurrently. Running them consecutively would have increased efficiency and made it less onerous to track and trace data and to modify the design, but would have elongated the timetable by months.

Pilot HEIs were permitted to make successive submissions of data. This was intended to assist them to keep to a tight timetable and to allow us flexibility to respond to different levels of data availability among pilot HEIs. With hindsight, the cost is stark, because multiple submissions increased the central workload disproportionately. The benefits of allowing multiple submissions, however, were: iterative development of data processing and cleaning techniques; flexibility to vary the requirements according to individual pilot HEIs' status; and an opportunity to brief pilot HEI staff on working aspects of the relevant data.

Extension of output data and disambiguation of author and staff names

Symplectic, a subcontractor to *Evidence*, focused its work around two major tasks: first, pilot HEIs' publication data were reconciled to Thomson Reuters Web of Science data in an automated fashion to produce a single, inclusive pool of publications; second, records were then matched with the academic staff lists.

Data records were classified into three categories: output with a Thomson Reuters record alone; output with an HEI record matched to a Thomson Reuters record; output with an HEI record alone. To maximise the linkage between staff and the publications data, an automated mechanism was needed to suggest potential links. Automated methodology suffers from two major drawbacks: first, the risk of identifying false positive matches, suggesting that authors have written more papers than is the case; second, the risk of missing matches, thus failing to suggest the authors of papers in the dataset.

A simple algorithm was devised to match outputs with their authors. This relied on matching institutionally supplied names and variations with 'searchable data' restricted to the portion of the article database associated with each staff member's home institution. The links supplied by each pilot HEI were then applied over these data to form a firm link of "approved" articles.

Any suggested link from the automated mechanism was a "pending" link, reviewed by pilot HEIs through a customised web interface or a downloadable spreadsheet. It became clear that a sampling strategy would be required where strategic approval methodology could be applied, in order to ensure maintenance of the data quality and to understand weak points. Several methodologies were applied to institutional data in order to help institutions with larger amounts of "pending".

Creation of database

The outputs of the Symplectic Publications system were recreated forms of the key REF data tables containing deduplicated staff information (Table 1); extended, cleaned and deduplicated publication records (Table 2); and more comprehensive links between staff and authors (Table 3). The data records and links processed by Symplectic Publications and accepted by the pilot HEIs were resubmitted to the secure server.

The final steps in creating the bibliographic database required for the REF pilot project were the association with the validated publication records of their relevant citations data and the normalisation of the citations data to enable comparative analyses. A later report will describe the development of the combined publication and citation database and the decisions made regarding normalisation.

1 A pilot study of bibliometric indicators of research quality

1 In June 2008, the Higher Education Funding Council for England (HEFCE) commissioned *Evidence* Ltd to initiate a project to develop and analyse a pilot bibliometric database to evaluate the feasibility of introducing quantitative indicators to its research assessment methodology.

This report

2 This report covers the work required to address the development of an initial bibliographic database from data supplied by a group of pilot higher education institutions (pilot HEIs). This is Task A, the first of three principal tasks within the overall project plan. Later reports will discuss subsequent work phases: Task B on collation and normalisation of the citation data so as to develop the bibliographic database (outputs only) into a bibliometric database (outputs with citation counts and indicators); and Task C on variant analyses.

3 Twenty-two pilot HEIs agreed to participate in and support the work reported here. What is described has been carried out in close liaison with those institutions, and the contractors wish to acknowledge at the outset the very considerable workload that has been accepted and absorbed by the pilot HEIs.

4 The contractors are particularly grateful to the many individual staff who worked with us for their untiring efforts to enable this work to progress. Without their commitment it is not feasible that so much progress could have been made in the time available, and nor could we have learnt so much about the issues that will need to be constructively addressed in any subsequent implementation.

Background

5 In 2006, the UK Treasury proposed that core research funds might be distributed to universities via 'metrics' rather than peer review. The (then) Department for Education and Skills and HEFCE carried out work to explore this proposal and develop a sound implementation pathway that accorded with system requirements.

6 In 2007, HEFCE commissioned various preparatory studies to consider the extent to which bibliometrics (the study and analysis of publications and citations) would prove a sufficient and appropriate source of indicators for purpose. This enabled a further round of consultation with the academic community on this Research Excellence Framework (REF) (see HEFCE circular letter 13/2008), following which HEFCE and the Secretary of State for Innovation, Universities and Skills announced further modifications to the broad structure of the proposed metrics process, but confirmed that it would be introduced after the 2008 Research Assessment Exercise (RAE2008).

7 This report is the first on a study to develop a bibliometric indicator for the REF, working with a small group of institutions to pilot a route for the higher education (HE) system. The steps in this work are the assembly of a database of published outputs, the development of an appropriate citation database for the articles and reviews among those outputs, and the analysis of the citation indices derived from those data.

8 Unless otherwise indicated, in this report the metrics discussed in reference to the REF are restricted to 'bibliometrics', which are the indicators created by an analysis of research journal articles and their subsequent citations. These are not the only research activity variables that would need to be utilised in the REF, but the additional input and activity measures are being developed separately in-house by the Funding Councils.

2 Aims of the pilot

9 In order to develop a bibliometric indicator that will command the confidence of the academic community, HEFCE committed itself to running a pilot exercise for which the aims were to:

- test processes and methods for generating and producing bibliometric indicators of research quality;
- investigate how bibliometric indicators can be used within the REF in a way that is robust and fit for purpose;
- identify the operational implications and costs of implementing the process (for HEFCE and pilot HEIs);
- inform an assessment of the REF's potential impact on the sector;
- assist pilot HEIs in understanding the implications of the bibliometrics process.

10 This report does not address the potential utility of bibliometrics. That was considered in earlier work commissioned by HEFCE ("Scoping study on the use of bibliometric analysis to measure the quality of research in UK higher education pilot HEIs" (CWTS, University of Leiden: 2007; HEFCE R&D Report No 18)) and will be looked at later in the light of the outcomes of the work reported here. However, among the challenges identified by HEFCE are the need to:

- explore the range of subjects to which bibliometric indicators should be applied, i.e. where such indicators would be consonant with other indicators of research quality;
- determine whether staff coverage should be restricted to 'principal investigators' or should include a wider cohort contributing to research, such as research assistants, postgraduates and technicians;
- decide whether research should be credited to author or institution at the time of publication;
- determine which categories of papers should be included in the REF;
- evaluate whether the inclusion of publications should be universal, to make a comprehensive assessment of research outputs, or selective as it is under the RAE;
- develop a process for assembling, checking and validating data on eligible staff and papers and, within this, create an understanding of the implied workload for pilot HEIs;
- define appropriate bibliometric indicators, including single metrics such as average impact and integrated metrics such as a 'quality profile', and describe methods for constructing them;
- determine how the indicators would most effectively be used by expert panels;
- explore what other information should be made available to pilot HEIs.

11 The pilot work was also designed to compare two variant approaches:

- address-model – a low-burden model for data collection and analysis, with data collated by address and then allocated to subject areas via journal categories;
- author-model – a more onerous model in which each output would be associated with subject areas via an explicit author-staff link, but in which pilot HEIs and their key staff would need to both support and then validate the development of author-staff identification and disambiguation.

12 As the pilot work was to be developed within a relatively compressed timetable, the project made use of readily available data, expertise and methods and therefore did not test or resolve all the issues raised in earlier work. The aim was to provide evidence on which firmer proposals could be formulated on specific points, but in other areas only to illuminate options that require further work. The policy environment itself continued to develop during the period of the pilot work.

13 The contractors faced three key tasks:

Task A – The assembly and development of an appropriate bibliographic database, including the collection and structuring of full and accurate journal article and review records for the pilot HEIs.

Task B – The collation and development of citation data and benchmarks for the article records assembled in Task A.

Task C – The development of variant analyses of the publication and citation data (for example, collated around authors or collated around addresses) to explore the effects of variant data combinations and to work towards a standard analytical methodology and a set of standard research performance indicators suitable for inclusion in a national REF methodology.

14 A critical first aim was to assemble a database fit for purpose and without which further development would be infeasible. The assembly of this database and the exploration of its characteristics is the subject of this report.

3 Prior assumptions and planning

15 The pilot work was launched in June 2008. It was expected that Task A, the development of the bibliographic database, would take up to six months.

16 As an address-model database had already been created by *Evidence* for each pilot HEI, the aim was to focus on the author-model and to collect, process, clean and structure institutional staff and output data by September, and then by December to assemble this into a cleaned and functional bibliographic database with institutional output records linked to named staff. This development would take place in parallel with Task B, which was the association with the publication data of the relevant data on citations, both for the counts per article and for the associated global benchmarks. Finally, with the publication and citation database structured and complete, the plan was to initiate Task C analyses on the bibliometric data early in 2009.

17 This was a challenging timetable, both for the pilot HEIs involved in pilot work and for the contractors. It is useful to put that challenge in the context of the recent RAE.

18 Response to the RAE2008 specification was preceded by a long period of discussion and preparation. Nonetheless, despite this anticipatory period, some HEIs were still left with problems over their final delivery in late 2007. Given this recent experience, it can be assumed that requesting a response to a new REF pilot specification in weeks rather than years would provoke some resistance. However, such a request and timetable would also identify areas where HEI systems were not (yet) designed to respond. To mitigate this, it was agreed that the pilot REF data specification should match the RAE2008 data collection as closely as possible. Where that mitigation was insufficient, and data gaps appeared, the exercise would reveal the challenge to implementation that HEFCE would need to take into account in a full national exercise in terms of the timetable to systematic implementation and the data specification and management strategies that would be required.

19 Address-based data were available immediately. HEFCE's preferred approach to pilot data collection for the author-based model was that:

- the pilot HEIs provide up-front data on relevant staff and papers, building on existing data collected internally in preparation for RAE2008 and existing institutional bibliographic databases;
- the contractors match HEI data to the citation database.

20 HEFCE did not expect data verification beyond the pilot HEIs' agreement to the inclusion of their datasets of matched publications. It was accepted that this might result in some level of inaccuracy or incompleteness, but that a full and complete verification would be unduly onerous for the purpose and perhaps infeasible in the time available.

21 Note that the coverage by HEI was determined by HEFCE in a separate exercise, and that the coverage by Unit of Assessment (UoA) was also determined by HEFCE after consideration of subject relevance and coverage in citation databases.

22 In initial discussions between the contractors and HEFCE a number of issues were considered around the strategy for data collection, such as whether it was better to aim for an initially comprehensive data collection, which would have some redundancy, or a selective data collection, which might need to be expanded. It was concluded that it was better to err on the comprehensive side, as repeat requests would slow down the project, irritate the supporting pilot HEIs and add extra work at all stages of data development. It was more efficient to process as much data as possible in one tranche.

23 It was essential at the outset to specify in precise terms: (1) the categories of staff and papers for which the pilot HEIs would be expected to provide data, and (2) the relevant fields of data (building on the approach set out in HEFCE's prior survey document). This was essential both to avoid confusion and ambiguity and to enable proper subsequent comparative analyses. In practice, experience revealed that the data specification would have benefited from more testing for ambiguities and assumptions. Furthermore, not all pilot HEIs were in a position to provide

24 None of the problems discussed in this report should be taken to be a criticism of the pilot HEIs or of the hard work of their staff in seeking to respond to the data requirements. The reason for reporting these issues is the need to alert stakeholders to the investment that will be required to prepare the HE research base to respond to a REF-type exercise. The detail, on which we enlarge below, varies greatly among institutions; the key point is the general challenge in responding in the near future, fully and accurately, to a comprehensive and verifiable data request that supports the purposes of the HE Funding Councils.

Planned outputs from Task A

25 The aim was to produce a database of publications, in total and for each pilot HEI, reconciled to commercial-source article records. For each article record, the following elements needed to be created to enable development from the address-model data to complete the author-model:

- cleaned and completed versions of original records from pilot HEIs;
- extension of submitted pilot data with additional presumptive data (see below for explanation of this);
- unique IDs from one or more commercial databases for all matched items;
- categorisation and summary analysis of non-matched items;
- from a collated dataset, the development of IDs to link original and presumptive article records to:
    - the named HEI
    - the author's unique identification, for example via Higher Education Statistics Agency (HESA) staff ID
    - the author's UoA, department or other subject identification;
- validation by HEIs of additional records and author-staff links.

26 The latter tasks (linkage and validation mechanisms) were the subject of a subcontract to Symplectic Ltd, which is reported in a later section. Symplectic has been working with a number of research-intensive pilot HEIs on the challenge of reconciling author names in publications to staff names on human resources (HR) records. Note that there will be a many-to-many matching: each staff ID will be linked to multiple outputs; most outputs will have multiple authors and hence linked HESA IDs and subject IDs.

4 The pilot HEIs and subjects

Pilot HEIs

27 Twenty-two HEIs participated in the pilot REF exercise. The list was determined by a prior exercise conducted by the HE Funding Councils. Interested HEIs responded to a survey sent out by HEFCE; participants were selected from these to provide a range of types of pilot HEIs across the country and to ensure coverage of the selected UoAs (see below). Pilot HEIs with a wide range of research management systems and processes (as indicated in their survey returns) were deliberately selected, as the Funding Council recognised varying ability to comply with likely data requests.

28 Reference to the web-sites of the pilot HEIs listed in Table 4.1 will show the widespread diversity they represent by region, size, history and subject portfolio. They provide an excellent challenge for a prototype national evaluation system to encompass.

Table 4.1 The 22 pilot HEIs

Pilot HEI name	Abbreviation in this report
Bangor University	Bangor
University of Bath	Bath
Queens University, Belfast	Queens
University of Birmingham	Birmingham
Bournemouth University	Bournemouth
University of Cambridge	Cambridge
Durham University	Durham
University of East Anglia	UEA
University of Glasgow	Glasgow
Imperial College London	Imperial
Institute of Cancer Research	ICR
University of Leeds	Leeds
London School of Hygiene & Tropical Medicine	LSHTM
University of Nottingham	Nottingham
University of Plymouth	Plymouth
University of Portsmouth	Portsmouth
Robert Gordon University	Robert Gordon
Royal Veterinary College	Royal Vet
University of Southampton	Southampton
University of Stirling	Stirling
University of Sussex	Sussex
University College London	UCL

Units of Assessment included in the REF bibliometrics pilot

29 The subject (discipline, field) structure for the REF has yet to be decided. For the pilot project, it was agreed that the categorical structure for data collection and analysis would follow the well-understood subject categories set out within the Units of Assessment used for the RAE. The UoAs selected by HEFCE were those which had 40% or greater coverage in principal commercial data sources (either Web of Science or Scopus) as determined by earlier work done by CWTS ("Scoping study on the use of bibliometric analysis to measure the quality of research in UK higher education pilot HEIs", CWTS, University of Leiden: 2007; HEFCE R&D Report No 18). This list is at www.hefce.ac.uk/research/ref/pilot/datacoll/

Table 4.2 The Units of Assessment used as subject areas for the REF pilot project

UoA reference	Discipline name
1	Cardiovascular Medicine
2	Cancer Studies
3	Infection and Immunology
4	Other Hospital Based Clinical Subjects
5	Other Laboratory Based Clinical Subjects
6	Epidemiology and Public Health
7	Health Services Research
8	Primary Care and Other Community Based Clinical Subjects
9	Psychiatry, Neuroscience and Clinical Psychology
10	Dentistry
11	Nursing and Midwifery
12	Allied Health Professions and Studies
13	Pharmacy
14	Biological Sciences
15	Pre-clinical and Human Biological Sciences
16	Agriculture, Veterinary and Food Sciences
17	Earth Systems and Environmental Sciences
18	Chemistry
19	Physics
20	Pure Mathematics
21	Applied Mathematics
22	Statistics and Operational Research
23	Computer Science and Informatics
24	Electrical and Electronic Engineering
25	General Engineering and Mineral & Mining Engineering
26	Chemical Engineering
27	Civil Engineering
28	Mechanical, Aeronautical and Manufacturing Engineering
29	Metallurgy and Materials
32	Geography and Environmental Studies
34	Economics and Econometrics
40	Social Work and Social Policy & Administration
43	Development Studies
44	Psychology
46	Sports-related Studies

HEI/UoA differences

30 Not all pilot HEIs elected to supply data for all UoAs. Some UoAs are covered by data from most pilot HEIs, but others from only a handful. The most frequently submitted UoA was 23 Computer Science (18 of 22 pilots), followed by 14 Biological Sciences (17 pilots). Other examples are 17 Earth Systems and Environmental Sciences, which was covered by 13 pilots, while 19 Physics was covered by 12 and 28 Mechanical, Aeronautical and Manufacturing Engineering was covered by only 11.

31 The diversity of UoAs covered by a single institution varied from just one (the Royal Veterinary College) to more than 20. Note that UoA coverage does not necessarily represent the span of

5 Data request to pilot HEIs

Actual and presumptive data

32 There are a number of potential sources of data on research publications, including: structured, usually selective, commercial databases; selective and structured academic databases developed by learned societies; and less structured but often comprehensive data accessible through web browsers, often drawing on institutional repositories as well as publisher information. Because the commercial databases are standardised and have a well-understood structure and coverage, they usually provide the most appropriate 'match' for a formal and repeatable evaluation exercise covering a wide subject range. They also contain the indexed and collated reference links used to create citation counts and indices.

What are 'presumed' data?

33 *Evidence* has been able, from extensive prior work with UK HEIs, to collate article records from each HEI that map to Thomson Reuters' indexed journal articles. This work therefore provided a potential 'presumptive' dataset to compare with and if necessary substitute for what each HEI was in practice able to provide. 'Actual' data supplied by pilot HEIs was unlikely to be the most complete dataset for the institution, even if the collection was restricted to the commercial sources. This is because many institutional publication databases are in the process of development and because researchers can be selective in depositing their material in such databases or repositories.

34 *Evidence* annually processes all of the article records from Thomson Reuters Web of Science® that contain at least one address with a UK element (including England, Northern Ireland, Scotland and Wales). This process means trawling through about 100,000 article records annually, some with several UK addresses for the authors. Each of these 'author addresses' for the period back to 1981 has been manually reconciled to a real 'organisational address' verified by checking the address, and usually tracking the author name, via the Internet. It is often a surprise that academic researchers might give only their departmental or group address, but for some universities this reconciliation work has enabled *Evidence* to increase the 'raw' tally based on main institutional address by as much as 40%.

35 The consequence of this prior work is that a detailed set of variant addresses for each pilot HEI is available, and these can be used to work up an organisational set of article records presumed to be linked to the institution because of those address elements. These 'presumptive' databases can be compared with the 'actual' submitted data from the same pilot HEI.

36 The presumptive data alone could form the database required for the address-model variant. This is potentially low-burden because, without requiring any further work by the pilot HEIs, the data can be collected and assigned to subjects via a mapping of journal categories to subject areas.

What are 'actual' data?

37 The actual data are those article records already collected by institutions from their staff and therefore explicitly validated as part of the publication record submitted for the REF pilot exercise.

38 For the publication data collected from each pilot HEI to be formally analysed, in terms of either their relative coverage or their citation performance, it is necessary to match original researcher publication records to a common, formal and standardised source. The diverse output data actually supplied by pilot HEIs needed first to be reduced to potential journal article records and then formally matched to other standard records. Not all 'actual' records could be verified for the REF pilot.

Data collection specification

39 An outline specification for key data from pilot HEIs was drawn up in June 2008 (see Annex B). This was discussed with HEFCE on July 1st and then presented to pilot HEIs at a briefing meeting on July 9th. Pilot HEIs were invited to comment on the outline and indicate where they

40 To enable the pilot HEIs to complete the task, the specification followed closely the model set by the returns for the RAE2008 since it was assumed that all pilot HEIs would have data that would accord with this profile. Some additional fields not required by the RAE were requested, however, in order to address specific aspects of the REF.

Specification for staff data (Table 1)

41 An early proposal to minimise the burden of data collection on pilot HEIs by sourcing staff data from the Higher Education Statistics Agency was not pursued, as the problem of linking named staff to specific UoAs was seen to be a critical challenge and names are not held by HESA. Furthermore, the scope of staff coverage was also greater than that which is normally accessible.

42 It was therefore necessary to solicit data from the pilot HEIs. It was assumed that a data format which followed the RAE2008 submission would prove amenable and, as a minimum, that such a format would elicit a complete and accurate tally for staff submitted. The aim was then to extend the actual REF submission as far as possible so as to place this 'submitted staff' cohort into a broader context of potential authors of research publications.

43 The RAE specification and definitions were therefore used as an initial starting point to collect the data. From this, the contractors and HEFCE worked towards a final data specification, with additional fields and a greater number of staff types. The final specification is provided as Table A1 in Annex A.

Fields additional to the RAE submission

44 Additional (non-RAE) data were requested to help in creating variant databases that might determine the effects of varying the staff selection (and hence the collated output data) on the values of indicators. It would be valuable to know, for example:

- whether potential authors might be early-career researchers whose publications would be fewer and less prominent than established staff;
- when staff had left or joined pilot HEIs and which institutions joiners had come from, so that the scope of the analytical data could be varied by name and by address.

Staff additional to the RAE submission

45 The reasons why we sought to collect data on the widest possible range of staff were two-fold:

- To acquire the most complete list of the potential publishing population so as to disambiguate the greatest number of potential authors on papers associated with each institution. To do this we needed to know not only about authors on the permanent staff cohort but also about researchers, research students, technical staff and other groups who might contribute to research publications.
- To explore the consequences of reducing the indicative coverage by making more selective assessments using only particular staff groups. If we collected data on only the select 'RAE submitted' staff group, such a variant comparison would clearly be infeasible.

Specification for output data (Table 2)

46 Prior work by HEFCE's contractors had confirmed that the potential bibliometric analyses which would lead to useful research performance indicators for the purposes of the REF were likely to focus around the conventional indicator set used in the many research evaluation studies which exploit data on, first, research journal articles and, second, on their citations.

47 The RAE specification and definitions were used as an initial starting point to collect the data pertaining to outputs published by members of staff in the selected pilot UoAs (see Table A2 in Annex A). Some additional data were requested (fields and definitions in italics) to help in matching the outputs to citation databases.

48 All information on outputs was retained, both to inform HEFCE of the ability of HEIs to report on their outputs and to indicate the extent to which indicators would be inclusive of their overall portfolio.

Output records additional to journal articles

49 Although REF analyses would be constrained to outputs published in journals, or more specifically to articles and reviews from journals catalogued on commercial databases, we sought the most comprehensive list of outputs from pilot HEIs to create a context for the indicators. This context is useful for current policy debates and is also needed for benchmarks of publication behaviour and to plot potential areas for development.

50 From a policy viewpoint, if the REF indicators give only a partial view of research performance then it is desirable that observers should be aware of how partial that view might be. Only if we know the due proportions of each institution's output that are books, chapters, journal material, conference proceedings, patents, reports or other work will we be able to come to a view about this.

51 It is valuable for institutions to work towards the most complete database or repository for their internal research management purposes. It would then be feasible to report this accurately and in a structured way, by research area and output type, in a REF submission.

52 For future development, a consideration is the dynamic nature of the available reference data. For example, it will likely soon be possible to broaden the range of indicators from journal-based analyses to include conference proceedings. Later, it may become feasible to extend analyses to other works, including books and the so-called 'grey' literature produced in public policy reports. Some benchmark of the latent scope for the REF would therefore be valuable to policy makers and planners.

Comment on staff-author links (Table 3)

53 Institutions were asked to provide a third table, showing the pair-wise association between the unique record identifiers from their HR systems (i.e. the staff ID) and the unique record identifiers from their output databases (i.e. the article ID, hence linking to the list of authors). Since each staff member is likely to have authored several papers and many papers are likely to have multiple institutional authors, pair-wise linking is essential to ready matching.

54 A necessary and central part of the proposed analyses for the REF is that it should be possible to associate output data with chosen disciplines within the institutional structure. For the author-model, this is achieved by linking outputs with staff (and staff categories) in these areas. For the final REF format, both the disciplinary structure and the staff coverage have still to be determined, but any model system would need to be able to respond flexibly to both requirements.

55 Publications bear author names. Unfortunately, it has not always been possible immediately to associate an author with a member of staff. This is because author names are not explicitly linked to an address, and many articles are co-authored by staff from several institutions. The author-address linkage deficit is now being remedied by both of the principal commercial data suppliers.

56 A further problem is that of synonyms, so that even within an institution there is ambiguity about assignment; this constitutes the biggest problem with current data. While it may be evident that a paper from Uttoxeter University is authored by J Smith, that institution employs more than one J Smith and they work in different areas. There is therefore a need to disambiguate these names so as to ensure a comprehensive mapping system, in the case of the REF pilot between the data in Table 1 (staff names) and in Table 2 (author names).

57 Data processing applied by Symplectic, as a sub-contractor to *Evidence*, was a critical part of the project planning in regard to author-staff disambiguation. We anticipated that:

- Many institutions would only be able to supply partial linkage data in Table 3, and we would therefore need to work with them to complete the linkage maps.

- We would be adding new article records, for which no prior institutional linkage existed, from the 'presumptive' data for each institution; we would have relevant staff information from Table 1 to identify the authors of these articles, but would need to search to find possible links and then validate these with the pilot HEIs.

58 The Symplectic Publications management system is one that has been developed to enable a comprehensive search for additional author-staff links. In the context of this project that would enable us to collect and present putative links to pilot HEIs for validation or rejection.

Initial response from pilots

59 At the first pilot briefing meeting, in the early summer of 2008, pilot HEIs confirmed anticipated concerns about the volume of data required, the specification – including the range of fields for each record – and the timetable for delivery.

60 Pilot HEIs agreed to respond as best they could, but quite reasonably emphasised the constraints they would face during July-September (the HE vacation period) and in assembling the data required. Their view was that collecting a comprehensive dataset was unduly onerous given, from an external perspective, that the additional analyses this would enable appeared to be of marginal value.

61 However, few institutions immediately saw any issues with the data specification. No immediate concerns were expressed by pilot HEIs about data quality or data format. Later, it became apparent that there were in practice some serious issues with the data received. This probably reflected differences in interpretation and a lack of clarity in the data definitions, despite a substantial number of calls and emails to clarify that specification.

62 Staff data (Table 1) were by and large felt to be uncomplicated. Some institutions warned us that they would not be able to provide specific parts of the data. For example, the 'prior institution' and 'destination institution' information was frequently cited as data that were not held in a readily accessible format. Information on Category B staff (eligible staff who had left prior to the RAE census date) was also limited.

63 It became clear quite quickly that some institutions would find the extended bibliographic data (Table 2) more difficult to produce, because it was not held centrally and because of academic staff absence.

Initial estimates of volume

64 An early circular asked the pilot HEIs for an initial estimate of the volume of staff and publication records. To this end, they were supplied with a form prior to the July briefing meeting containing the questions in annex C.

65 Some pilot HEIs were not immediately able to make an estimate of the total data volume. This was because they were aware that additional holdings beyond the RAE submission were available within departments, but were unsure how accessible these would be or what quality of data might be discovered. This indicates that some institutional databases remain dispersed rather than being managed centrally.

66 Six pilot HEIs indicated that their total REF submission would not be more than they had previously submitted to the RAE, even if more time were available for collection of information. This indicates that some institutional databases probably do not extend beyond the RAE requirement.

67 Several pilot HEIs indicated that they expected a roughly four to five-fold additional data submission compared to RAE2008. One pilot HEI estimated that it held about 10 times as much data as were submitted to RAE2008. This indicates that at least some institutions do have substantial, perhaps comprehensive, centrally managed databases that are likely to contain a diversity of output types extending well beyond journal articles. However, these must be seen as pathfinders rather than typical examples.

68 In every case, pilot HEIs' estimates of the number of journal article records they held were less than *Evidence*'s 'presumptive' estimate, on average by about 30%. This confirms the indications that most institutions do not yet have comprehensive collections.

69 The presumptive data cover all subject areas, whereas the pilot HEIs submitted only selected UoAs. However, the REF UoAs are explicitly those where journal outputs are the predominant mode of publication, and hence they are the subject areas that make up the bulk of the institution's publication record. This suggests that the presumptive data were likely to add additional information not presently held by the pilot HEIs.

70 In the event, most HEIs were able to extend their submission beyond solely RAE data. During the summer period they were able to collect substantial additional information from the academic authors and assemble this for their REF submission.

71 The work carried out by pilot HEI staff during this vacation period was extremely helpful to the later analyses and much improved the contextualisation of the journal article data.

6 Receipt and processing of data from pilot HEIs

72 The original timetable from HEFCE, imposed by the policy timetable, meant that data collection needed to start as soon as feasible. All parties accepted that this was far from ideal since it meant that much work occurred during the summer vacation.

73 Within the limits of the timetable, flexibility in data submission became necessary to enable pilot HEIs to supply the data that were available. To achieve this, when the available records could not meet the original quality expectation, it was decided that data cleaning and processing would be carried out centrally. The value to the pilot process of accepting and then analysing the deficits across the data submissions is that HEFCE now has an insight into the quality of the data that pilot HEIs hold. For the contractors, however, this had a subsequent impact on workload and thus on resources.

74 This does not imply 'fault' on the part of the pilot HEIs. They had little time to respond to the specification and staff were thin on the ground. It is not surprising that what was received was of variable quality. For HEFCE, this will be valuable planning information: the constraints and the diversity of what a representative sample of UK HEIs would be able to supply at short notice is now much clearer.

Data management

75 Pilot HEIs were asked to supply data across the three key tables (Table 1: staff, Table 2: outputs, Table 3: links between these): first, by the beginning of August, all of Table 1 and an initial Table 2; second, by the end of August, any additions to Table 2 plus Table 3.

76 Pilot HEIs were permitted to supply the data in a range of formats (for example, Excel, Access or xml), via the HEFCE extranet.

77 A total of 85 data submissions were made by the 22 pilot HEIs. In total, the data submissions included:

- 87,641 staff and other researcher records in versions of Table 1, of which 44,136 cleaned and deduplicated records were passed to the Symplectic Publications system;
- 678,077 article and other output records in versions of Table 2, of which 328,136 cleaned and deduplicated article records were passed to the Symplectic Publications system;
- 872,132 links between staff and authors in versions of Table 3, of which 433,447 properly indexed and deduplicated links were passed to the Symplectic Publications system.

78 This compares with the roughly 50,000 staff records and 200,000 output records handled within the RAE system (based on estimates from 2001 data; the indication is that 2008 was a somewhat but not significantly larger submission).

79 We planned that, by the end of September, pilot HEIs would:

- check the additional indicative outputs identified from a search of the Thomson Reuters database;
- add links to staff from authors of the newly identified additional indicative outputs.

Interactions with HEIs

80 During the initial period after provision of the specification there was active and frequent interaction between the key staff at the pilot HEIs and the contractors. This particularly involved issues regarding: the scope of the data collection, which as noted was wider than most pilot HEIs had expected; the detailed fields in the specification, which were more numerous and specific than the pilot HEIs had expected; and the collection timetable, which was problematic by both its brevity and its timing in the summer vacation.

81 Pilot HEIs were extremely cooperative and supportive. Because of the seasonal timing, not all staff were available throughout the period and this gave rise to some continuity issues. Similarly

82 The role of HEFCE as an active participant (rather than simply a customer) was a valuable and, indeed, essential element of the process. Ultimately, the Funding Councils will manage the entire REF process from data specification through collection and processing to the analytical evaluation of research performance. This can only be implemented through their understanding of opportunities and threats, which comes through direct involvement rather than observation. The contractors were also grateful to the close attention given by HEFCE's team to the operation of the pilot process. Regular video-conferences maintained close communication, supplemented by frequent face-to-face meetings.

83 An email support system (REFPilotSupport) was established via HEFCE. This allowed pilot HEIs to raise questions about their data in a fashion that should have enabled a timely response. However, such queries frequently meant that data had first to be checked in order to address the enquiry and provide a response, a solution or a further request. This was frequently a time-consuming process in itself, and many responses were necessarily of a sensitive nature which required thoughtful handling.

84 REFPilotSupport mailboxes were set up early as a point of contact that allowed all partners to see issues reported by pilot HEIs, and responses. In practice, the common route was frequently by-passed and there was also significant traffic to individual mailboxes and via telephone, both from pilot HEIs with questions and from staff in *Evidence* seeking clarification or chasing data. Other traffic to REFPilotSupport included matters of a technical nature, for instance regarding the provision of extranet keys, usernames and passwords.

85 Many of the early issues were straightforward and easily addressed, often relating to the specification. Once some data had been supplied, the focus moved to checking between multiple data files, provided at different times, and processing to bring them to the standard specification while correctly retaining the institutional information. Even with only 22 institutions it was challenging to keep track of the process as it related to individual pilot HEIs.

86 It would have been inappropriate not to have recognised the burden that was being imposed upon a relatively small group of key HE staff. The interactions between the contractors and the pilot HEIs were therefore at times necessarily informal and iterative. We feel that this was a valuable aspect of the work which helped to achieve a satisfactory outcome.

Review of actual timetable

87 When data first began to be submitted, issues quickly emerged regarding file formats, the completeness of the data, the number and structure of the fields supplied, and the specification and accuracy of the data in the fields.

88 With hindsight, it might have been preferable to have insisted on the original data standards. If this had been done, it would have distributed the workload back to the pilot HEIs. It might then have been difficult for some HEIs readily to have met the challenge, since they had already endeavoured to supply the best available data. That challenge would have impacted on the timetable and significantly reduced the final data pool. It would certainly have raised some additional concerns among managers at the pilot HEIs.

89 The choice for the project was either to reject data that did not meet the original specification, sending it back to the pilot HEIs for revision, or to handle the data as received. Since the contractors had experience in managing such datasets, they were well positioned rapidly to identify management solutions. An early decision to accept problematic data was therefore a pragmatic solution, to employ prior experience to improve data quality on a systematic basis. The principal driver for taking on this task, however, was the external timetable.

Consequences of additional data processing

90 Because of the greater level of central data processing and cleaning than had originally been planned, the project became increasingly engaged with data management. This affected

91 An interim consequence was that some pilot HEIs went without contact or feedback for some time. Responses to pilot HEIs were often slower during the middle part of the data development period than would have been optimal, because of the need to check each indicative response with all parties, on e.g. the data specification, the xml schema, access to the Symplectic Publications system, and the guidelines on data checking.

92 As academic staff returned from summer study visits, so the possibility of additional data submission, modification and revision emerged. Thus, at the same time (September) as the contractors were seeking to absorb a larger and more complex data-management task than originally planned, so the pilot HEIs were also encountering additional data. In some cases this led to permissions for further submissions, but in other cases it led to requests to delay submissions so that the pilot HEI could absorb the newly available material.

93 The 'learning environment' in which pilot HEIs, contractors and HEFCE were evolving their view of objectives and practice was a further reason not to apply and enforce a rigid and inflexible technical specification. The full REF implementation schedule will no doubt now be able to take account of the time that will be required for institutions to respond.

Use of a secure server

94 The project made use of a secure server at a remote location to increase protection of pilot HEI data, maintain accessibility for all parties, ensure that a common dataset was used, and ensure that data collected solely for the use of this project could be identified and deleted when the work was complete.

95 It was desirable to store institutional data on a server that was not owned by or physically situated in the consultants' offices. It was also necessary, given the potential sensitivity of the data and analyses, to put in place security protocols that the pilot HEIs would have found difficult to implement locally.

96 The clear advantage of this approach was that the project required a secure route to sharing data between HEFCE, *Evidence*, Symplectic and the pilot HEIs. The secure server meant that the main datasets were permanently accessible to HEFCE as the consultants worked on them. This also enabled ready access by pilot HEIs when they were invited to verify additional article records and author-staff links.

Generic data processing

97 The details of data processing and issues arising from the decision to manage data quality centrally are indicated in Annex D.

7 Management of staff data – Table 1

98 Data specifications were developed jointly by *Evidence* and HEFCE to meet proposed analytical requirements and evaluate future system needs while at the same time seeking to minimise institutional burden. Initially, the prospect of using HESA staff data was explored, but it was decided that all data would need to be requested from pilot HEIs directly.

Summary of staff data processing

99 The process of building the combined staff table was an iterative one of importing, reviewing, returning to pilot HEIs and re-importing records. The data in Annex E show the staff record count by each pilot HEI following initial cleaning. The Annex extends this brief description and discusses the nature of and remedy for data deficits where they occurred.

100 Files provided by pilot HEIs were fitted into the standard Excel template if not supplied in that format. Where necessary, changes were applied to standardise the order of the fields in the pilot HEI data and to change field names to the standard specification. Excel files were converted to text files (.txt) and imported to an Access database. The data were then appended to a common, central table. Data were consolidated, checked for accuracy and completeness, and modified (cleaned) where necessary.

101 The central table was queried in Access to provide summary statistics. Required fields were extracted from the central table for subsequent use in the Symplectic Publications system, to enable matching between staff names in the HR databases supplied by the pilot HEIs and the author names on article records matched to Thomson Reuters (Table 2 expanded).

Issues emerging

102 The quality of the data submissions was affected by two things. First, the haste to supply information meant that it was often submitted in a form that, had more time been available, the pilot HEIs themselves would have corrected (for example, fields were confused or mislabelled). Second, pilot HEIs' systems could not readily retrieve data in the specified format, which may limit their internal management reporting.

103 Beyond data quality, there are some fields which would be required for a complete REF analysis that are at present not generally maintained electronically by institutions. For example, the location of any prior employment tends to be filed in hard copy.

104 Each field was assessed for data quality and steps were taken to resolve more serious issues:

- Some data issues were resolved in-house where they affected a single pilot HEI and involved data structure, for example:
  - extraction of the first name from the last name field
  - swapping the contents of fields which had been incorrectly filled (such as the RAESubmitted and HESAStaffID fields).
- Other data issues were resolved by requesting further information from pilot HEIs where they affected many institutions and/or could not be resolved by the contractors, for example:
  - missing HESA staff IDs
  - missing data on RAE eligibility.

105 The variable outcome was in part because of differences in the approach that pilot HEIs took to supplying data and because of differences in the availability of those data. This raises questions about what can and will have to be done in national implementation of an evaluation system. On the one hand there are issues about the different interpretations of eligibility which institutions might have, and how these can be reconciled to a commonly agreed standard. On the other hand, there are questions about the current state of institutional HR systems and their capacity for delivering data for external, perhaps also for internal, management purposes. That said, we also recognise that the submissions reflected what we asked for and, in order to move the process forwards within the timetable, what we were prepared to accept.

106 The most critical piece of data that was widely absent was any information about the prior employment record of staff currently employed at a pilot HEI. While such information is held by institutions, it has not previously been a part of normal electronic database records. For the REF, the significance of this is in identifying and examining output data prior to current employment. If this information is to be a standard part of analysis, then there will need to be a systematic and systemic change in the way it is captured.

Handover of staff data to Symplectic

107 The cleaned staff data were extracted from the Access database and transferred to Symplectic via the secure server, for use in creating additional staff-author links in the Symplectic Publications system. Initially, all staff were included. Later, it was decided to restrict subsequent analysis to RAE-eligible staff only (after initial data review, it was determined that staff who were ineligible for the RAE appeared to have relatively few publications that were not co-authored with an RAE-eligible member of staff).

108 *Evidence* transferred 44,136 cleaned staff records to the Symplectic Publications system. Once the data had been submitted to the Symplectic Publications system, it became a straightforward task to alter the staff list so as to vary which staff from Table 1 were included in analyses.

Lessons learned

109 Lessons emerging from this part of the project are collated in Chapter 10. Detailed analyses of data cleaning are set out in Annex E.

8 Management of output data – Table 2

110 Data were submitted via the HEFCE extranet in the agreed formats (Excel, Access or xml). Serial submissions of data were allowed, with two main stages: an initial table (currently available data on publications, including RAE2008 submitted outputs and other data available to the pilot HEI from work for that submission), growing into an extended database (additional submitted output records plus presumptive data).

111 In reality, this process took longer than anticipated, because of both missing and erroneous data and because it involved more updating than was expected. During the process, four pilot HEIs submitted data once, 10 pilot HEIs submitted two datasets and the remaining pilot HEIs serially submitted data as many as six times. This included supplementary datasets, some complete updates and some partial replacements. Additions were not always in the same format or with the same ID numbers as the original data.

Development and restriction of data

112 Pilot HEIs were asked to supply records for all output types (see Table F1, Annex F), so as to provide a background context for the REF. The dataset for matching outputs to citation data was restricted, however, to those outputs defined by the pilot HEIs as a journal article or output type 'D' (RAE definition). The dataset was restricted at an early stage since there was little point in cleaning the page numbers of conference proceedings or books when these outputs would not be included in quantitative analyses.

Article records as a proportion of outputs

113 We noted earlier in this report that the context for evaluating research performance using indicators based on journal publication and citation data is bound by the range of other outputs that a research-based organisation produces. We need to know what proportion of output is being incorporated in our indicators before we can consider their value to an overall assessment of research outcomes.

114 To this end, the original data request to pilot HEIs asked them to submit not just their records of journal articles, from which matches might be made to commercial citation databases, but also their records of non-journal outputs. From this information we might be able better to throw light on the broader context:

- Pilot HEIs submitted a total of 328,136 output records, some of which were labelled according to the RAE schema of output types, but many of which were labelled in more idiosyncratic ways where some guesses had to be made for reconciliation.
- About 250,600 records appeared to be from research journals:
  - 222,470 were labelled as output type 'D' (the RAE identifier for journal articles and reviews)
  - 26,468 were labelled as JOUR
  - 1,663 were labelled as Article
  - 953 were labelled as Journal Article and 76 as Review.
- Of the somewhat fewer than 80,000 residual output records:
  - 24,492 appeared to be books or chapters in books
  - 40,194 appeared to be conference contributions or proceedings.

115 The overall balance of output types is important and is summarised under main modes (Annex F, Table F1). There is some subjectivity in the assignment of some material to these headings, but doubt only applies to smaller parts of the data total, except in the case of one university where an exceptional frequency of 'other' material remains to be explored.

116 Articles and reviews in leading journals can be mapped to commercial databases, as we have done in this project. Conference proceedings can also be mapped with increasing success and

117 What do the data in Table F1 tell us about the likelihood that research evaluation using bibliometric analyses will be a useful exercise, in the sense that it will be a generally good reflection of research impact based on output analysis? Recall first that this pilot exercise focuses on those subject areas where journal articles are the predominant publication mode. This data collation has little to tell us about those subjects where books are a more important mode.

118 The table shows that some institutions have either selectively supplied journal data or have databases that contain little or no other data. Zeroes for books, patents and reports and zero or very low counts for conference proceedings are markers of this. Where there is a more diverse offering, it remains generally true that articles are almost certainly disproportionately represented. While books may also be presented as very significant markers of esteem, they are on the whole less likely in these scientific and technical fields to represent a major contribution to original research findings. They may, however, be essential syntheses: milestones that can influence the direction a field takes over many years. Their impact is then undeniably greater than all but a few single research articles, but these are the exceptions.

119 On balance, however, it seems reasonable to estimate that articles and reviews account for 65-70% of significant outputs in the subject areas under examination. Conference proceedings would account for another 10-15%. These proceedings will soon be subject to much improved evaluation through the same methodology as has been applied to journal material. Thus, together, journal and conference output records would allow the REF process to explore academic impact (in the sense of impact reflected in citation counts and indices) for upwards of 80% of the potentially available material.

120 This tally is unlikely to provide anything like a full assessment of the true range and diversity of what the pilot HEIs actually produced during the relevant period. It is not yet commonly the practice for institutions to maintain a full and accurate record of all outputs and, where they have approached this, the printed form is often considered to be more likely to be recorded than other formats. However, in the subject areas considered here, the printed form remains the most important mode of dissemination for research results.

121 These statistics suggest that, for these pilot HEIs and for subject UoAs for which data were gathered, records of journal-based outputs do indeed represent the most significant part (more than three-quarters) of published output. This should be an entirely satisfactory sample of output for these institutions and subjects.

Creating the main database

122 The overall outputs dataset contained 250,603 outputs of type ‘D’, as defined by the pilot HEIs, and was used without further processing as the source outputs data for the Symplectic Publications verification system. This has meant that the pilot HEIs will have had sight of some of the less common ‘errors’ in the outputs data which were not corrected, such as article titles in journal title fields and invalid volume and page number formats. This may be useful feedback.

123 The data received from pilot HEIs were not initially of sufficient quality to attempt assignment of citation data from any commercial databases. The data validation and manipulation following this stage took significantly longer than had been anticipated.

124 Although it was tempting to apply customised cleaning procedures to each institutional dataset before collating into the final database, a decision was made to apply cleaning as a standardised and comprehensive regime for the final dataset. In this way, all data were treated equally and without prejudice with regard to any other information about the submitting pilot HEIs. It must be considered, however, that a ‘common’ procedure may have led to other bias. It necessarily meant that work concentrated on the most frequent errors. Errors which were locally more frequent in a small dataset (typically from smaller pilot HEIs) could still be less frequent overall than those from the larger datasets. Such errors may therefore not have been fully registered and rectified.

125 Annex F, associated with this chapter, sets out the detailed analysis of the data supplied and the field-specific cleaning procedures. Cleaning regimes were applied to selected fields in the outputs database, concentrating on those essential for a satisfactory match to be made to commercial citation databases. This prioritised fields such as journal titles, volume and page numbers and unique identifiers, including DOIs (digital object identifiers) and Thomson UTs (unique tags).

126 The outputs dataset was used for the development and assessment of matching strategies and journal coverage of both Thomson Reuters and Scopus databases. These complementary analyses were done in parallel but independently, by *Evidence* and HEFCE respectively, and the results later compared by HEFCE.

127 At the end of the cleaning stage, the proportion of the outputs with 'valid' data which could be used to match the outputs to the citation databases of Thomson Reuters and Scopus was estimated at around 86%. This was based on the output having either a DOI of the format '10.[string]' or a combination of standardised journal name, valid year in the range 2001 to 2007 and cleaned volume and page data.

128 The data were combined with the Table 1 staff data and used as the subsequent outputs dataset for the Symplectic Publications verification system.

9 Reconciliation of journal outputs to the Web of Science

129 *Evidence* was responsible for reconciling the article records supplied by pilot HEIs to the article records in the Thomson Reuters Web of Science commercial database. Reconciliation of pilot HEI data to Elsevier's Scopus database, an alternative source of commercial supply, was carried out by HEFCE. This, and comparison between the two will be reported elsewhere.

Matching strategy to Thomson Reuters citation database

130 *Evidence* has significant experience in matching publication records, or bibliographic data, to citation databases. In the case of relatively small databases (for 20,000 records or fewer) this can be done for all the records in the database, including those not covered by the database, which can be identified as not abstracted. This methodology would use a number of logically constructed strategies such as those below, coupled with manual verification where duplicate matches were obtained and where a single piece of data differed from that submitted.

131 For the purposes of the REF pilot study, however, such a time-intensive process (for 250,000 records or more) would not be feasible. More fundamentally, it would not be possible to extend such a process to later REF implementation across the complete higher education sector. It was therefore necessary to apply a more robust process for the REF pilot while considering the processes that would be required to underpin later systemic application.

132 Three different matching strategies were applied to the outputs database to link outputs to citation databases. This work was shared with HEFCE, who matched independently the Scopus data and will report separately on the outcome. After each matching strategy was applied, all matches were dropped unless they resulted in a unique match; that is, where the output linked to only one record in the citation database on those criteria. All duplicated matches were dropped entirely and no manual verification was employed.

133 A number of differences between the Thomson Reuters and Scopus databases may affect the outcomes, the most obvious of which would be the number of outputs not included in the UK address-restricted Thomson database owing to non-UK addresses or no address data.

134 Annex G associated with this chapter describes field-specific matching strategies and summarises success rates, including a figure which shows the proportion relative to the total number of outputs type 'D' for the Thomson Reuters-matched data:

- DOI data were available for 49% of outputs type 'D'; 57% of outputs with DOI data were uniquely matched to Thomson data, giving an overall matching success of 28% or 70,147 outputs.
- Journal and article title data were available for 85% of outputs type 'D'; 73% of outputs with journal and article title data were uniquely matched to Thomson data, giving an overall matching success of 62% or 155,986 outputs.
- Journal title, volume and pagination data were available for 78% of outputs type 'D'; 78% of these outputs with journal, volume and pagination data were uniquely matched to Thomson data, giving an overall matching success of 61% or 152,440 outputs.

Thomson Reuters' unique article identifier

135 The gold-standard, and easiest, method to assign the citation data from any commercial citation database would be to collect the outputs' unique identifiers (the UT in the case of Thomson Reuters' Web of Science) at the time of data submission.

136 There is currently a limitation on this, as only 10 out of the 22 pilot HEIs were able to provide even partial data, which amounted to less than 13% of outputs type 'D' overall. No Scopus article identifiers were provided by any of the pilot HEIs.

137 *Evidence* undertook a limited audit of the Thomson UT data supplied. Of the 32,000 Thomson UTs supplied, 2,300 were not matched to data where the records were restricted to authors listing at least one UK address on the publication. Citation data for non-UK records were sourced

138 Some doubt must apply to reliance on the current accuracy of the UT data tags in institutional systems, but this could be much enhanced by automatic downloading of full article records (which would in itself improve data supply for institutional purposes) and identification tags.

139 Of those outputs where the UT from the pilot HEI was linked to the Thomson database, it is interesting to note that almost 6,700 did not then match on the pilot HEI record of the truncated article title (40 characters). This usefully illustrates the inadequacy of the article title for matching these data. *Evidence* has not quantified these mismatches, but some common causes were identified:

- Use/omission of quotation marks around the “article title”. Without reviewing the printed output it is not possible to ascertain whether this is an inputting error or an abstraction error.
- Different denotations of non-standard characters such as beta. Thomson has a standard notation for these characters which may differ from that input by the author, but the range/stability of inputted characters was extremely variable.
- Poor title inputting (presumed to be institutional). For example, it is visually obvious that “Hox gene mutation that triggers Nonsense…” is the same as “A Hox gene mutation that triggers nonsense…” but these records will not match electronically.
- Misspelling. For example, “A first version of the Caenorhabditis…” is not logically matched to “A fist version of the Caenorhabditis…”. In this instance the error was with the abstracting service, but in many instances such spelling and format errors were institutional.

140 As well as these ‘errors’, it was apparent that some Thomson UTs had indeed been wrongly associated by pilot HEIs with the rest of the output record.

141 The matching rate and accuracy can be significantly enhanced where institutions download from Web of Science and store not only the core article record but also, as part of the associated information, the unique identifiers such as the UT. Further analysis is then essentially automatic.

10 Issues arising from data gathering and processing

142 This chapter reviews issues that arose from the overall process of gathering data from the pilot HEIs, managing that process in interaction with the pilot HEIs' staff, and cleaning and structuring the data records.

143 We need to reiterate at the outset that these issues are reported to provide guidance on potential constraints and pitfalls for any future exercise. The problems that arose in the pilot should in no sense be seen as a criticism of the pilot HEIs, whose staff worked with exceptional diligence and enthusiasm to support an onerous exercise under a pressured timetable. It was that very pressure which helped to reveal problems that might arise where historical systems have not been required to produce data of a flexibility, scope or validated accuracy that would be required effectively to support the REF.

144 Many issues that arose during the pilot exercise are less likely to arise during a full-scale national implementation. The specifications for data collection would have been reviewed extensively over some prior period and then studied and elaborated within institutions. The data templates would have been locked down so that data entry would necessarily conform to specification. Institutions would have had longer to collect, review and prepare their data. The data checking, validation and cleaning would be simpler because of those high-quality inputs. And there would have been much more time to test and execute the necessary procedures.

145 Nonetheless, all the problems that did occur will have to be taken into account. Many of them reflect a systemic issue across the entire HE research base: currently, most institutions are not readily able to supply the data that would be required to support the kinds of analyses that the REF would require. This constraint is not limited to any particular group. Nor does it apply to any particular type of data. It affects both internal data about people and shared external data about outputs.

146 It may seem surprising that, after more than two decades of research assessment, relatively few institutions have comprehensive, structured research management information systems. In part, this state of affairs reflects the extent to which research remains an activity driven by the principal investigators rather than managed by institutions. This is right and proper and is a keystone of the success of the UK research base. It will therefore be necessary for HEFCE to consider how best it can prosecute its exceptional needs without damaging the excellence it seeks to evaluate.

147 We have listed the various issues in some detail because they are a useful source of information for research managers. This is not a template against which the HE research base should develop a uniform or centrally managed response.

General issues

148 Some generic issues, about the data collection process and about data quality management, were associated with many pilot HEI submissions. The tables in Annex H provide an aide memoire and an overview for the many pilot HEI staff who helped the contractors to solve the conundrums that emerged from the submitted data.

149 A problem immediately encountered was that the pilot HEIs had not been directed to:

a. A standard notation for recording missing data.

b. A clear statement about recording specific numeric fields (for example, page numbers as a single value with no preceding characters, brackets or punctuation).

This led to a wide range of actual indications which then had to be decoded. This is clearly something that can and must be addressed in any implementation.

Database structure

150 We asked for a simply structured, normalised database consisting of two tables linked by a third – Table 1 for staff records, Table 2 for publication records and Table 3 for staff-publication links:

- Each record in Table 1 should have referred uniquely to a member of staff. If data were drawn from several institutional databases, then it should have been combined into a single record before being submitted. In practice, de-duplication of staff records was required and would have been best done by the pilot HEI, being more familiar with its own database structure and content and better placed to pursue inconsistencies.
- Each record in Table 2 should have referred uniquely to a publication. Again, records had to be extensively de-duplicated so that jointly authored papers appeared once for each pilot HEI.
- Each record in Table 3 should link a publication from Table 1 to an author from Table 2.

151 We requested that the data be submitted in this format so that referential integrity could be enforced. This means ensuring that there is at least one record in Table 1 linked to every record in Table 2 (every publication has an author), and at least one record in Table 2 linked to every record in Table 1 (every author has a publication). Table 3 links must refer only to items in Tables 1 and 2 and contain only ID fields.

152 We provided table templates which specified the names and properties of the fields within each table, and our initial estimate of the scale of the data-processing task was based on the assumption that data would be supplied in this format. This assumption proved incorrect, and pilot HEIs were unable to comply with this expectation in the time available. As there were 22 pilot HEIs, and each had interpreted the data specification in a slightly different way, this led to a plethora of error types.

153 Incorrectly structured data had to be restructured (by creating unique identifiers, re-ordering fields etc) before they could be assessed for completeness. Successive submissions meant that this restructuring process had to be repeated with every new batch of records received.

Effects of having stages running concurrently

154 Owing to the constrained timetable of the pilot project, three stages were running concurrently: the period during which we were receiving successive submissions of data, the data cleaning and validation process, and the development of the database. Running such stages consecutively will increase efficiency and make it less onerous to track and trace data and design-modifications, but for the pilot this would have elongated the timetable by months. Audit analyses of data quality and content also had to be reproduced as closely as possible at successive stages using modified databases, which extended reporting time.

Serial submissions

155 Pilot HEIs were permitted to make successive submissions of data to assist in keeping to a tight timetable and allowing flexibility in response to varying levels of data availability.

156 Multiple submissions increased the central workload disproportionately and also increased the workload on pilot HEIs, who helpfully sought to highlight version changes. In addition to the initial 22 submissions, a further 38 separate files were processed. Only three pilot HEIs supplied all data without subsequent amendments. Three large pilot HEIs each made five submissions, and one institution made seven. In each case, every additional file had to be checked, cleaned and compared against previous submissions before then being incorporated into the master dataset.

157 The benefits of allowing multiple submissions were: iterative development of data processing and cleaning techniques; flexibility to vary the requirements according to individual pilot HEIs' status; and an opportunity for us to build working relationships with pilot HEIs' staff.

158 The original intention, to analyse differences between first and final versions of data, turned out to be infeasible because of the lack of uniform compliance with the standard template.

Prioritisation of validation

159 On data quality, it was necessary to ensure that the REF pilot data were fit for current objectives, but there were neither resources nor a requirement from HEFCE to aim for 100% accuracy.

160 Because data matched the template inconsistently, the validation task was considerable. The fact that this was a pilot process added complications to prioritising data quality issues. We had three broad objectives: to ensure the successful implementation of the Symplectic Publications system so that pilot HEIs could review and amend records; to develop a database from which to create citation analyses for Task B; and to assess the readiness of pilot HEIs to provide suitable data.

161 It was also possible that the address-model might eventually be selected as the preferred REF methodology. If this were the case, then the criteria for inclusion of publications would be address-based rather than author-based, so there would be no need for a staff table or a links table. The pilot project therefore aimed to highlight data-collection issues as they occurred, but not necessarily to solve them.

Data availability

162 Pilot HEIs' difficulties in providing data as specified must have stemmed in part from the current status of their institutional data systems, not from any lack of willingness to comply. From our interactions with the invariably responsive and helpful liaison staff, we detected a lack of connectedness in information systems across human resources, research management and publication databases. This was affected in some institutions by a paucity of information about the exercise, which had been introduced and implemented in a short period, and uncertainty as to where corporate responsibility lay.

163 The present lack of data connectedness is unsurprising. Staff and research databases have evolved independently to serve separate needs. The effort required using present systems to combine data for other purposes, such as the REF, should not be underestimated. However, the compromised ability to respond to the REF data request is slightly more surprising given the recent RAE2008, to which the REF request was designed to be as similar as possible.

Table-specific issues

164 Summary aspects of specific features of general data processing and the field-specific data issues associated with Table 1: staff and Table 2: outputs are attached in the relevant Annex (Annex H).

11 The Symplectic Publications system

165 Symplectic, a subcontractor to *Evidence*, focused its work around two major tasks in terms of the creation of a database for the purposes of analysis:

- Actual publication data provided by pilot HEIs and the presumptive Web of Science data provided by *Evidence* were reconciled in an automated fashion. The aim was to produce a single, inclusive pool of journal publication data to be used as the total set of potential data that could be associated with the pilot HEI staff lists.
- Publications data were matched with the academic staff lists provided by pilot HEIs.

First, this chapter explains the methodology behind the addition of presumptive data to the actual pilot HEI data. The next section then addresses the issue of name disambiguation.

Addition of presumptive Web of Science data

166 The 22 pilot HEIs provided their own publications data as described above. *Evidence* staff cleaned these data and implemented a standard format across all datasets. These data were then restricted to include only journal publications, stored in CSV (comma separated values) format, and passed to Symplectic for processing together with the presumptive dataset, an extract from *Evidence*'s address-rectified version of Thomson Reuters' UK National Citation Report (NCR). The presumptive data constitute what is in fact a 23rd dataset for Symplectic purposes. This dataset was treated in a distinctly different way to the other 22 datasets and must be thought of as a distinct resource.

167 The NCR extract included all journal publication output data presumed, by virtue of the address data associated with each output, to be linked with at least one of the 22 pilot HEIs for the period 1st January 2001 to 31st December 2007. The data fields associated with articles which were provided in the NCR extract are set out in Tables I1, I2 and I3 in Annex I, and the relationship between these tables is shown schematically in Figure I4 in that Annex.

168 Specifically, the NCR database holds data associated with the institutional (and other) addresses associated with each article. This includes both standard institutional data and variant address data.

169 In the context of the data supplied by the 22 pilot HEIs, it is appropriate to emphasise that data quality, even after basic data cleaning performed by *Evidence* staff, required additional work in order to make the data viable for automated treatment. Even though the format of data had been standardised at the level of field naming and field length, the consistency of data contained within those fields varied both across and within institutions. To illustrate this, the DOI field often contained data which did not correspond to a DOI format or which held additional prefix characters such as "DOI:" or "http://dx.doi.org/". Additional processing had to be done to remove all such data inconsistencies prior to disambiguation processing.

170 Many outputs had not only been submitted by multiple HEIs, each in their own datasets, but also multiple times within single institutional datasets. Additionally, the presumptive dataset contained an additional copy of many of the outputs returned by HEIs. Hence, the aim of the initial activity was to create a single database entity for each physical journal article, henceforth referred to as an 'output'. This output would then be associated with authors in their respective organisations. A complexity to this task was that some authors appeared multiple times in the combined Table 1 data supplied by HEIs. Hence, the same author name (but distinct staff records sourced from multiple institutions) might be associated with the same output multiple times if the author held relationships with multiple institutions during the census period.

171 Symplectic applied a multi-stage process to the datasets to create a single publication data pool:

- Data from the pilot HEIs were stitched together into a single data table.
- Then, each row in this dataset was rectified against the *Evidence*/Thomson Reuters dataset.

- Where an appropriate matching row appeared in the Thomson Reuters data, the UT (unique tag – the Thomson unique article identifier) was associated with the row in the institutional dataset.
- A number of criteria were used to achieve this match. Records were matched on DOI or UT where those fields had been supplied by the pilot HEI. Then, for those records which contained neither DOI nor UT data, a match was sought via a combination of journal name, volume of publication, pagination, year of publication and title keywords.
- Next, duplicate data were removed from the dataset.
- Data from each individual institution were checked and duplicate article data were removed from each dataset. This de-duplication was based on the UT and DOI fields for those articles with such data.
- For those records without either of these two unique identifiers, similar criteria were used as in the earlier part of the de-duplication exercise, focusing on journal name, volume of publication, pagination, year of publication and title similarity.
- Finally, inter-institution duplication was removed by the same mechanism.

The overlap between the presumptive database and the pilot HEI databases for each institution is shown in Annex I (Table I5 and Figure I6).

172 The outcome of the processing activity described above was a classification of data records into three distinct categories:

- output with an *Evidence*/Thomson Reuters record alone;
- output with an institutional record that had also been matched to an *Evidence*/Thomson Reuters record;
- output with an institutional record alone.

173 Annex I (Figure I7) shows this breakdown schematically. It should be re-emphasised that many data rows from the original distinct institutional datasets had been replaced by a single set of bibliographic data describing each output. Each data row in the amalgamated institutional data pool may have a relationship to a single record in the presumptive dataset.

174 In order to add Table 3 (links) data to the process at a later stage, mapping information was maintained at all stages of this process. That is, whenever two institutional article records were matched together, the unique identifiers originally assigned by *Evidence* were recorded as being associated with the same output. This additional process allowed the correct mapping of Table 3 data into the de-duplicated data pool created from the processing described in this section.

175 The picture that should be borne in mind is that there are two distinct bibliographic data pools: the amalgamated institutional data and the presumptive Thomson Reuters data. As a result of data processing, much duplication was removed from the institutional dataset and a mapping between the two datasets was established, linking some of the records in one dataset to some of the records in the other.

176 In the following section we discuss the methodology for associating these data to named people in the pilot HEI staff lists.

Disambiguation of author and staff names

177 The second task for Symplectic was to take the Table 1 data on staff names provided by pilot HEIs and to rectify this to the journal publication dataset described above. This section describes the process applied to the various data to suggest a link between authors and articles where none previously existed.

Rationale

178 It should be recalled that pilot HEIs had been asked to provide, where possible, a link between their staff (Table 1) and the output publications (Table 2) in the form of a list of links (Table 3). It should also be recalled that not all pilot HEIs were able to supply such data for all records, and

179 It is important to understand the motivation of the methodology chosen to carry out this task. In light of the difficulty which most pilot HEIs found in supplying the type of data required for this project, it was assumed that the data provided in terms of links between publications and authors were correct insofar as they went, but not sufficiently complete to form the basis of a detailed and comprehensive analysis.

180 In order to seek to maximise the linkage between staff and the publications data, it was decided that an automated mechanism was needed to suggest potential links. The application of automated methodology suffers, however, from two major drawbacks. First, because of (for example) synonyms, there is the potential risk of identifying false positive matches – i.e. suggesting that authors have written more papers than is in fact the case. Second, because of (for example) name variants for any one individual, there is a risk of missing matches – i.e. failing to suggest that an author has written several papers when those publications are in the presumptive dataset. Since the process of identifying papers written by authors at a particular institution is of general interest, it was decided to carry out the current analysis in a way that gave the largest possible scope for a later analysis of appropriateness of matching algorithms.

181 In order to keep the spirit of full scientific methodology, a simple algorithm was devised to match publication outputs with their authors. This relied on matching institutionally supplied names and name variations from that institution's Table 1. Additionally, for each user, the 'searchable data' were restricted to the portion of the data associated with each staff member's home institution (sourced from either the address data supplied as part of the *Evidence*/Thomson Reuters address-disambiguated database or the data supplied by institutions themselves in Table 2). The result of this activity was intended to be a highly inclusive dataset of potential associations of outputs with staff: a suggested but highly extended Table 3.

182 The Table 3 links supplied by each institution were then applied on top of these data. Any Table 3 record was considered to form a firm link between an author and an output. Outputs in this category were listed as "approved" articles in the Symplectic interface. Any suggested link from the automated mechanism described in the last paragraph was treated as a "pending" link, and institutions were invited to review the data through the Symplectic data-checking mechanism.

Data modifications

183 The initial data collection and the dataset prepared according to the methodology described in previous sections included all potential output Type 'D' journal items, as would the typical RAE data submission. This meant that it included not only standard research journal articles but also other sub-types of publications such as reviews and abstracts, as well as letters and editorials.

184 Pilot HEIs pointed out that some of these publication types were potentially inappropriate for an evaluation exercise, despite being readily available through commercial bibliographic databases. In response to these arguments, the dataset was restricted to include only the data associated with mainstream journal articles and reviews. The more minor publication types (not including articles and reviews but including conference proceedings and conference abstracts) were removed prior to release of the data to pilot HEIs.

185 A three-stage process, shown in Annex I (Figure I8), was used to create a list of links between outputs and authors. The first stage was to apply the automated author/paper-matching algorithm to the staff data (Table 1) and the records in the outputs dataset. For each staff record in the amalgamated staff dataset a search was performed on the part of the outputs dataset associated by address with the institution which supplied the staff record. Following this processing, data could be grouped into two classifications: "suggested link" and "unlinked".

186 The additional author/paper link data supplied by the institutions (Table 3) was then applied to the staff and outputs data. The effect of this additional data was to introduce a new classification to the underlying data. Data could now be classified as "associated", "suggested" or "unlinked". As a result of processing Table 3, some articles moved from the "suggested" status into an "associated" status, while other articles might have moved from an "unlinked" status directly into "associated". In the context of the data-checking software, "associated" corresponds to an

Data-checking mechanism

187 The staff-author disambiguated data were presented to pilot HEIs in two classifications:

- *For each academic* in the institution there was a list of "accepted" publications – those which were in the journal publication list provided by the pilot HEI and to which they were already able affirmatively to link the academic in question (i.e. via the data provided in Table 3). Some, but not all, of these records would have had Thomson Reuters data associated with them from the first part of the matching process.
- A second class of articles were those which had been suggested to academics and had a "pending" status. These articles could have components from the combined institutional dataset, or from the *Evidence*/Thomson Reuters presumptive data, or both. These suggested outputs were found using the criteria outlined above.

188 Pilot HEIs were able to review data through a customised web interface to the system and change the status of articles from "pending" to "approved" or "declined" depending on whether they felt that the suggested association was appropriate. Annex I (Figure I9) shows three example journal articles in the Symplectic data-checking system. It is clear to the reviewer where the data have been derived from for each article (institutional data, presumptive data or a mix of the two) and also whether the article is in an "approved", "pending" or "declined" status.

189 For smaller pilot HEIs, it was practical not only to use the data-checking mechanism for review of the articles but also to work through the lists of articles in "pending" status and move them to "approved" or "declined". For larger pilot HEIs this approach was not considered practical because of volume and so they were offered the ability to download a CSV-format spreadsheet (see Annex I, Figure I10) of the status data for each link. Based on feedback in meetings with pilot HEIs, Symplectic extended the number of columns in the spreadsheet report to include DOI and publisher links and Thomson UT where available. Institutions were able to edit the status of an article by altering the spreadsheet and returning that spreadsheet to Symplectic for upload into the system. Annex I (Table I11) shows the numbers of suggested articles for each HEI.

190 It became clear that a sampling strategy would be required to employ strategic approval methodology in order to ensure maintenance of the data quality and to understand weak points. Several methodologies were applied to institutional data to help institutions with larger amounts of "pending" links. Not all strategies may have been applied to all data but they were as follows, in order of application:

- Institutions could focus their efforts on disambiguating the outputs for staff in Table 1 who had been classified as being "RAE eligible".
- Outputs with only a single institutional address associated with them from the *Evidence*/Thomson Reuters database were accepted.
- Outputs where the institutional named author was the first author were accepted.
- Links to a particular Unit of Assessment already identified through another member of staff were accepted.
- Remaining links were declined.

Table I11 shows the numbers of "approved" articles as a percentage of "pending" (suggested) articles after manual intervention and application of strategies.

Issues associated with author-staff disambiguation

191 The approach adopted for the REF pilot project had inbuilt issues surrounding the amount of data which some pilot HEIs would have to handle. Within each institution there would naturally be a number of synonymous authors at the level of the data held for this pilot exercise. It has been pointed out that a number of strategies could be used to more accurately disambiguate authors by using the Table 3 data provided by pilot HEIs. There are, however, two reasons why this was not

192 Individual pilot HEI datasets varied significantly in terms of the volume of the suggested links (Annex I, Table I11). The scaling factor between pilot HEIs is driven by several factors:

- The most important factor is that if one works in an institution with a large number of researchers then there is a high probability of a second researcher sharing at least one's surname and possibly also one's initials. This probability is related to the 'surface area' of the research community and hence an approximate square-like scaling may be assumed.
- Larger and more research-intensive institutions tend to have article production levels that are, on average, higher than at smaller institutions. There are often more papers per researcher.
- Larger institutions also have capacity to collaborate more widely. This entangles with another effect which is that, since address data associated with papers are not directly linked to specific authors (although this is now being rectified), the names of collaborators not at the institution may be readily but incorrectly matched with collaborators who are at the institution but not associated with that paper.

All these factors mean that the synonym problem rapidly became very difficult to deal with for larger pilot HEIs.

193 An issue which affected smaller HEIs more than their larger counterparts was the accuracy of the data presented. In several cases Table 3 links were supplied which apparently linked staff to articles where either staff or output did not appear in Tables 1 or 2 respectively. This, together with the issue of pilot HEIs which were unable to supply any Table 3 data, led after processing to a corpus of approximately 8% of the total data which could not be linked to any author supplied in the Table 1 return.

12 Creation of bibliometric database

194 The data records and links processed by Symplectic Publications and accepted by the pilot HEIs were resubmitted to the secure server. This constituted the required development of the bibliographic database.

195 The final steps, to create the bibliometric database required for the REF pilot project, were:

- association with the validated publication records of their relevant citations data;
- normalisation of the citations data to enable comparative analyses.

196 Normalisation is a process of adjusting actual citation counts to an indexed value that takes into account both the year of publication and the field to which the article's journal is assigned. Normalisation, or rebasing, is essential because older papers inevitably have more time to accumulate citations and some fields have underlying citation rates that are much higher than other fields. There are a number of levels of aggregation at which such normalisation could occur.

197 A later report will describe the development of the combined publication and citation database and the decisions made regarding normalisation.

13 ANNEX A, Field specifications

Table A1 Detailed field specification for the pilot HEI staff data

Field	*Type*	*Expected values*	*Mandatory*
Institution ID	String	HESA institution code	Yes
HESAStaffIdentifier	String	<13 characters>	Yes (for Cat A and B staff)
InstitutionalStaffIdentifier	String	<24 characters>	Yes
UnitOfAssessment	Integer	1-67	Yes
Unit, department, school or other location	String	<Unlimited>	
Title	String	<50 characters>	
First name	String	<100 characters>	
Other first names	String	<500 characters>	
Initials	String	<50 characters>	Yes
LastName	String	<500 characters>	Yes
Alternative or prior surname(s)	String	<500 characters>	
Alias or 'known as' for publications	String	<100 characters>	
email address	String	320	Yes
Submitted to 2008 RAE	Boolean	True/false	Yes
RAE staff category	Character	A, B, C, D	Yes
Early career researcher	Boolean	True/false	Yes
Start date (if started after Jan 1st 2001)	Date	2001-01-01 to 2007-10-31	Yes (if started after 2001-01-01)
Prior institution (if started after Jan 1st 2001)	String	HESA institution code	
Leave date (if left before Oct 31st 2007)	Date	2001-01-01 to 2007-10-31	Yes (for Cat B and D))
Destination institution	String	HESA institution code	

Table A2 Detailed field specification for the pilot HEI outputs data

RAE field?	*Field*	*Type*	*For output type = 'D'*	*Expected values*
Yes	Institution	String		Your HESA institution code (including a campus code if applicable)
Yes	InstitutionalUniqueOutputId	String		<24 characters>
Yes	Year	String		2001-2007
Yes	OutputType	Character		A-T, a-t
Yes	LongTitle	Text	Article title	Text
Yes	ShortTitle	String	Vol number/ edition	<512 characters>
Yes	Pagination	String	pp	<64 characters>
Yes	Publisher	String	FULL journal title	<256 characters>
Yes	Editors	String		<256 characters>
Yes	ISBN	String	ISSN	<24 characters>
Yes	PublicationDate	Date	M/Y or D	2001-01-01 to 2007-12-31
Yes	DOI	String		<256 characters>
Yes	IsInterdisciplinary	Boolean	Default = no	True/false
Yes	IsSensitive?	Boolean	Default = no	True/false
No	ListOfAllAuthors	String		<512 characters>
No	Indexed by Thomson Reuters (ISI)?*	Boolean		True/false
No	Thomson unique identifier (UT or ISI_LOC)	String		<15 characters>
No	Indexed by Scopus?	Boolean		True/false
No	Scopus author identifier	String		N/K
No	Scopus article identifier	String		N/K

14 ANNEX B, REF Bulletins

The following URL provides a link to the main resources which relate to the development of the Research Excellence Framework:

www.hefce.ac.uk/research/ref/resources/

The first briefing report on “The REF pilot study: Bibliometric indicators of research quality” is at:

www.hefce.ac.uk/research/ref/resources/biblio_indicators.doc

The second briefing report on “The REF pilot study: Bulletin 2: revised data collection” is at:

www.hefce.ac.uk/research/ref/resources/Bulletin.doc

15 ANNEX C, Chapter 5: Data collection

Pilot HEIs were asked to detail by Unit of Assessment the following information:

- whether they would be supplying data in that UoA to the REF pilot;
- how many staff were submitted to RAE2008;
- how many additional staff would be returned to the REF pilot;
- how many outputs were submitted to RAE2008;
- the number of additional outputs for which the pilot HEI had data;
- an estimate of the proportion of their total outputs that the pilot HEI expected would be found in journals indexed by the commercial databases.

On average, pilot HEIs estimated that they would be able to supply about three times as much data to the REF as they had submitted in their RAE2008 returns. Unsurprisingly, this gross figure disguises very significant variation by HEI and by UoA within HEIs.

Some institutions reserved their position pending discussion with academic departments.

The full figures on estimated and actual data provision are shown in Annex E.

16 ANNEX D, Chapter 6: Generic data processing

First data tranche

Pilot HEIs were asked to supply data across three key tables. First, by the beginning of August:

- Table 1 covered staff – all researchers who were eligible to be submitted to RAE2008 (using the same criteria and census date of October 31st 2007) and all staff eligible to be submitted as Category A and B, and any Category C and D staff for which data were readily available. We invited pilot HEIs to augment these core data with any additional information about people who might have published or been co-authors on publications, to enable the most complete identification of author names.
- Table 2 (initial) covered outputs – all research outputs eligible to be submitted to RAE2008 that were authored by the Table 1 staff (published from January 1st 2001 to December 31st 2008). We asked that outputs by Category A and B staff should be as complete as possible and that outputs by Category C and D staff and any other researchers named in the REF Table 1 submission should be included where readily available.

Second data tranche

Pilot HEIs were able to enhance their initial submissions by additional data collections and work with academic departments, although this was necessarily over the summer when many staff were absent. They were asked to supply, by the end of August:

- Table 2 (final) – with additional outputs beyond those submitted to the RAE;
- Table 3 listing authorial links between Table 1 (staff) and Table 2 (outputs).

Once the pilot HEIs' most complete lists of their known outputs were to hand, we could draw from these lists the specific material published in journals. This could then be checked against the commercial databases to make a match with definitive article records. From that analysis we could then determine if we had presumptive material, associated with the pilot HEI's address but not yet in its database. These additional records could subsequently be offered to the pilot HEIs.

Principal issues

Table 1: staff data. The first major problem raised by many pilot HEIs was the coverage of Table 1. It rapidly became evident that the limit to what could be provided by many pilot HEIs without significant additional effort was a table that corresponded to the staff list for RAE2008. This would therefore cover staff who were eligible and submitted to the RAE. Going beyond this to include a wider range of staff who were research active but not RAE eligible, and who might be potential authors or at least co-authors, was very much more difficult. Indeed, in many cases only detailed cooperation from academic departments would have enabled this to be completed in the way originally intended.

Table 2: output data. The second major problem was the extension of the collection of Table 2 data beyond that already held by the pilot HEI centrally for the purposes of RAE2008. Although more extensive data were often understood to be available, these were held by other parts of the institution and to a non-standard format. Academic departments had often processed their own data and then supplied a selective file to the centre to support the RAE submission. It became evident that few institutions have in place a system for the regular submission of standard and comprehensive publication data or content by academic departments to any central database or repository. The collegial nature of the institutions also tends to work against the ready imposition of any formal mandate to achieve such an objective, as advocates of repositories and open access have discovered.

Data management

HEFCE, as the responsible agent, retrieved the files submitted by pilot HEIs from the extranet, placed them on the remote server and informed the contractor.

- Each file was retrieved from the remote server by secure FTP (file transfer protocol) – a time-consuming process for large files – and temporarily saved locally for easier manipulation. A log file was then updated detailing the filename, the institution, the date and a description.
- Commonly, the file comprised an Excel workbook consisting of four sheets, one containing notes on the data, and one sheet each for Tables 1, 2 and 3. The file was opened in Excel 2007. Any remarks that might have been made on the data were noted, and the first sheet containing data was examined by an operator.
- The first step was to check whether the data were in an Excel list format (contiguous data containing no blank rows or columns). Where they were not in such a format, the data were amended by removing the blank rows or columns to make it so.
- The data were then checked by inspection for the goodness-of-fit to the original template specification. For the fit to be good, we ensured that:
    - the field order matched the common template. Where it did not, the data were corrected by moving fields relevant to the template to their required positions, with any additional fields placed to the right of the last template field;
    - the field names were correct. Where they were not, the relevant labels were copied across from the template.
- If the generic fit was not good – for example, if the data were provided UoA by UoA each in its own sheet, or author data provided from different systems in different sheets – these data were consolidated first. Sometimes this required that data-points in a single cell be separated out into new records. New VBA (Visual Basic for applications) procedures were written or existing ones amended to automate this.
- An auto-filter was applied to the field headings.
- For each field, the contents of the field were checked to note any inconsistencies (for example, for the institution field, that it contained only one code and that the code was a four-digit number beginning with 0).
- If the field contained other data, those other data were noted and checked visually for a pattern (for example, if the contents of the field were surrounded by HTML tags, or if the contents of the field should have been in a different field). If a pattern emerged subsequently (when checking another sheet, perhaps from another institution), the pattern was noted at that point.
- This process was repeated for each field in the sheet: 20 fields for Table 1, 20 fields for table 2, and four fields for Table 3.

The count of records that met certain criteria was enumerated.

- For Table 1, a simple count of records, and a count of records meeting the following criteria:
    - HESAStaffID is blank
    - InstitutionalStaffIdentifier is blank
    - UoA is blank
    - Eligible is not blank
    - Submitted is not blank
    - RAE category equals A to D (separately)
    - Early career is not blank
    - HasStartDate is not blank
    - HasPriorInstData is not blank
    - HasLeaveDate is not blank
    - HasDestinationData is not blank.

- For Table 2, a simple count of records and a count of records meeting the following criteria:
  - Publication type equals A to T (separately)
  - LongTitle is blank
  - ShortTitle is blank
  - Pagination is blank
  - Publisher is blank
  - SourceYear is blank
- For Table 3, a simple count of records and a count of records meeting the following criteria:
  - HESAStaffID is blank
  - InstitutionalUniqueOutputId is blank.

These data were stored in the DataLog workbook, so as to provide descriptive statistics on the data institution by institution.

When this was done, the data were copied from the file and pasted as values into the template. This file was then saved as [UniversityName]DataInTemplate.xlsx

The process was repeated for any subsequent sheets. The process was then repeated for each file. Because many institutions submitted many tranches of data, there was considerable duplication of effort, as each stage was necessarily repeated for each file. There were approximately 80 files in total.

The data in each institution's [UniversityName]DataInTemplate.xlsx were imported into Access for batch cleaning and processing.

Central data management: delays and consequences

Chapter 6 described the pragmatic decision to accept all data submitted by pilot HEIs rather than to return any records not meeting the data specification and quality requirements. This decision was taken to meet the timetable requirements and to address deficits via central batch solutions, but it had a significant impact on the resources available for original tasks.

The greatest impact was on the speed of development of the core database.

There was, first, a much greater delay than could have been anticipated in collating 22 sets of submissions into a coherent and uniform set of files. Even after data moderation and cleaning, it is almost certainly the case that some records remained uninformative for later analysis (e.g. because of missing information) while other records were passed as valid and complete but could, if sufficient checking had been possible, have proved to be ambiguous or inaccurate.

A second area of delay was in the processing of additional records from the presumptive data. Because it took longer than originally expected to match pilot HEI data to Thomson Reuters data, it was not feasible, as intended, to offer the supplementary records to the pilot HEIs to review in a separate exercise. Instead, these records were passed to the Symplectic stage of staff-author disambiguation and then offered to the pilot HEIs alongside the data on new, indicative links. This meant that staff in the HEIs were presented with two verification tasks at the same time rather than having these tasks presented separately.

Because of the underlying deficits in data quality, the task given to Symplectic was itself more onerous and complex than had been intended. This meant that a great deal more work was required to manage and process the data in the Symplectic Publications management system in seeking to link author and staff names. Indeed, this task became so complex that emerging characteristics of the data were still being investigated well after the indicator analyses (Task C) had started.

The combined effect of delayed data handover between *Evidence* and Symplectic and the poor relative quality of the data at that point exacerbated the delay in offering enhanced data to pilot HEIs for verification. It had been intended that this should have started by October. In practice, it was

December before substantial data were offered and the deadline for completion of the verification task was necessarily extended into early 2009.

The scale of the verification task for the pilot HEIs was itself much greater than had been anticipated. This is discussed in more detail in section 11, but the key aspect was the fact that larger and more research-intensive institutions have, naturally, both more staff and more outputs per staff. As a result the number of possible author-staff links increases geometrically in leading research universities. This is obvious with hindsight, but the specific impact of data quality on the relevant workload within the REF pilot project was not fully appreciated until the data challenges actually presented themselves.

The root cause of the serious delays in the project lies in the quality of the data supplied by the pilot HEIs and accepted by the contractors, but there is no sense in which the institutions bear any blame for this. First, there was a common but unduly optimistic view that the HE research base holds better research information than is in fact generally the case. Second, there was an erroneous assumption about the ability to replicate let alone extend the submissions made less than a year earlier for the RAE. Third, the timetable and timing of the project were less than ideal for the pilot HEIs. Fourth, the scale and impact of the additional work in data cleaning was initially underestimated. Had it been correctly evaluated, this might have led to changes in planning decisions at an earlier stage.

Important lessons have, hopefully, emerged. The central one is that most institutions will require a very clear and extended implementation pathway before the REF could be introduced on a national scale.

17 ANNEX E, Chapter 7: Staff data

Significant time was spent giving thought to what minimum data would be needed specifically to enable the planned analyses (as distinct from what might ideally have been requested for more comprehensive but speculative analyses). Similar consideration was given to the data that pilot HEIs would be able to supply within the time constraints. It was also expected that prior data collections, for RAE2008 or for HESA, would be easier to access and supply than novel data (for example, designations such as 'alias', early career researcher), so particular emphasis was put on using existing specifications and adopting them for the REF pilot data collection as closely as possible.

Data-cleaning routines were created in the form of a repeatable series of about 50 automated steps. This was initially based on procedures used by *Evidence* in previous work with organisations to develop staff-author links for their databases. The steps were progressively extended, modified and re-run as additional data-quality issues came to light during the database development.

The first version of the database utilised linked Excel tables so that data could be updated automatically, but this meant that some data types were not correctly imported. For the second version, Excel files were converted to text files, and then imported into Access. After importing the data into Access, the first tasks were to clean the HESA InstID field, delete the records which were blank and remove those which had some data but no LastName.

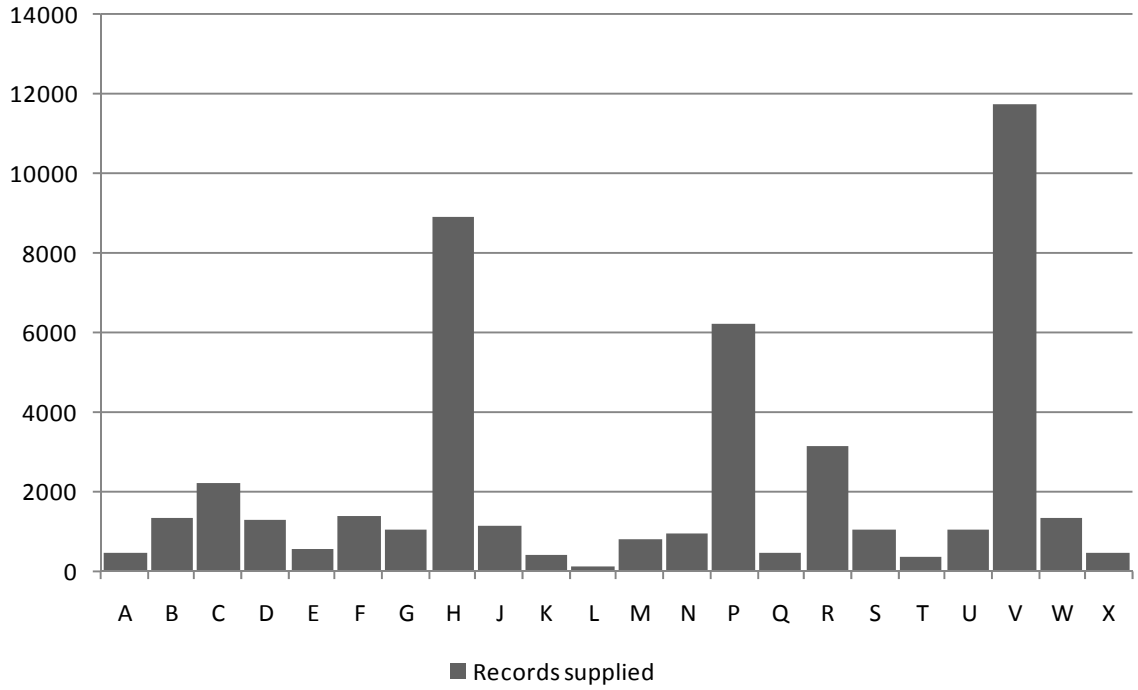
The estimated and total number of records supplied is shown in Table E1 and Figure E2.

Table E1 Summary of estimated and final staff data provision by the pilot HEIs

Institution	Estimated staff records	% RAE	Staff records supplied	% RAE	Institution	Estimated staff records	% RAE	Staff records supplied	% RAE
A	156	94.9	425	89.6	M			774	63.6
B			1311	68.9	N			946	27.2
C	506	78.3	2175	18.1	P	7798	21.4	6203	26.9
D	922	97.8	1274	70.6	Q	413	92.7	441	86.8
E	584	95.7	521	95.6	R			3119	34.9
F	1486	67.4	1354	70.8	S	1160	84.6	1000	89.6
G	928	26.5	1040	23.3	T			349	26.1
H	1600	86.2	8898	16.4	U			1007	14.4
J	356	47.2	1137	14.7	V	1606	67.6	11713	12.4
K	140	76.4	375	36.8	W	408	72.5	1308	24.1
L	96	61.5	92	77.2	X			416	23.3

Notes: % RAE indicates the percentage of the supplied records that were RAE-submitted staff; a lower percentage in the supplied data indicates the extent to which pilot HEIs were able to enhance the data over and above their RAE submission. Column 1: Estimated staff records = (headcount of staff submitted to RAE2008 + estimated headcount of additional staff to be included in the pilot); Column 2: % RAE = headcount of staff submitted to RAE2008/Column 1 * 100; Column 3: Staff records supplied = record count from Table 1; Column 4: Staff records supplied for RAE-eligible staff/Column 3 * 100

Figure E2 Number of staff and researcher records supplied by each institution



The following sections review the data quality and draw attention to specific data issues.

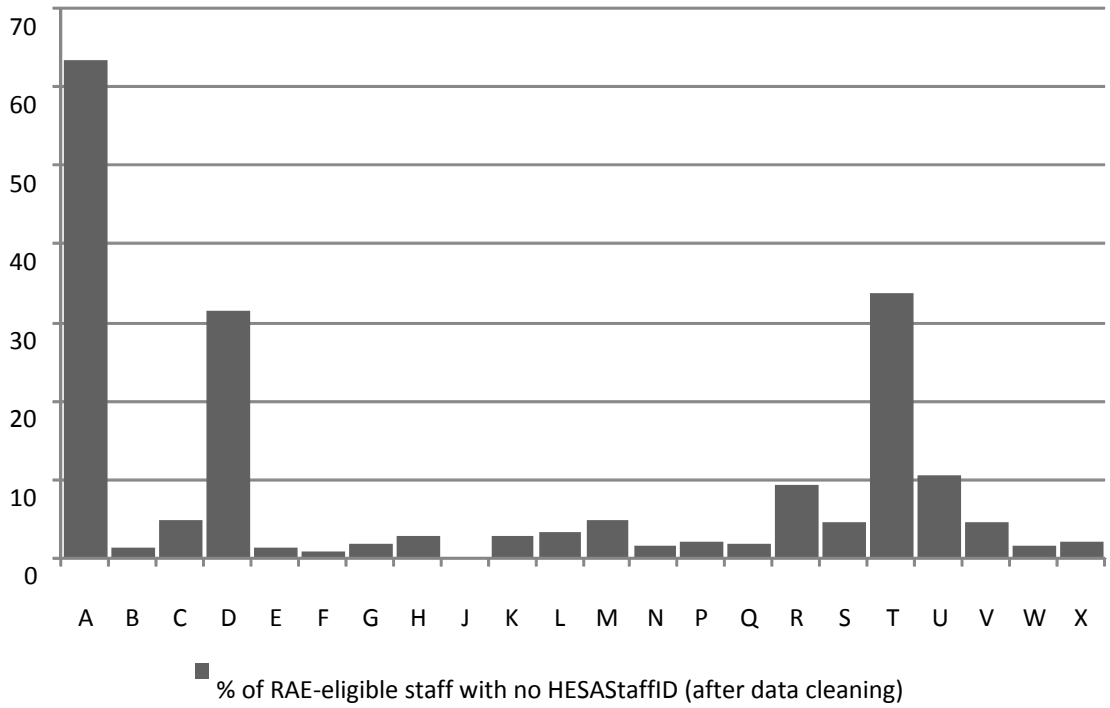
HESA staff identifier

Some staff contributing to an institution's research profile (e.g. those classified for the RAE2008 return as Category C) would not have a HESA staff ID, so this field was not mandatory for all staff. It was therefore anticipated that records would not necessarily and always have this identifier.

In practice, several pilot HEIs, of different sizes, supplied HESA staff IDs for virtually all staff, whereas the HESA IDs for similar pilot HEIs were missing from the majority of records.

The unavailability of these identifiers was not recorded in a standard way. For example, records with no data in the HESAStaffIdentifier field were indicated by: Null; "0"; ""; "unknown"; "none"; "no number available". It was therefore not immediately obvious where identifiers were absent because they did not exist (for Category C staff), or where they were missing or not recorded on the originating HR system. Figure E3 shows the spread of missing data across pilot HEIs.

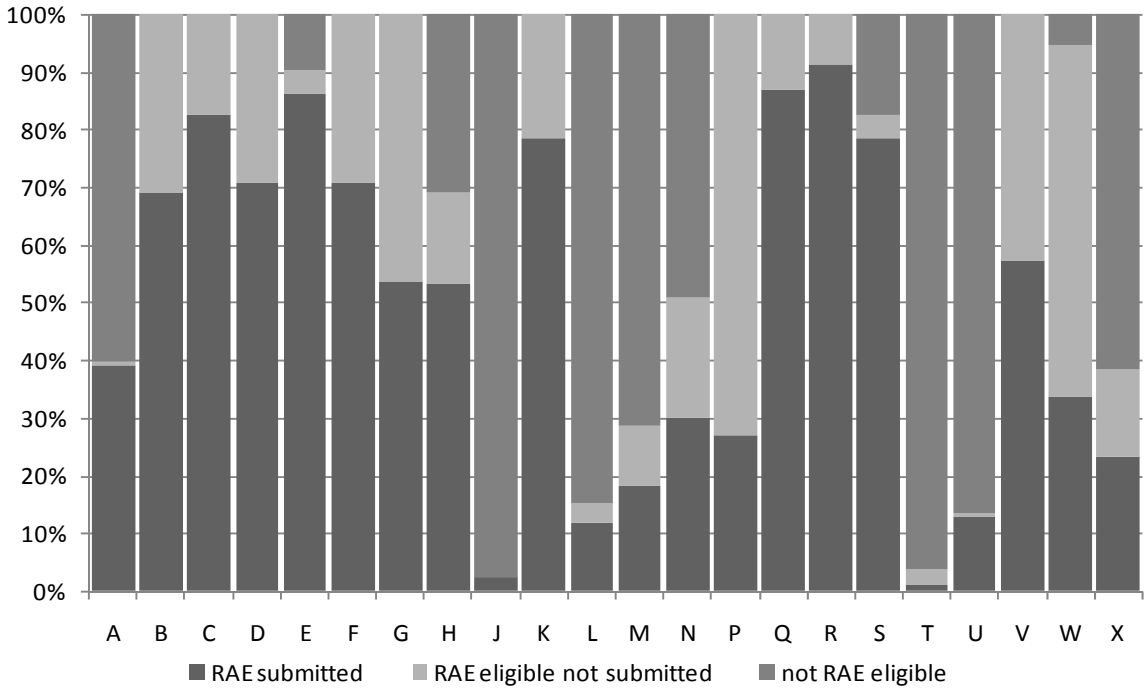
Figure E3 Percentage of RAE-eligible staff with missing HESA staff ID



RAE-eligible staff

Four pilot HEIs initially failed to supply data in this field. They were contacted to check which staff were eligible and their records were then updated accordingly (Figure E4).

Figure E4 Percentage of staff records for RAE-eligible and RAE-submitted staff by institution



Overall, 50% of records supplied were for RAE-eligible staff. Ten pilot HEIs supplied records (staff or output data) only for RAE-eligible staff. Of the remaining 12 pilot HEIs, the RAE eligible as a percentage of total staff records supplied varied from 15% to over 90%. This wide variation would suggest the application of differing criteria for inclusion of staff in each pilot HEI's dataset. For

example, some pilot HEIs were able to supply staff data for very large numbers of post-doctoral researchers who were very likely to be co-authors on publications (which would help with author disambiguation), but who would not have been eligible for inclusion on an RAE submission. Other pilot HEIs either did not choose to or were unable to supply such data, or did not interpret the data request as being so wide-ranging.

The number of RAE-eligible staff per pilot HEI varied from 92 to 6,154. The cleaned database contains records for 22,134 RAE-eligible staff across the 22 pilot HEIs.

Unit of Assessment

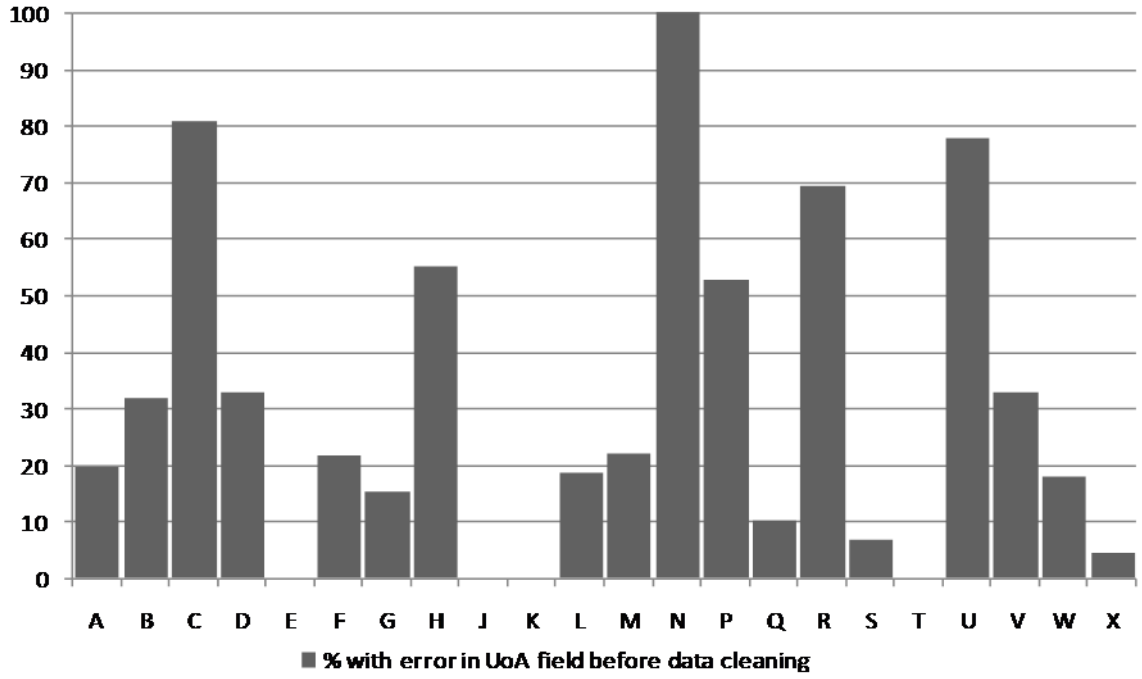
For the Unit of Assessment field, we requested a two-digit entry, corresponding to one of the REF pilot UoAs. We were clearly unable to assign staff to a subject area without this information from the pilot HEIs. This link, added to the author-staff links, enables the collation of publications to subjects.

Only 57% of the initial records had a valid entry in the UoA field. A further 8% were blank or "0", while the rest of the entries had been incorrectly formatted.

After data cleaning, and by iterative consultation with pilot HEIs, we ultimately achieved a valid entry in the UoA field for each member of staff who was RAE eligible.

Only staff in pilot UoAs were included in later stages of the project. About 95% (21,045) of the staff submitted were in pilot UoAs, while other staff initially submitted were later withdrawn as they worked in non-pilot subject areas.

Figure E5 Percentage of staff records supplied with errors in UoA field



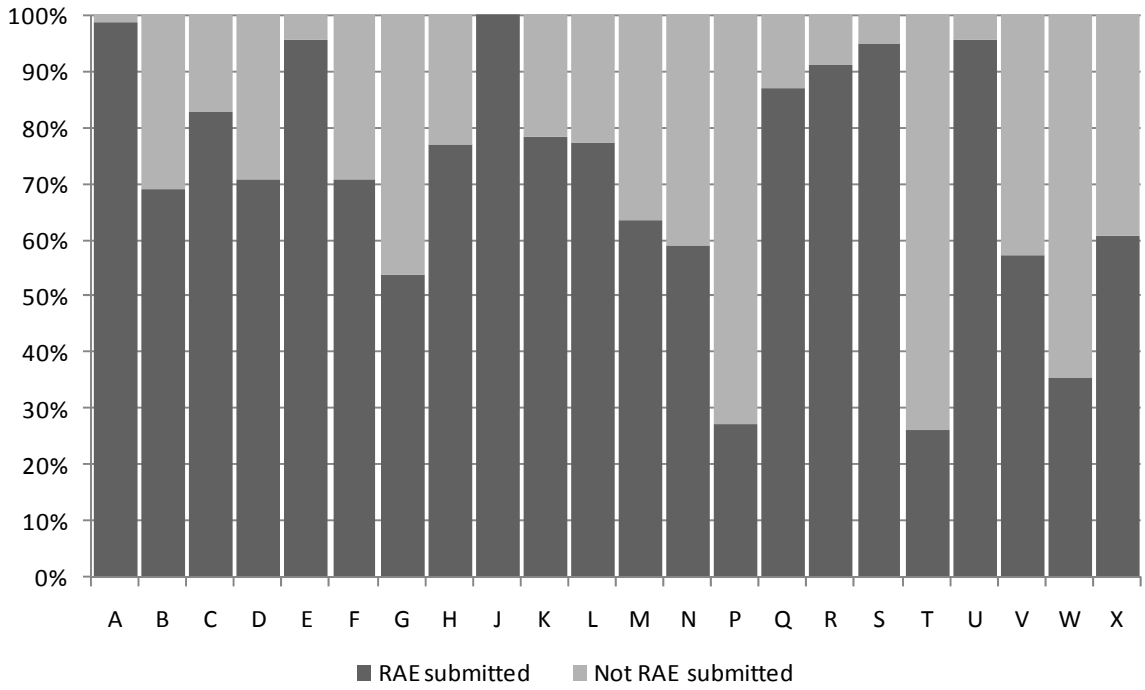
The breakdown by UoA (for REF pilot UoAs only) is shown in Table E9.

RAE-submitted staff

The RAE submitted field was required as it would be used in later stages of the project analyses to select staff for inclusion in particular model variants.

The proportions of RAE-eligible staff who were flagged as RAE submitted varied across pilot HEIs from 26% to 100% (Figure E6). This proportion reflects the varying approaches pilot HEIs took to submitting RAE eligible staff. We are not in a position to judge, from these data, whether pilot HEIs included all of their RAE-eligible staff in the data they supplied for this pilot.

Figure E6 RAE-submitted staff as a percentage of RAE-eligible staff



RAE staff category

The RAE staff category field (Category A, B, C, D) was required as it would be used in the later stages of the project analyses to select staff (and hence their outputs) for inclusion in particular model variants.

All pilot HEIs provided valid data in this field (for their RAE-eligible staff), apart from one where data were missing for around three-quarters of records and another where data were missing for about one-third of records.

Early career researchers

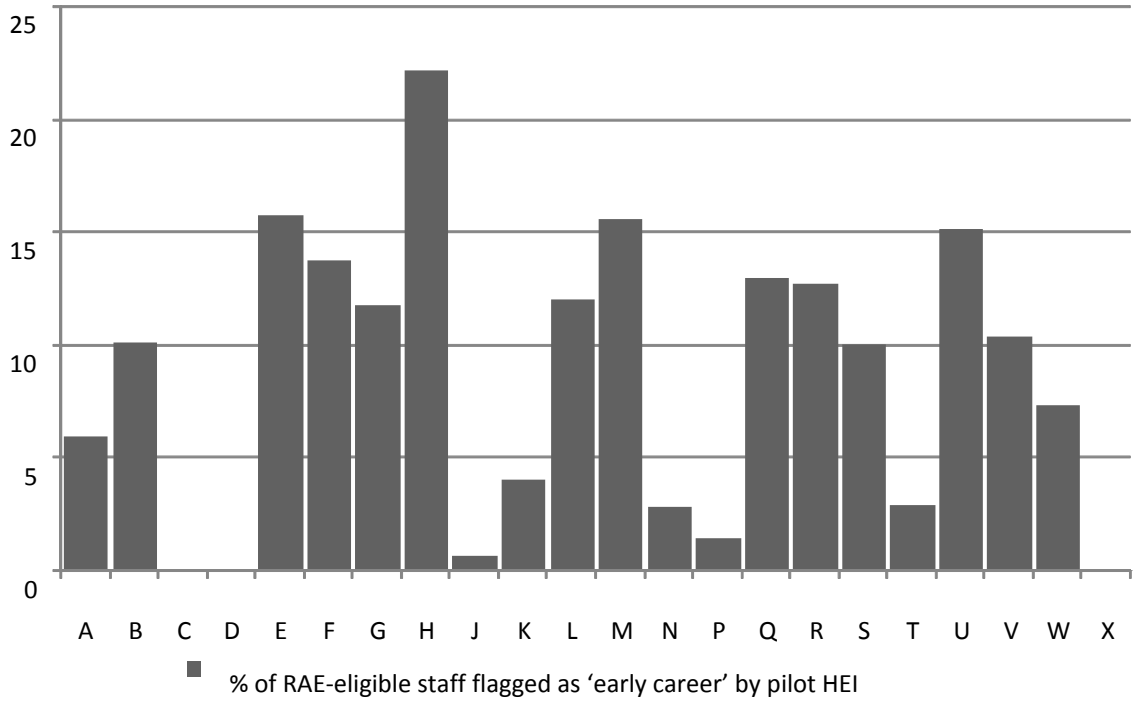
Most of the pilot HEIs produced data to identify which of the submitted staff could be interpreted as 'early career researchers' (Figure E7). Three pilot HEIs provided no relevant data.

This category is important because those staff who are at an early stage of their career might be reasonably expected to have a publication profile that is somewhat shorter and less frequently cited than their established colleagues. This might be detectable in variant analyses.

The percentage of RAE-eligible staff who were identified as 'early career researchers' varied across a fairly narrow range, from 10% to 15%. There was one high outlier at 22% and five pilot HEIs below 5%.

Such variation suggests that pilot HEIs' criteria used for including staff and/or for defining them as 'early career' varied, but did not do so in a seriously divergent fashion.

Figure E7: Percentage of RAE-eligible staff flagged as early career researchers



Start date and prior institution

The start date and prior institution fields were required as they would be used in the later stages of the project analyses to select staff (and hence their outputs) for inclusion in particular model variants.

The key issue here is about staff joining from another institution during the census period and whether their outputs prior to recruitment should be included in the profile of their current employer. Unless both start date and prior institution are known the data cannot be sieved to examine the effect of this.

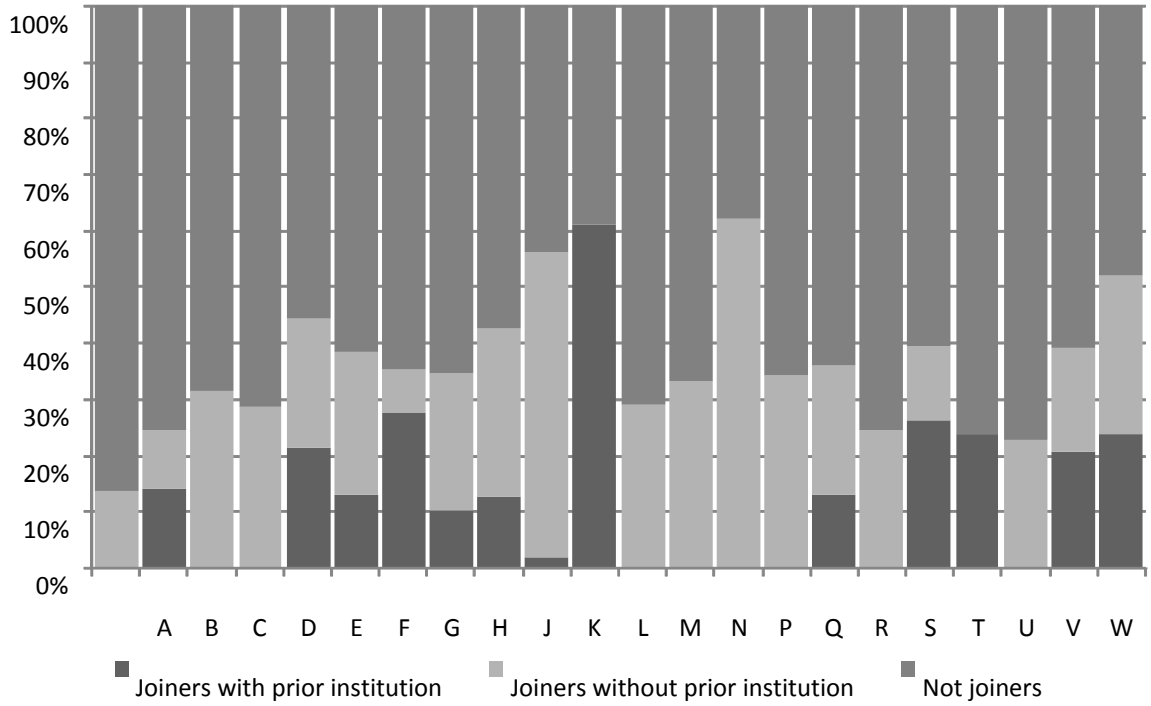
Relevant start dates (after the end of 2001) were provided for 40% of RAE-eligible staff. The actual percentage by pilot HEI varied from 14% to 62%. As was the case for other fields, discussed above, we suspect that the variation among institutions reflected not only a varying pattern of staff mobility but, at least in part, somewhat different criteria in the identification of RAE-eligible staff. An important third factor is, we believe, variation among institutional HR systems in the availability of data on start date.

For each staff record with a start date within the census period there should have been accompanying information on the prior institution at which that individual was employed. In practice, these particular

data were extremely scarce: there were 15,543 records with no data, 378 nulls, 109 with a HESA ID of 0000 and 76 'not known'.

There were 302 name and code variants for subsequently identifiable prior UK HEIs, which covered some 1,680 staff. Of these, the most frequently recorded were leading research institutions.

Figure E8 Availability of prior institution for 'joiners' (staff with start dates after 2001)



Note that Figure E8 shows which records have a non-null entry in the prior institution field. Of these records, only one-third contained data in the specified format (four-digit HESA institution identifier). The rest contained data that were not immediately usable and required data cleaning. In a significant number of cases the problem arose because the prior organisation was not a UK HEI, either because it was not an HEI or because it was an overseas university.

Nine pilot HEIs produced no data on prior institutions. Twelve of the pilot HEIs were able to produce a spread of information across UoAs.

Overall, only a small number of UoAs had a significant cluster of staff for whom prior institution data were available across a reasonable number of the pilot HEIs. No UoA had a concentrated sample across many pilot HEIs.

Given the high proportion of joiners without prior institution, we asked pilot HEIs whether they would be able to address this particular deficit given time. A frequent response was that the data were only kept in hard-copy files and that a person-by-person manual search and data entry would therefore be required. Any analysis of the effect of including data on publications associated with prior institution will therefore need to be based on this sub-sample.

If this information is required for an evaluation system to be implemented nationally, then institutions will generally need not only to amend their current practice but also to review their currently held records and update them.

Table E9 Final tally of usable records for RAE-eligible staff, by pilot HEI and pilot UoA

HEI	Total staff	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	32	34	40	43	44	46
A	**343**										11	36	84	19	10			34		4		19		18		3		1	16		23		3		43	19
B	**1,242**	85	68	66	63		46			20	55		17		142		126	53	56	69	39	21	25	45	57			23	57		23	38	7		41	
C	**477**													66	49				38	29	16	25	32	31	35		21	27	67					26		15
D	**1,274**	44	74	70	110		21	32	30	37	43				79			33	45	72	43	31		47	37		41	46	38	35	54	33	53	20	67	39
E	**521**														49			38	53	82	22	32	13	25		40					67		59		41	
F	**1,354**	51	78		73		32	61			58	52			199		13	107	63	59	36	40	16	53		78		36	64		74		50		61	
G	**451**							113							39		66	53	34					15	24								31		50	26
H	**1,783**	111	73	164	285		56		13	83					149	22	1	22	73	148	31	54	23	67	64	85	56	68	96	39						
J	**167**											21					35	23						10								9	28		28	13
K	**176**																176																			
L	**92**				17							22												13		24					16					
M	**774**		45		15	9	25	17	50	10	28	24	16	31	55	28	26		50	71	15			37	43			37	39		30		28		45	
N	**437**			57			293	87																											
P	**6,154**		222	211	1791		213	50	222	535	147	9	76		500	331		75	192	419	42	43	36	182	108		74	90	59		103	123	26		275	
Q	**441**				44										42	44			40	40		26		51		33					34	20	11		56	
R	**1,194**			20	106	25	31		7	48		54	28	59	57	67	124		44	52	26	58	12	32	29	28		40	52		46	53	48		48	
S	**944**					69									383		50	67						56		183	47			42				47	
T	**349**												193											60		96									
U	**152**		112												40																				
V	**1,703**		43		210			42	10			211	122		68	6	1	109	57	65	29	21	47	67	75	1		46	170		61	63	100		68	11
W	**860**				58		6	51			8	126	75		64		37	86				11	12	37	38			25	32		49		98		47	
X	**157**						23	41				12	36	2	17			4														3	5		14	
	21045	**291**	**715**	**588**	**2772**	**103**	**746**	**494**	**332**	**733**	**350**	**567**	**647**	**177**	**1942**	**498**	**655**	**704**	**745**	**1110**	**299**	**381**	**216**	**846**	**510**	**571**	**239**	**439**	**690**	**116**	**580**	**342**	**547**	**46**	**931**	**123**

Readers should refer to the main text (Table 4.1) for pilot HEIs' names (Table E9: left column) and Table 4.2 for a list of the UoAs (Table E9: top row by number).

It will be seen from Table E9 that some UoAs were widely populated but others were relatively sparse. For example, UoA 14 (Biological Sciences) was submitted by 17 of the pilot HEIs with almost 2,000 staff in total. By contrast, UoA 16 (Agriculture, Veterinary and Food Science) was submitted by 11 of the pilot HEIs, of which two submitted just a single individual. However, for the Royal Vet(erinary College) this latter UoA was the only UoA submitted and contained all the pilot HEI's researchers.

It is also evident by inspection that some UoAs generally had relatively few researchers and were submitted by few pilot HEIs. For example, UoA 5 (Other Laboratory Based Clinical Subjects) had only 103 RAE-eligible staff from just three pilot HEIs.

This pattern of staff distribution and concentration has implications for further analysis, because sparse data may be subject to relatively greater influence from outliers in publication and citation data.

18 ANNEX F, Chapter 8: Output data

Most institutions were able to supply substantially more data records than they originally estimated from their RAE2008 returns, reflecting the large amount of material available within institutions, and which provides a context for the evaluation analyses. It should be recognised that most pilot HEIs will have focused on maximising the returns on journal articles and that further material would become available in a full national exercise where institutions had a longer period for preparation.

Output type balance

Particular note was taken of the balance of output types, since this provides a context for the evaluation of journal articles for which citation impact can be calculated (see main text for discussion).

Table F1 Output balance (total count and % by type, ranked by % journal articles) in REF pilot HEI submissions

	Count	Books	Articles	Proceedings	Patents	Reports	Other
Robert Gordon	2019	4.3	47.2	0.0	1.7	3.4	43.4
Bournemouth	1803	9.5	59.4	24.3	0.0	4.8	2.0
Southampton	36873	12.7	60.0	20.8	0.0	0.0	6.4
Nottingham	37944	13.6	66.8	15.4	0.2	2.2	1.8
Stirling	4652	12.8	67.1	10.4	0.0	5.5	4.2
Bangor	4584	9.9	68.5	17.0	0.4	4.3	0.0
UEA	3068	9.0	71.1	19.6	0.0	0.0	0.4
Bath	18545	7.9	74.1	16.6	0.0	1.2	0.1
Sussex	6605	10.4	74.4	11.8	0.4	0.5	2.4
TOTAL	**327964**	**7.5**	**76.9**	**12.1**	**0.2**	**1.0**	**2.2**
Imperial	54389	5.0	76.9	16.3	0.7	0.8	0.2
UCL	47685	8.1	77.1	10.6	0.2	1.1	2.9
Durham	9427	8.7	77.8	7.5	0.1	0.8	5.0
Leeds	8534	2.8	81.0	11.8	0.2	0.0	4.1
LSHTM	9117	8.4	81.8	5.7	0.0	1.7	2.4
Glasgow	21520	5.0	82.7	11.6	0.0	0.1	0.5
ICR	3928	4.7	83.3	3.9	0.0	0.0	8.0
Queens	7471	0.0	84.8	15.2	0.0	0.0	0.0
Plymouth	6232	9.7	86.6	3.3	0.0	0.3	0.2
Portsmouth	494	0.8	91.3	4.5	0.0	0.0	3.4
Cambridge	24652	2.3	95.9	0.0	0.6	1.3	0.0
Royal Vet	1873	0.6	98.7	0.6	0.0	0.0	0.1
Birmingham	16549	0.0	100.0	0.0	0.0	0.0	0.0

Import strategies

Individual import strategies were developed to collate all the outputs from the various pilot HEI Excel (and other format) templates into a uniform database (Access format) with the following fields:

EvidUniqueOutputID – created for all pilot HEIs by preceding the InstitutionalUniqueOutputID with the HESA code as more than one pilot HEI used the ID number 1

Institution – HESA code with preceding 0, without H-

InstitutionalUniqueOutputID – as supplied

Year – as supplied

OutputType – as supplied

LongTitle2 – as supplied but with leading/trailing spaces/punctuation removed

ShortTitle2 – as supplied but with leading/trailing spaces/punctuation removed

Pagination2 – as supplied but with leading/trailing spaces/punctuation removed

Publisher2 – as supplied but with leading/trailing spaces/punctuation removed

DOI2 – as supplied but with leading/trailing spaces/punctuation removed

ThomsonUT – as supplied but some corrupted values resolved

ISBN – as supplied

Table F2 Summary of output type 'D' data received from pilot HEIs

Pilot HEI (anonymised)	**A**	**B**	**C**	**D**	**E**	**F**
InstitutionalUniqueOutputId	all	all	incomplete	all	all	all
Year	incomplete	incomplete	incomplete	all	all	incomplete
OutputType	incomplete	all	all	all	all	all
LongTitle (OutputTypeD)	incomplete	incomplete	incomplete	incomplete	incomplete	all
ShortTitle (OutputTypeD)	incomplete	incomplete	incomplete	incomplete	incomplete	incomplete
Pagination (OutputTypeD)	incomplete	incomplete	incomplete	incomplete	incomplete	incomplete
Publisher (OutputTypeD)	incomplete	incomplete	incomplete	incomplete	incomplete	incomplete
PublicationDate	incomplete	incomplete	incomplete	all	incomplete	incomplete
Thomson UT	no data	incomplete	no data	no data	incomplete	no data

FieldName	**G**	**H**	**J**	**K**	**L**	**M**
InstitutionalUniqueOutputId	all	all	all	incomplete	all	all
Year	incomplete	incomplete	incomplete	all	incomplete	all
OutputType	all	all	incomplete	all	incomplete	all
LongTitle (OutputTypeD)	all	all	incomplete	all	all	all
ShortTitle (OutputTypeD)	incomplete	incomplete	incomplete	incomplete	incomplete	incomplete
Pagination (OutputTypeD)	incomplete	incomplete	incomplete	incomplete	incomplete	incomplete
Publisher (OutputTypeD)	all	incomplete	incomplete	all	incomplete	all
PublicationDate	all	incomplete	all	all	incomplete	incomplete
Thomson UT	no data	no data	incomplete	incomplete	no data	incomplete

Pilot HEI (anonymised)	**N**	**P**	**Q**	**R**	**S**
InstitutionalUniqueOutputId	all	all	all	all	all
Year	all	all	incomplete	all	all
OutputType	incomplete	all	all	all	all
LongTitle (OutputTypeD)	incomplete	incomplete	all	incomplete	all
ShortTitle (OutputTypeD)	incomplete	incomplete	incomplete	incomplete	incomplete
Pagination (OutputTypeD)	incomplete	incomplete	incomplete	incomplete	incomplete
Publisher (OutputTypeD)	incomplete	incomplete	incomplete	incomplete	all
PublicationDate	all	incomplete	all	all	incomplete
Thomson UT	incomplete	no data	no data	no data	incomplete

FieldName	**T**	**U**	**V**	**W**	**X**
InstitutionalUniqueOutputId	all	all	all	all	all
Year	all	all	incomplete	incomplete	incomplete
OutputType	all	incomplete	all	incomplete	all
LongTitle (OutputTypeD)	all	incomplete	incomplete	incomplete	all
ShortTitle (OutputTypeD)	incomplete	incomplete	incomplete	incomplete	incomplete
Pagination (OutputTypeD)	incomplete	incomplete	incomplete	incomplete	incomplete
Publisher (OutputTypeD)	all	incomplete	incomplete	incomplete	incomplete
PublicationDate	incomplete	all	incomplete	incomplete	incomplete
Thomson UT	no data	incomplete	no data	incomplete	no data

After data receipt, the following cleaning regimes were applied to database fields.

Year

The format requested was a single year, within the range 2001 to 2007.

Some outputs submitted were outside this range and others had no publication year given. Missing data were updated, where possible, from the PublicationDate field.

Following this step, data were missing in this field for less than 0.5% of outputs type 'D'.

Journal titles

The format requested was full journal title.

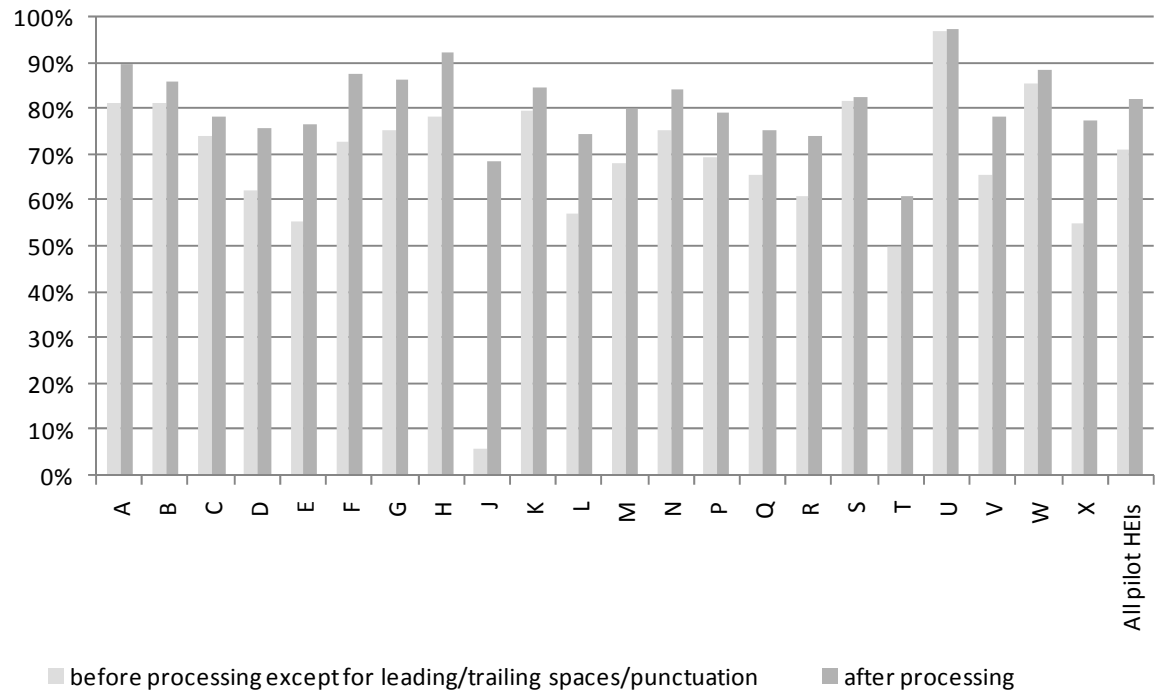
The data received were very diverse. They included full titles, Thomson standard abbreviations, commonly used abbreviations and non-standard variants, and typographical errors.

Data were missing in this field for less than 2% of outputs type 'D'.

The raw data as provided by the pilot HEIs contained over 31,000 journal title variants, of which some 25,000 occurred fewer than five times. A standardised journal name was added to each publication in the database using a customised list informed by previous *Evidence* work on similar datasets and work carried out specifically for the HEFCE REF pilot study. This list assigned variant journal titles used at least five times to either the standardised journal name or recorded the item as 'non-Thomson'. Standardised journal titles were also added using the ISSN where available.

These strategies reduced the number of journal title variants to around 6,500.

Figure F3 Percentage of outputs type 'D' with standardised journal title or non-Thomson designation by pilot HEIs before and after processing



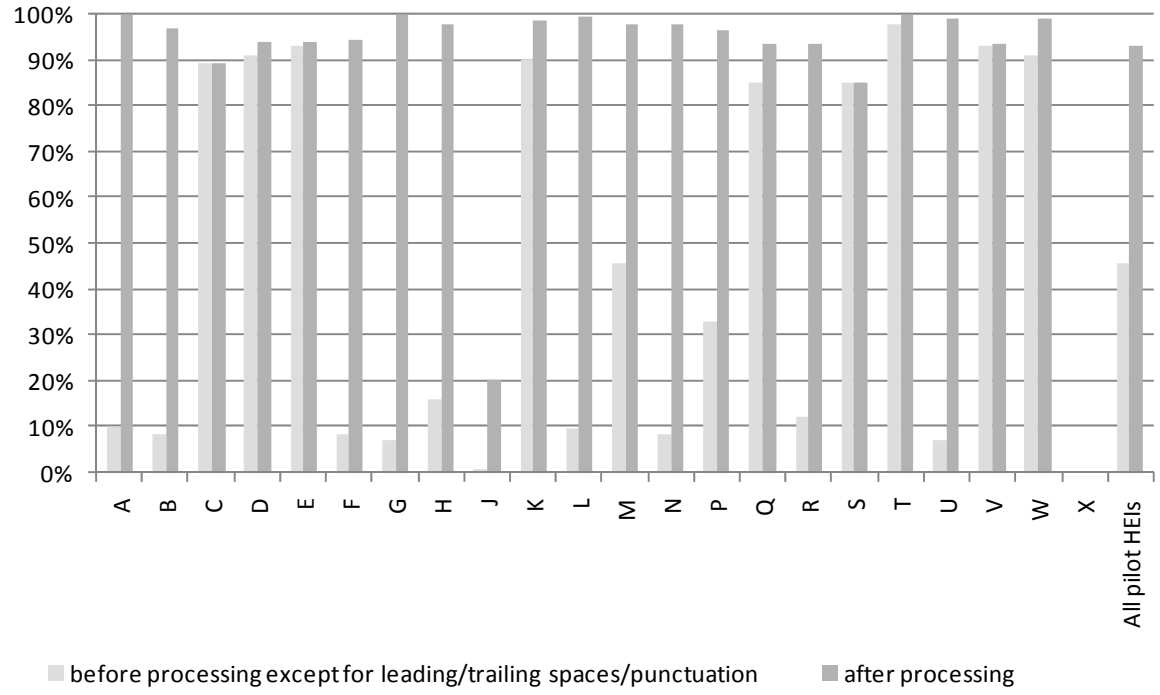
Volume

The format requested was “numerical only”.

The data received were of various formats, of which around 130,000 – just over 50% – were in the format requested. The most common deviation was the inclusion of the issue number of the journal, rather than simply the volume number. The issue was often denoted by parentheses () but also by /. The other common problem was that numerical volume data were preceded by text, usually either ‘vol’ or ‘volume’; this occurred in 3% of publication records.

Data were missing in this field for 4.7% of outputs type ‘D’.

Figure F4 Percentage of outputs type ‘D’ with volume number by pilot HEIs before and after processing



First page number

The format requested was “numerical only” for a single value indicating the first page of the article.

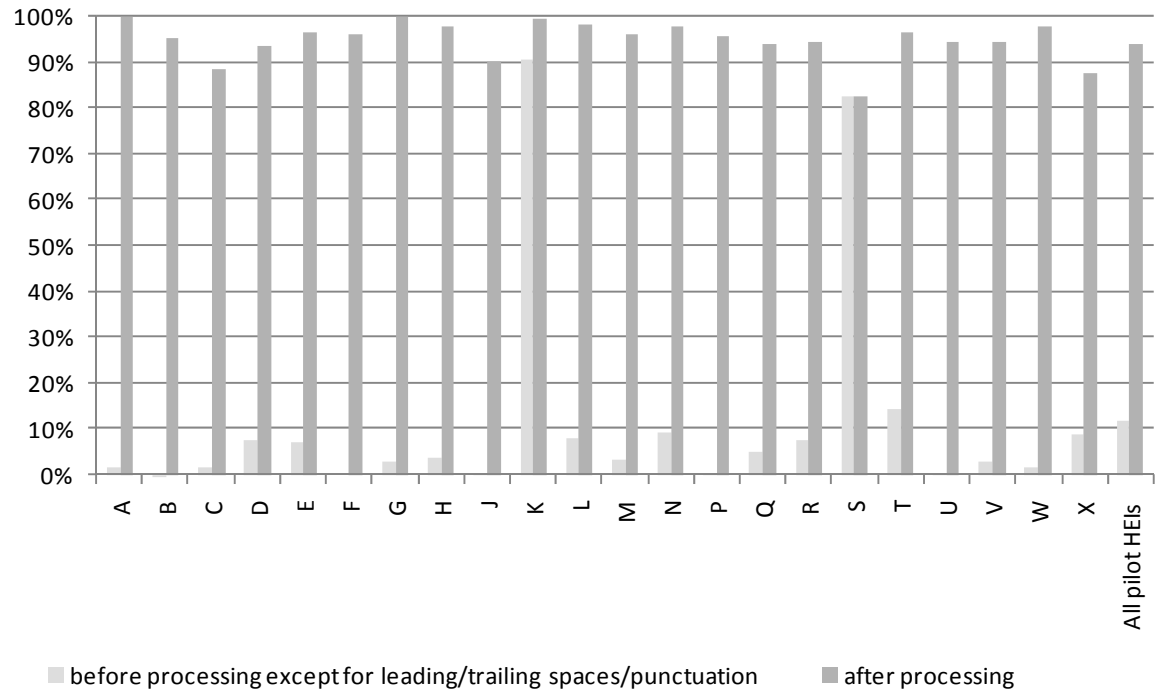
The data received were of various formats. Around 190,000 – over 75% – consisted of a page range rather than a single numerical value for the first page.

Experience has indicated that matching to commercial citation databases is simpler if the pagination is limited to the first page for the output. This reduces the problems of variable formats (e.g. pages 2045-2049 are often abbreviated to 2045-49 or even 2045-9). Using the ‘first page only’ methodology can lead to multiple matches, but almost without exception these would be restricted to very brief items: meeting abstracts, editorials and other short communications not used in typical bibliometric analyses. Where page ranges are submitted it is essential to impose strict formatting rules.

After page ranges, the most common problem was prefixing of page numbers with variants of p/p./pp/pp./page/pages. Such markers were found in around 5% of outputs. This is more of a problem than might at first be assumed, because some publications are not denoted by page numbers but by article numbers. This is especially true in physics journals. Article numbers are often prefixed with one or more alphabetic characters (around 3% in addition to the abbreviations for page).

Data were missing in this field for 4.7% of outputs type ‘D’. Although the numbers of outputs with missing data for volume or page numbers were similar, they were the same outputs in only around half of these.

Figure F5 Percentage of outputs type ‘D’ with first page number by pilot HEIs before and after processing



Article titles

This can be a useful field to validate matching strategies.

Apart from the cleaning at the point of importing, to remove leading/trailing spaces, no further cleaning was attempted except for one institution where all the articles were processed to remove xml tags derived from EndNote.

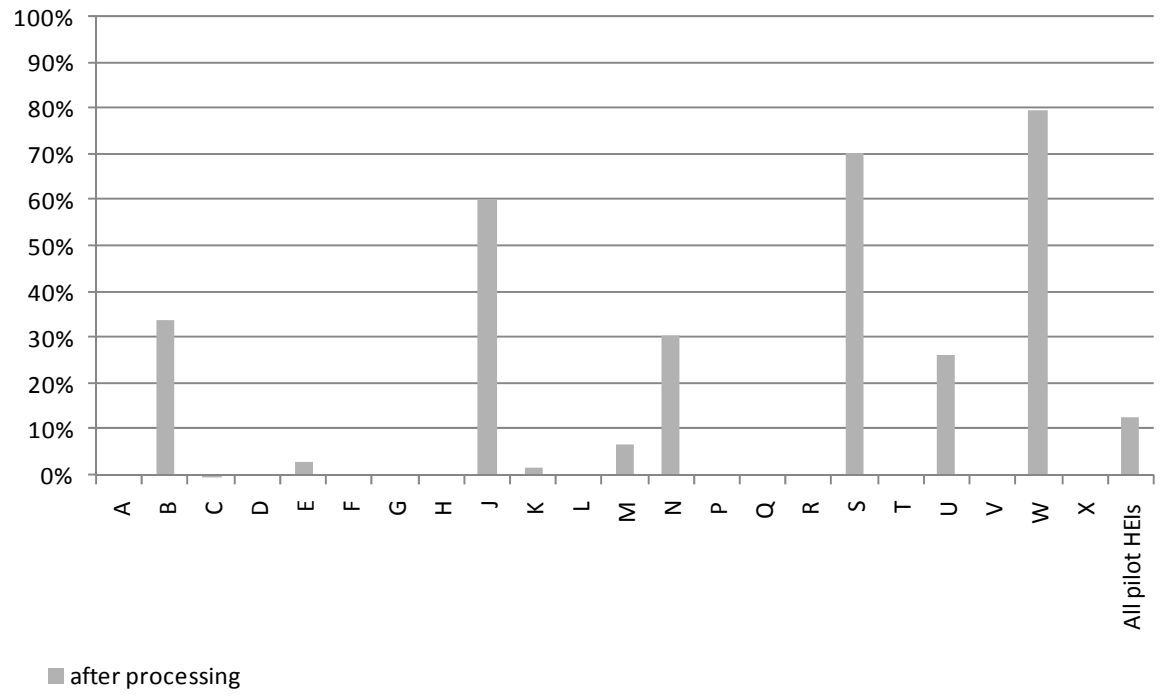
The first 40 characters of the article titles were used as visual checks for the matching strategies.

Thomson UT

The format requested was a string of 15 characters, which is the known format of the Thomson UT identifier.

This data was optional and was in practice included by only 10 pilot HEIs. Some cleaning other than removal of leading/trailing spaces/punctuation was done on these data. This was done principally to remove URL tags and to convert the data to 15 characters where this format had been lost during processing either by the pilot institution or in-house.

Figure F6 Percentage of outputs type 'D' with Thomson unique identifier (UT) by pilot HEIs after processing

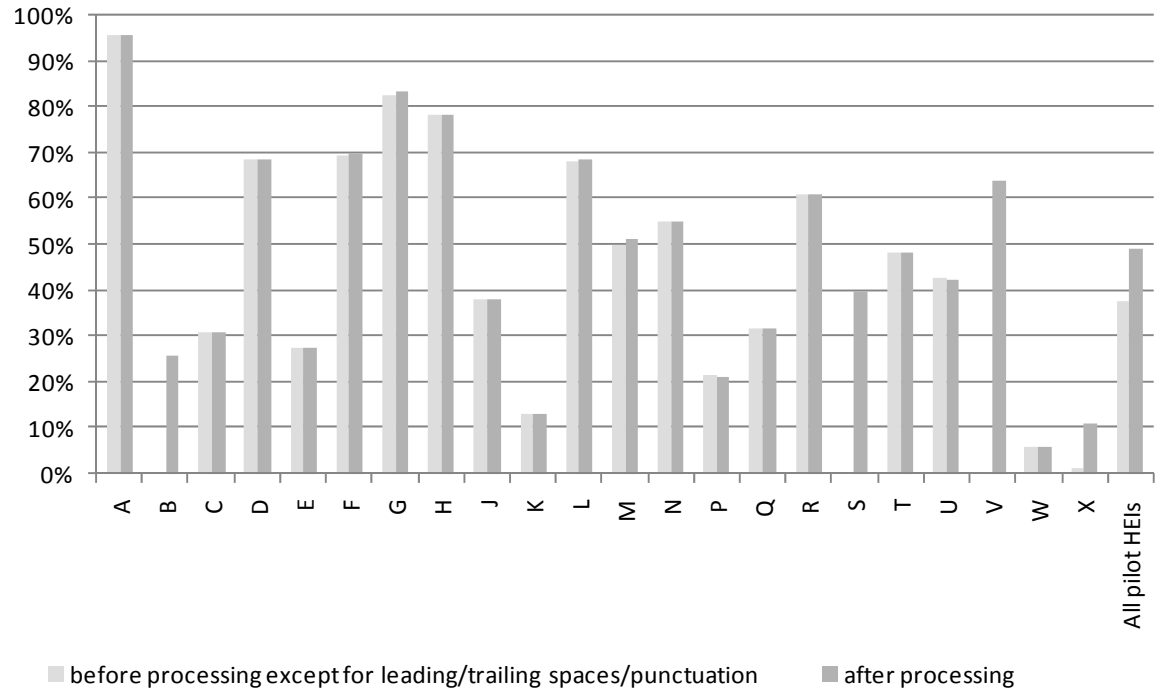


DOI

The format requested was a string of up to 256 characters.

Around 30,000 (11.3%) outputs included some HTML/URL tags preceding the DOI. After processing to remove these extraneous data, just under half of the outputs type 'D' had an associated DOI. In general, where provided, these data were apparently valid (validity was assumed if the data began '10.' as there is no other standard format).

Figure F7 Percentage of outputs type 'D' with digital object identifier by pilot HEIs before and after processing



19 ANNEX G, Chapter 9: Matching strategies for Thomson Reuters citation database

Match using DOI data

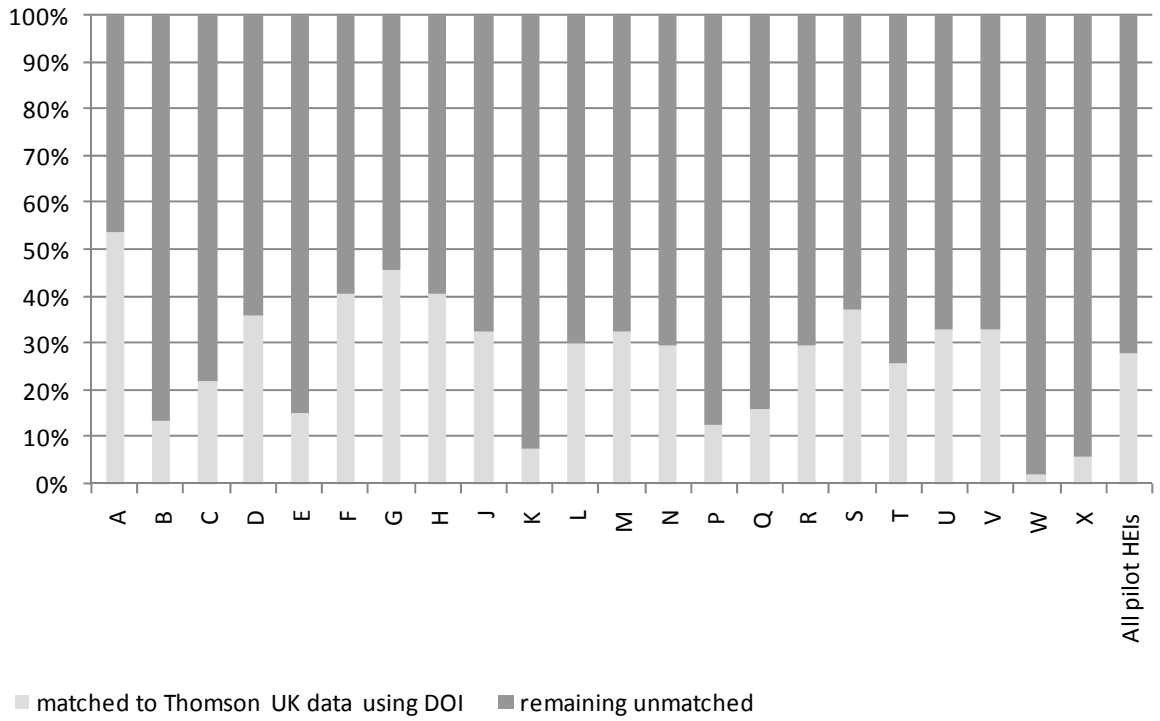
In the absence of a commercial citation database unique output identifier, the next most precise and simple method to assign the citation data from any database would appear to be via the digital object identifier, or DOI, for the outputs at the time of data submission from the pilot HEIs. These data were part of the submission for RAE2008 and consequently it was expected that these and some additional outputs held by pilot HEIs would have those data. In fact, only just under half of output type 'D' records had DOI data of the format '10.[….]' and, of these, not all turned out to be both valid and accurate after cleaning.

Results indicate the variation in success of matching outputs type 'D' using DOIs. Success rates tended to be higher for some pilot HEIs, but there was no association with the proportion of records submitted which had the DOI field completed. There is therefore no indication that any institutions tended to pay particular attention to this field. Overall matching success was limited to around one-third.

Using DOI data the proportions of outputs type 'D' matched for three pilot HEIs were very low overall, despite these data being submitted for substantially more outputs. This presumably reflects an aspect of the way in which this field had been completed.

Considering all pilot HEIs, DOI data were available for 49% of outputs type 'D'; 57% of these outputs with DOI data were uniquely matched to Thomson data, giving an overall matching success of 28% or 70,147 outputs.

Figure G1 The percentage of pilot HEI output type 'D' records matched to Web of Science data using only DOI



The Thomson databases have only recently collected DOIs routinely (only 3% of articles and reviews in 2001 were abstracted with a DOI, increasing to 64% and 66% in 2006 and 2007).

Although it was not part of the matching strategy, *Evidence* visually checked some of the outputs where the DOI was linked to the Thomson UK-based citation database. More than 6,000 did not

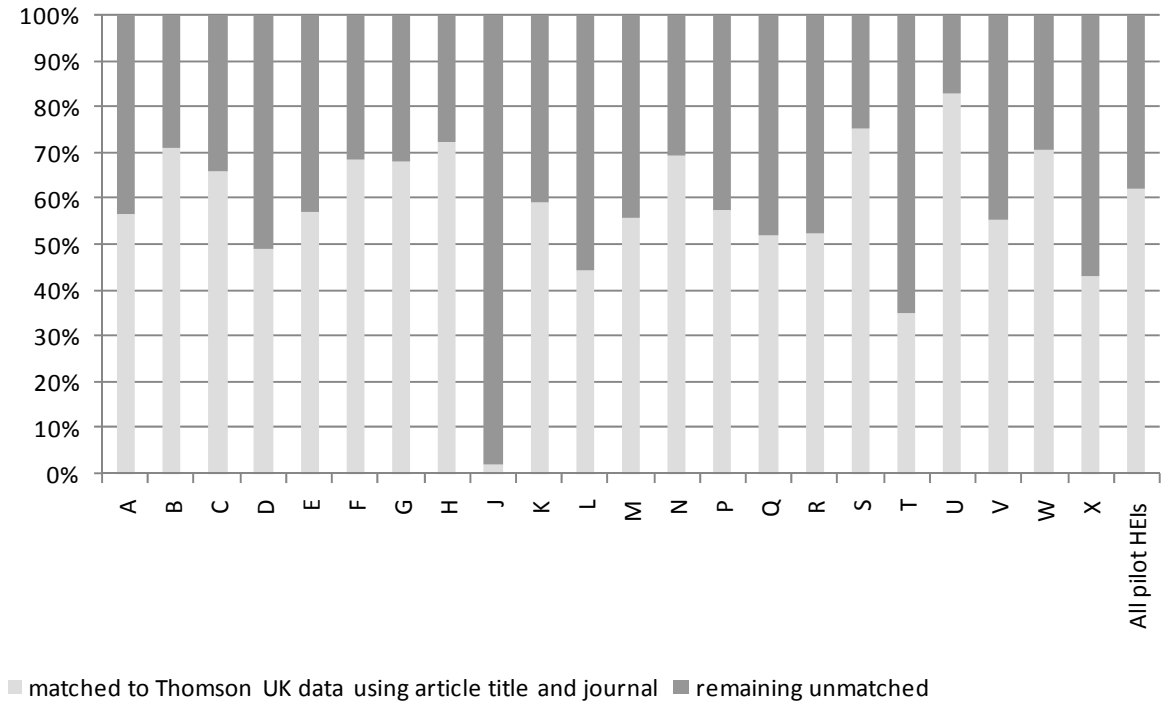
match on the truncated article title (40 characters). Although most of these could be visually confirmed as the correct match, there were more incorrect matches than using the UT data. Consequently, the submitted DOI data would cause the citation data from the wrong article to be assigned to an output. This was especially problematic for one of the pilot HEIs where citation data for a substantive research output were linked to an output which could not be verified or was found as a meeting abstract.

Match using article title truncated to 40 characters and journal data

Although the data for article titles could be much less clean than other data, this would have led to matches not being made rather than incorrectly matching data. It also used only two fields of data, so was the next preferred matching strategy for most bibliographic data. These data were provided for the majority of outputs type 'D'; 85% of outputs type 'D' had some data in both these fields, but again not all were necessarily accurate and valid.

Figure G2 indicates the variation in matching outputs type 'D' to commercial citation data. Considering all pilot HEIs, journal and article title data were available for 85% of outputs type 'D'; 73% of these outputs with journal and article title data were uniquely matched to Thomson data, giving an overall matching success of 62% or 155,986 outputs.

Figure G2 The percentage of pilot HEI output type 'D' records matched to Web of Science data using only journal and article title



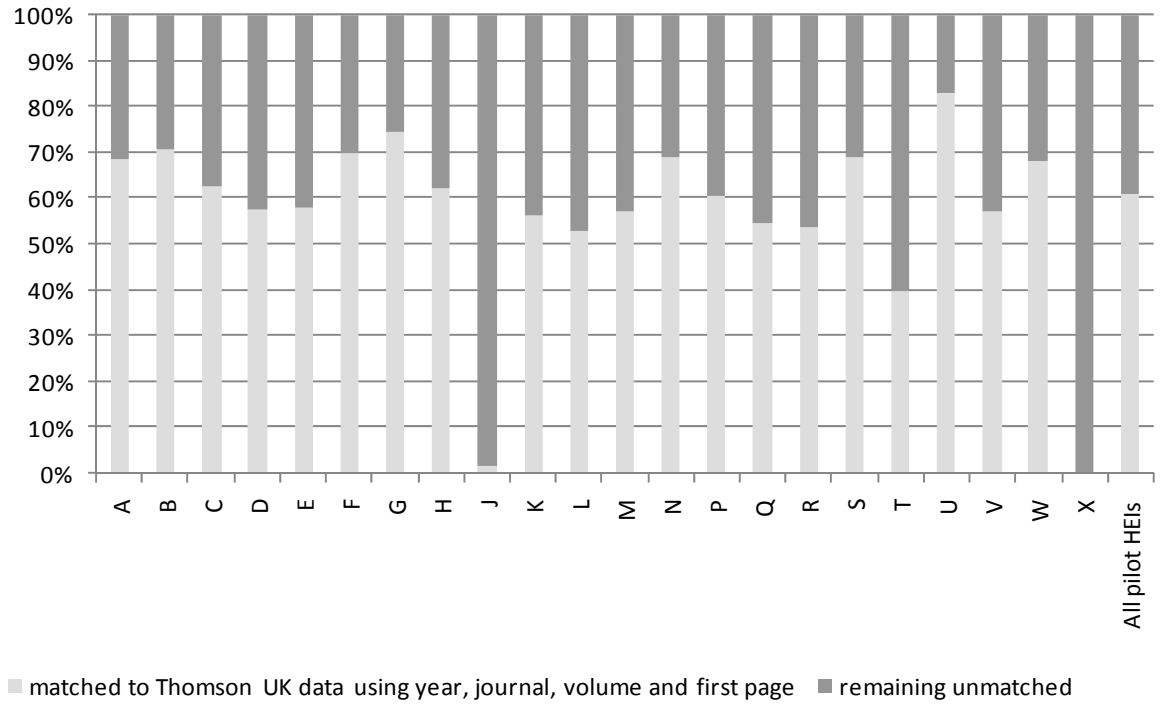
Match using year, journal, volume and first page number data

This was the least robust matching strategy used on the REF pilot outputs data. It used four fields of data, all of which had to be cleaned to a significant extent just to enable some matching to take place. These data were provided for the majority of outputs type 'D'; 85% of outputs type 'D' had some data in these fields, but again not all were necessarily accurate and valid.

Figure G3 indicate the variation in matching outputs type 'D' to commercial citation data. With one exception, the Thomson UK database matched more outputs to unique records than the Scopus global database. One pilot HEI provided no volume data for its publications, so no outputs were matched using this strategy. Another pilot HEI's dataset was extremely limited, in that very few outputs had journal data; consequently, the overall matching success was limited to less than 2%.

Considering all pilot HEIs, year, journal, volume and pagination data were available for 78% of outputs type ‘D’ and these were uniquely matched to Thomson data, giving an overall matching success of 61% or 152,440 outputs.

Figure G3 The percentage of pilot HEI output type ‘D’ records matched to Web of Science data using only year journal title, volume and first page number



20 ANNEX H, Chapter 10: Summary of issues arising from data gathering and processing

As noted in the main report, it should be anticipated that many of the issues arising were a consequence of the compressed timetable, the lack of time for institutions to interpret the project requirements and prepare data to respond to the staff and outputs specifications, and the main focus of activity being in pilot HEIs over the summer. There was also a fundamental over-optimism about the state and content of research data management systems in institutions, which led to a much more onerous task of data cleaning than had been expected.

In full national implementation the longer time for preparation and response will overcome many of the issues below. However, by airing these in detail now it is anticipated that institutions will generally be able to anticipate some of the likely changes needed.

Two other reports on the REF pilot process were commissioned. One was commissioned by the Joint Information Systems Committee (JISC) from Stuart Bolton Associates and is available on JISC's web-site (www.jisc.ac.uk/). The other was commissioned by HEFCE from Technopolis Group and will be available on the REF web-site (http://www.hefce.ac.uk/Research/ref/biblio/).

Specific aspects of data gathering

Table H1 General issues arising from the process of pilot HEI data submission to the REF pilot exercise

Issue	Comment
Multiple submissions	The pilot process allowed multiple submissions of information. This was intended to be additive, working towards completion. In practice, with varying delivery dates, it became confused. Some submissions overlapped, creating multiple partial versions which then had to be collated and deduplicated, preserving always the most complete version of any record. Version control therefore became a problem, so that the accurate identification of records became a challenge. Multiple submissions should obviously be avoided in full implementation, but in a compressed pilot exercise this may be infeasible.
Assumptions as to how the data would be managed	Although the management of data from 22 pilot HEIs, with tens of thousands of staff and hundreds of thousands of outputs, might seem to imply that data management would be electronic, this was not always recognised in the submission process. It was evident that assumptions had been made that some data variants would be visually reviewed and 'similarities' would then be appreciated, or that 'obvious errors' would be picked up. In fact, such visual checking should have been redundant, as the expectation that data would be accurate and complete was made clear in briefing meetings.

Issue	Comment
Divergence from template	The contractors had a clear picture of the form in which they would most readily be able to manage the data submissions. It was intended that this would be reified in the data template presented and circulated to pilot HEIs, but that communication was not always effected. In a one-to-one operation with a single organisation this vision would have been readily developed and communicated. The justification for uniformity in data presentation became obscured, however, when many pilot HEIs were involved and the technical importance of using the same set of fields and field specifications was lost. Some pilot HEIs provided data UoA by UoA, which needed consolidating. Some pilot HEIs simply submitted that data they had to hand which approached the outline the contractors had requested, while two pilot HEIs submitted data that differed significantly from the structure of the template. It was evident that the significance of a data template to effective data management had been completely misunderstood. Good data management required well-identified single data values. As a challenge to such data management, an example of the more complex data treatments which the contractors were required to create was to strip multiple data-points out of a single cell from a pilot HTML document into new records in Excel.
Relict data	The REF pilot had a well-defined census period, starting in 2001. For the pilot analysis to be appropriate, data were required for staff records and for outputs that covered years from the start of the census period. This should have been feasible since it matched the recent RAE2008 census period. In practice, many pilot HEIs have had a change in their systems in the meantime, making older staff data difficult or infeasible to access. Where data was accessible, it could be stored in a problematic format where codes in use at the time (such as the UoA subject structure of the RAE2001) have subsequently been modified.
Field order	For data from many institutions to be handled (visually checked or electronically processed) in a coherent way, it was essential that pilot HEIs should present their data in the same set of fields and, preferably, using the same labels for those fields. Many pilot HEIs returned their data in ways that diverged from the template. Commonly, fields were renamed, missing or in a different order, or new fields – unfamiliar to the contractors – were inserted. The consequence was that much time had to be spent matching the observed data to the expected template.
Default values	Although a specification for the data to be inserted was identified, default values were not defined for every field. This was, with hindsight, an error on the contractor's part. This led to problems because when missing values were encountered their meaning was unknown. This was particularly problematic when dealing with nulls and zeroes: they might represent quite different things and therefore needed to be distinguished.

Issue	Comment
Missing values	Data which were implicitly required were sometimes absent. For example, if a staff member was coded in one field as RAE Category B, they should have had a leaving date in the corresponding appropriate field, yet such data were frequently absent.
Data entry errors	It was apparent that far too much data was entered manually, either by typing or by cutting and pasting from another source. Both processes introduced errors, particularly where the cut-and-paste carried over superfluous characters, including punctuation and non-printing characters that were only detected when errors appeared. Mistyping even a single character or digit could lead to problems; for instance, data entry errors sometimes meant that there was more than one institution identifier for a given institution. It is essential that data are checked and validated in institutional databases prior to submission. They can then be accessed systematically and error-free for any internal or external reporting purpose.
Data appearing in the wrong field	The appearance of data in an incorrect field (e.g. year data in a volume field) sometimes came about because of poorly formed CSV. This particularly affected fields containing text strings where the text included commas. Where a comma was found, everything after it was pushed into the next field, displacing that field's content into the next, and so on. Occasionally data had been repeatedly entered in the wrong field. This was particularly an issue for publications data, where data were required in different fields according to output type. It had been anticipated that the methodology would be familiar, if not automated, from the recent RAE submission, but this was generally not the case.
Non-printing characters	Some fields contained 'unseen' characters, such as soft returns, or leading or trailing spaces. These inevitably generated unexpected results prior to detection, especially during matching procedures.
Xml (Extensible Markup Language)	xml presented particular issues. A notable problem was produced by exporting from EndNote libraries. These will not read into Excel or Access in a predictable fashion. For example, in Access each field is split into a table, whereas in Excel there is an addition of extraneous fields and blank records.
Excel	Some data were handled in unhelpful ways in Excel, particularly dates, large numbers and numbers separated by hyphens. Without changing the underlying data, Excel sometimes converts large numbers into scientific notation, and dates into date serial numbers. It was the numbers separated by hyphens which caused the most problems, however, as Excel assumes these are dates and converts them into a date. This changes the actual value within the cell; for instance, "1-2" becomes the date 01/02/09: 1st of February (1st of the 2nd of the current year). This was a frequent problem in pagination data.
Access	Import specifications work only for text files. In order to force Access to import Excel data as text, it was sometimes necessary to insert blank rows at the start of Excel tables.

Issue	Comment
EndNote	Exports from EndNote in xml proved very difficult to process. Neither Access nor Excel handled xml data properly, particularly when the xml was generated from EndNote. Again, very large numbers (such as those in HESA staff identifiers) were found to be especially problematic in this context. Other issues included HTML tags surrounding data in particular fields, particularly the LongTitle field. Occasionally the only content in a field was HTML tags. EndNote libraries provided 'as is' were unproblematic, though time-consuming to transfer to Access.

Staff data in Table 1

Data cleaning would have been significantly reduced if data had been submitted through a restrictive interface with built-in rules to disallow invalid or incorrectly formatted data, and to validate within and between fields. For the purposes of the REF pilot, however, we felt that such restrictions would have increased the burden on pilot HEIs, possibly to the point where they were unable to supply data in the time available.

Institution

This field was designed to contain each institution's unique HESA institution code so it could be used to assign records to pilot HEIs in the central database. The data collection form should have been designed to allow one record per person, with mandatory and validated fields for HESA staff identifier and HESA institution identifier.

We had to amend the field for five pilot HEIs, usually where the leading zero had been omitted. For one institution, staff had been assigned to HESA codes from other pilot HEIs, while for another the field was either blank or contained just two digits.

Staff institutional ID

For one institution there was a considerable amount of duplication (approximately 500 out of 2,700 records were duplicates). This occurred because records had been extracted from more than one system, resulting in two or more records per person. A considerable amount of work was required to condense these data into single records per person.

The problem of managing staff institutional IDs was exacerbated by the fact that there was no common field with which to identify each staff member. Records which apparently applied to the same staff member had staff IDs in different formats (e.g. one version with leading zeroes, one without). We overcame this by matching electronically on surname and first initial and using visual checks to identify typographical errors in the surname field and false matches on common surnames.

Dates of starting and leaving

The data template specified the date format as YYYY-MM-DD, but date formats varied not only among pilot HEIs but, in some cases, within a single pilot HEI's dataset. The most common variants were: nn.nn.nn; nn/nn/nn; nn/nn/nnnn and nnnn/nn/nn, where n is an integer.

The start and leave date fields should have been empty where staff were Category A, but one institution created a default leave date for each member of staff, which was entered as 1900-01-01.

All staff identified as Category B or Category D would require a leaving date, correctly formatted and within the census period. All staff with a leaving date would require a destination and all staff with a joining date would require a prior institution. Where the destination or prior institution was a UK HEI, this field would have to contain a valid HESA institution identifier.

For dates, formatting variants in the data field may appear minor, and a number of staff pointed out that a visual inspection would readily reveal what the field should contain. Nonetheless, a considerable amount of cleaning work was required to ensure that dates were formatted consistently. Visual inspection with the volume of data in an exercise of this kind, and a consistent response, was infeasible in the time available.

Subject area

Each staff record should have been attached to a subject area, for which the UoA was used as a convenient substitute.

The UoA field should have been a two-digit text field requiring entries to be between 01 and 67.

Confidentiality

Confidentiality issues around staff data added to the effort required to clear discrepancies and missing data.

Staff data had to be maintained in protected files and kept on the secure server. This worked well for the initial submission of data, but became irksome where there was a significant requirement for iteration with pilot HEIs over minor queries.

Table H2 Issues arising from management of data in specific fields in the Table 1 staff data

Issue	Comment
Institution	The institution code should have been a four-character value starting with a zero. This zero was sometimes missing. In other instances there were no data at all in this field.
HESA staff identifier **[YY][institution ID][institutional staff ID][check digit]**	All academic members of staff in UK HEIs should have a HESA staff identifier which uniquely identifies them and travels with them within the UK HE system. In practice, one person might have several HESA staff identifiers. The identifier should comprise 13 characters in the form indicated. Thus, 0501249876541 would be a staff member who arrived at Leeds (0124) in 2005 (the initial 05) with an internal ID of 987654 and a check digit of 1. As well as missing values and badly formed values (typically fewer than 13 characters), there were a number of duplicate values that appeared to refer to different members of staff.
Institutional staff identifier **Pilot HEIs took different approaches to this**	Sometimes the identifiers formed part of the absent HESA staff identifier (see above), while others used separate codes and still others were without values. Where present, these values were non-unique across the system, so we needed to append the institution ID to the code. There were several duplicate values, which sometimes referred to the same member of staff (this could be valid, as staff may have left and returned), but had sometimes been re-used and applied to different staff. Where we were given more than one record for a member of staff (perhaps because some information we asked for was held in legacy systems), the institutional staff identifier differed between the records.

Issue	Comment
Unit of Assessment	This should have been a two-digit number, with a leading zero where necessary. The leading zero is why UoAs should, as requested in our specification, have been provided as text values. Where leading zeros were missing, they had to be re-established. Some information, particularly older information retrieved from legacy systems, used RAE2001 UoA codes, which needed translating to 2008 codes. There were a number of missing values. Pilot HEIs presently have to provide the UoA of academic staff to HESA, but some pilot HEIs had trouble assigning UoAs to staff to whom they had not previously needed to assign UoAs.
Unit, department, school or other location	We discovered that 'location' would have been better collected as a separate field, to ensure that the data returned were more consistent.
Initials	No guidance was given as to how initials were to be separated (i.e. with spaces, full stops, commas or with no separation). Data received ranged across the entire spectrum. Because this was a field used for critical name matching, rigorous cleaning was needed.
Alias or 'known as' for publications	Data in this field were sparse. This was unfortunate. Where available, such data can be very helpful in disambiguating synonyms.
Email address	It was intended that this might be used as a user-name for the Symplectic Publications system, but it was agreed that asking researchers to confirm their own publications was going to be infeasible in the period of the pilot project. These data raised particular data protection implications.
Submitted to 2008 RAE	These binary data were coded in a number of different ways (yes/no, true/false, 1/0) which then had to be standardised.
RAE eligible	Half of the pilot HEIs returned data only on eligible staff. For the half that supplied a more complete and informative dataset, eligible as a percentage of the total ranged from 15% to 68%. This is a surprising disparity and suggests differing interpretation of the criteria for inclusion. This indicates a need for careful standardisation of staff definitions for REF implementation.
RAE staff category	These should have been coded A, B, C or D, as for the RAE. A wide range of approaches were taken. Some pilot HEIs provided Category A staff only, while others included almost all their staff and many postgraduate research students. Extraneous codes were present and it was discovered on validation that some data had been incorrectly entered. For instance, if a member of Category A staff left after the RAE deadline, and was given a leave date, they were listed as Category B.
Early career researcher	This category seemed to be a challenge to interpretation. Some pilot HEIs were not able to provide any relevant data from their existing systems, except insofar as it related to RAE-submitted staff.

Issue	Comment
Start date (if started after Jan 2001)	Data were returned in many different date formats, some of which were not interpreted as dates by Excel and which therefore had to be visually checked and systematically amended. A number of individuals joined during the period, left and then joined again. We considered setting up the template to capture these, but decided it would be too complex. Note that this cyclical employment was raised as an issue by some pilot HEIs.
Prior institution (if started after Jan 2001	Many pilot HEIs supplied no data or were only able to find sparse records. For many pilot HEIs this was (and will be) a manual task, involving going back to hard-copy CVs in HR files. Where present in plot submissions, these data were often not properly formed. We were expecting valid HESA institution IDs, as for the pilot HEI's own HESA ID, but in fact the range of information was in such diverse formats where it was provided that the text and IDs had to be manually resolved.
Leave date (if left before Dec 2007	See start date.
Destination institution	See prior institution.

Output data in Table 2

As a pilot study, it was anticipated that the pilot HEIs would not always have their output data held in the ideal and appropriate format for the project. For this reason, the specification stayed as close as possible to the format that all institutions had been required to use for RAE2008 submissions. It was expected that this part of any output record would be provided in a consistent and accurate form.

The quality of output data is of paramount importance to the future application of bibliometric methodology to the REF. The points below are intended to indicate where the quality of the data would need significant improvement:

- InstitutionalUniqueOutputID must be unique. Some pilot HEIs supplied data with duplicated InstitutionalUniqueOutputIDs. Sometimes these were not duplicates of the same publication but wholly different publications that had been assigned the same ID number.
- Problems arise in maintaining ID numbers when moving between bibliographic management software such as EndNote or Reference Manager and other formats such as Excel. HEFCE and the pilot HEIs need to be aware of this.
- Pagination formats must be standard. There should be no extraneous characters in page (such as pp, pg), volume (no extraneous characters such as vol.) or article titles (no quotation marks).
- Journal names used by institutions should use standard or commonly agreed formats (such as Lancet or The Lancet).
- Validation will need to work both ways between Funding Council and institutions. Data, or large samples of data, will always need to be mutually validated at some point within the system.
- Unique article identifiers, such as Thomson UTs and DOIs, were incorrectly supplied by pilot HEIs. If the errors had not been identified this could have led, in some cases, to a serious mismatch between publication and citation data.

Receiving bibliographic data from HEIs

Here, as elsewhere, there is a critical question as to whether earlier and more rigorous identification of data that did not meet the published specification (e.g. where journal names or pagination departed from the required format) should have been applied. This would have led to much data being

returned to the relevant pilot HEI, amended and then resubmitted later, but this would have reduced the significant additional time spent by the contractor. HEFCE will need to decide how the data will be linked to any commercial database and then in implementation prioritise the critical fields to ensure that submitted data are as clean as possible.

Table H3 Issues arising from management of data in specific fields in the Table 2 outputs data

Issue	Comment
InstitutionalUniqueOutputId	Similar to institutional staff identifier. Where these data were missing, codes had to be created.
Year	There were many missing values. Where present, the data sometimes indicated years from outside the census period.
Output type	This field frequently included extraneous codes and miscoded items. For example, outputs type 'D' (journal articles) were also submitted as articles, journal articles, reviews and so on. Visual scanning was required to reinterpret the labelling.
Long title	There were extensive spelling mistakes, the inclusion of practically identical titles for quite different items, and errors thrown up by the inclusion of HTML tags.
Short title	This field contained characters other than numeric for output type 'D' (volume/edition).
Pagination	Although necessarily numerical, this field contained a wide range of extraneous characters.
Publisher	For journal articles, this field is used in the RAE data specification for the journal title. There were extensive journal abbreviation issues, the appearance of leading and trailing spaces and/or full stops. There was an inconsistent use of '&' and 'and'. Many spelling mistakes were detected, e.g. 'Agroforesty Systems' not 'Agroforestry Systems'. There was insufficient information to disambiguate some journals, and this was especially noticeable in journals with multiple parts with which the authors should have been familiar, e.g. Biochimica Biophysica Acta, Acta Crystallographica. There was a widespread use of non-standard abbreviations, e.g. BMJ. Some journal acronyms were non-unique and had to be disambiguated. There was an inconsistent use of leading 'The', e.g. The Journal of Urology, The Lancet.
Editors	Data confused by separators: commas, semi-colons, etc
ISBN or ISSN	Pilot HEIs used many non-standard formats.
Publication date	This field was confused by different date formats and a range of dates from outside the census period.
DOI	This field seems to be very problematic for many institutions at present. Data were frequently incorrect and non-standard.
Is interdisciplinary	From RAE specification – not used by Evidence
Is sensitive	From RAE specification – not used by Evidence

Issue	Comment
ListOfAllAuthors	This was made difficult to interpret by a variable use of separators and a diverse set of such separators.
Indexed by Thomson Reuters	Frequently absent, so presumably not usually stored in institutional systems.
Thomson unique identifier (UT)*	Frequently absent, so presumably not usually stored in institutional systems.

21 ANNEX I, Chapter 11

Table I1 Fields associated with UK NCR data used for Symplectic reconciliation with data submitted by pilot HEIs: output data

Field name	***Data type***	***Notes***
EvidUniqueOutputID*	nvarchar(255)	Unique identifier with respect to total database of all articles (journal and non-journal) held by *Evidence*
Year	Int	Year of publication
LongTitle	nvarchar(255)	Full title of output
ShortTitle	nvarchar(255)	Abbreviated title of output
Pagination	nvarchar(255)	Full pagination details
Publisher	nvarchar(255)	Name of publisher
ThomsonUT	nvarchar(255)	Thomson unique identifier
DOI	nvarchar(255)	Digital object identifier
EvidJRNL29	nvarchar(29)	*Evidence*-rectified 29-character journal name
EvidVolume	nvarchar(255)	*Evidence*-rectified journal volume
EvidFirstPage	nvarchar(255)	*Evidence*-assigned first page number
EvidAssignedUT	nvarchar(255)	*Evidence*-assigned Thomson unique identifier

*primary key

Table I2 Fields associated with UK NCR data used for Symplectic reconciliation with data submitted by pilot HEIs: link data

Field name	***Date type***	***Notes***
HESACode	nvarchar(255)	Institutional HESA code
HESANumber	Int	Numeric part of the HESA code
YBPreferredName	nvarchar(255)	Standardised institution name
EVIDORG	nvarchar(255)	*Evidence* organisational identifier
FullName	nvarchar(255)	Full name of organisation
Abbreviation	nvarchar(255)	Abbreviated name of organisation
REFPilotShortName	nvarchar(255)	Name for purposes of REF pilot project

Table I3 Fields associated with UK NCR data used for Symplectic reconciliation with data submitted by pilot HEIs: address data

Field name	*Date type*	*Notes*
ISI_LOC*	nvarchar(15)	ISI article link
ORG	nvarchar(255)	Organisation
DEPT	nvarchar(255)	Department
LAB	nvarchar(255)	Laboratory
SECT	nvarchar(255)	Section
CITY	nvarchar(255)	City
PROVINCE	nvarchar(255)	Province
STATE	nvarchar(50)	State
ZIP_CODE	nvarchar(255)	Post/zip code
COUNTRY	nvarchar(255)	Country
EVIDORG	nvarchar(255)	*Evidence* organisational identifier

Figure I4 – Schematic representation of Tables I1, I2 and I3 showing their relationship

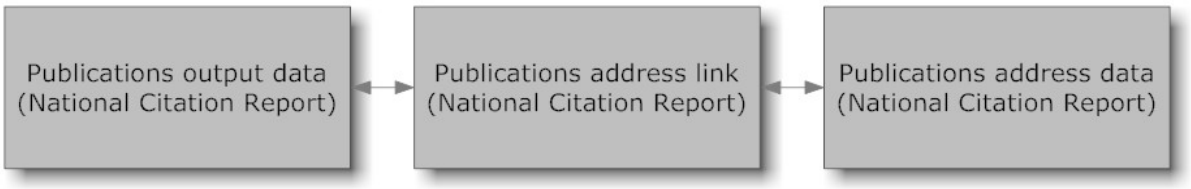


Table I5 Overlap between 'actual' institutionally supplied output type 'D' journal article data and 'presumptive' Evidence/Thomson Reuters Web of Science™ bibliographic data address-rectified for UK organisations

Pilot HEI	***Pilot HEI data***	***Thomson Reuters data***	***Overlap[2]***
Bangor	2575	2802	1470
Bath	7415	4943	3656
Birmingham	16084	14029	8441
Bournemouth	1037	522	203
Cambridge	18842	33240	10754
Durham	6625	7469	3824
UEA	2042	4256	452
Glasgow	15957	13090	9084
Imperial	40764	28861	20506
ICR	3222	2210	1809
Leeds	6843	12899	4204
LSHTM	7076	4697	3766
Nottingham	23456	12205	9427
Plymouth	5284	2961	2373
Portsmouth	448	1879	246
Queen's	5718	7213	3381
Robert Gordon	947	566	314
Royal Vet	1694	1378	913
Southampton	20435	13392	10013
Stirling	2827	2083	1693
Sussex	4545	5215	1965
UCL	35747	31558	19045
TOTAL	***229583***	***207468***	***117539***

[2] Those articles supplied in the databases provided by institutions which are additionally found in the *Evidence*/Thomson Reuters address-rectified data (referred to as presumptive data for pilot HEIs).

Figure I6 Plot of data in Table I5

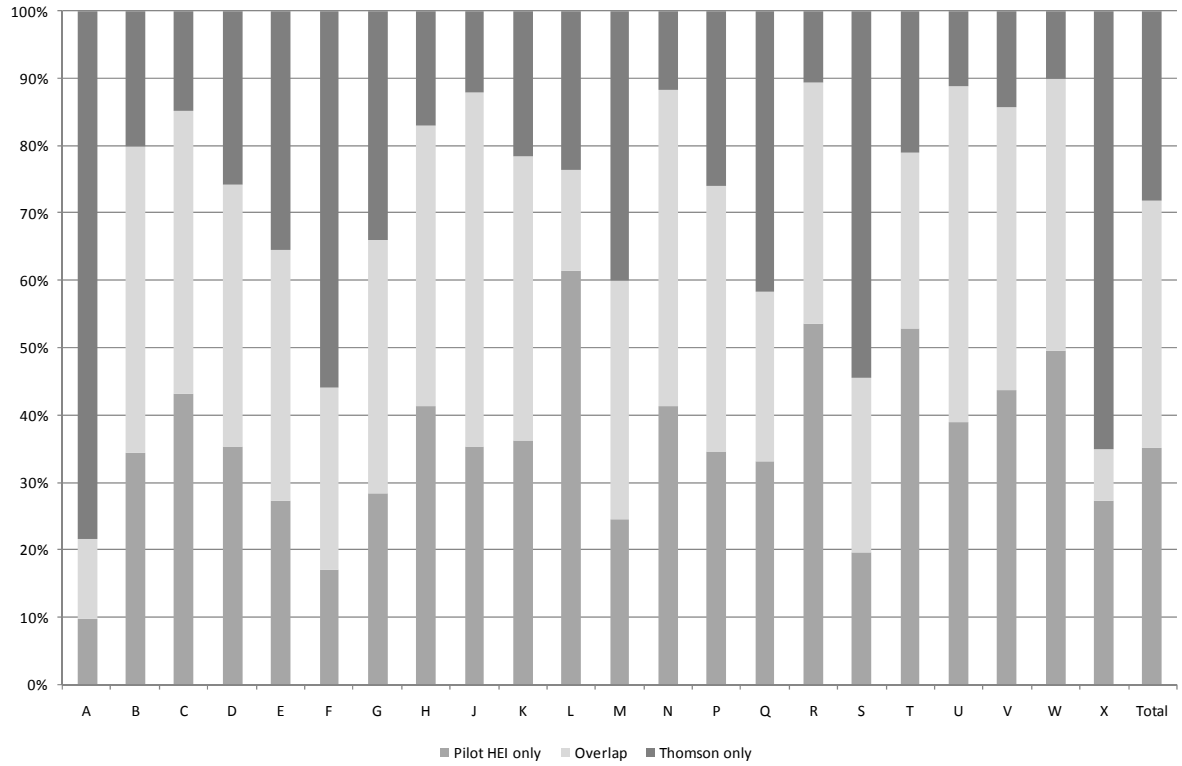


Figure I7 Relationship between presumptive data and amalgamated institutional journal data (see Table I5 for institutional breakdown)

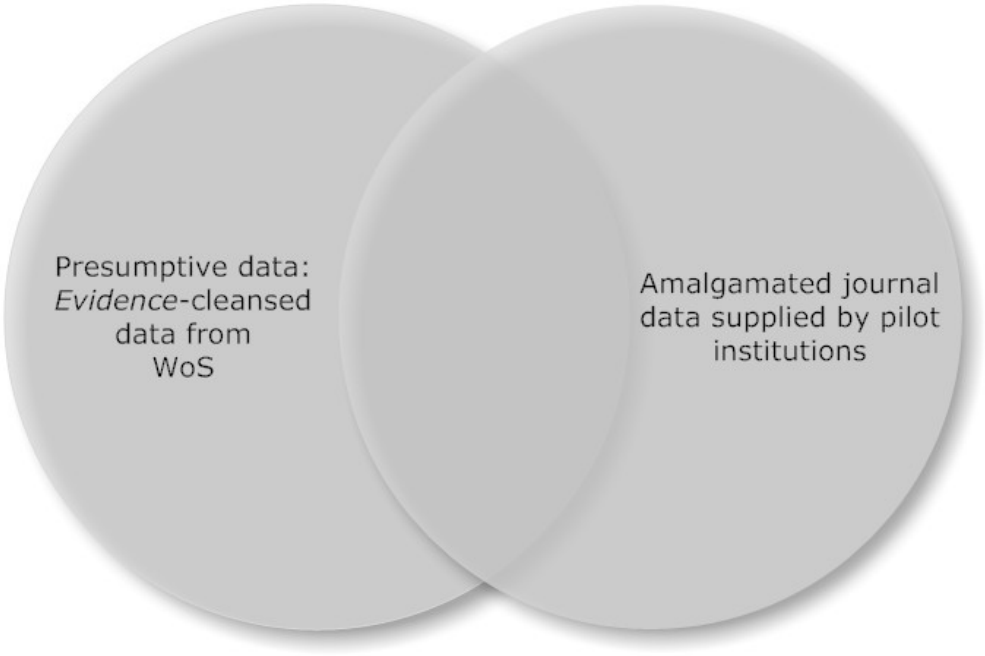


Figure I8 Process for matching publications to institutional staff

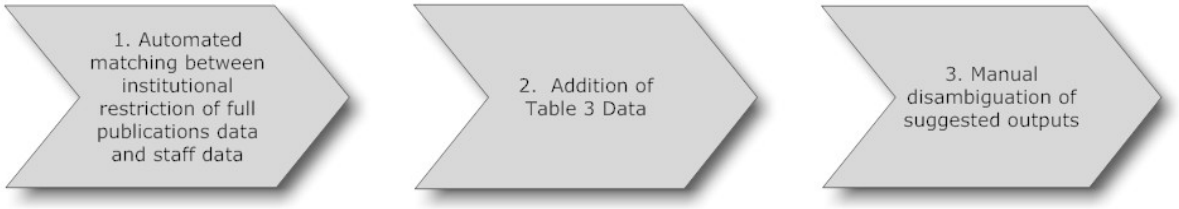
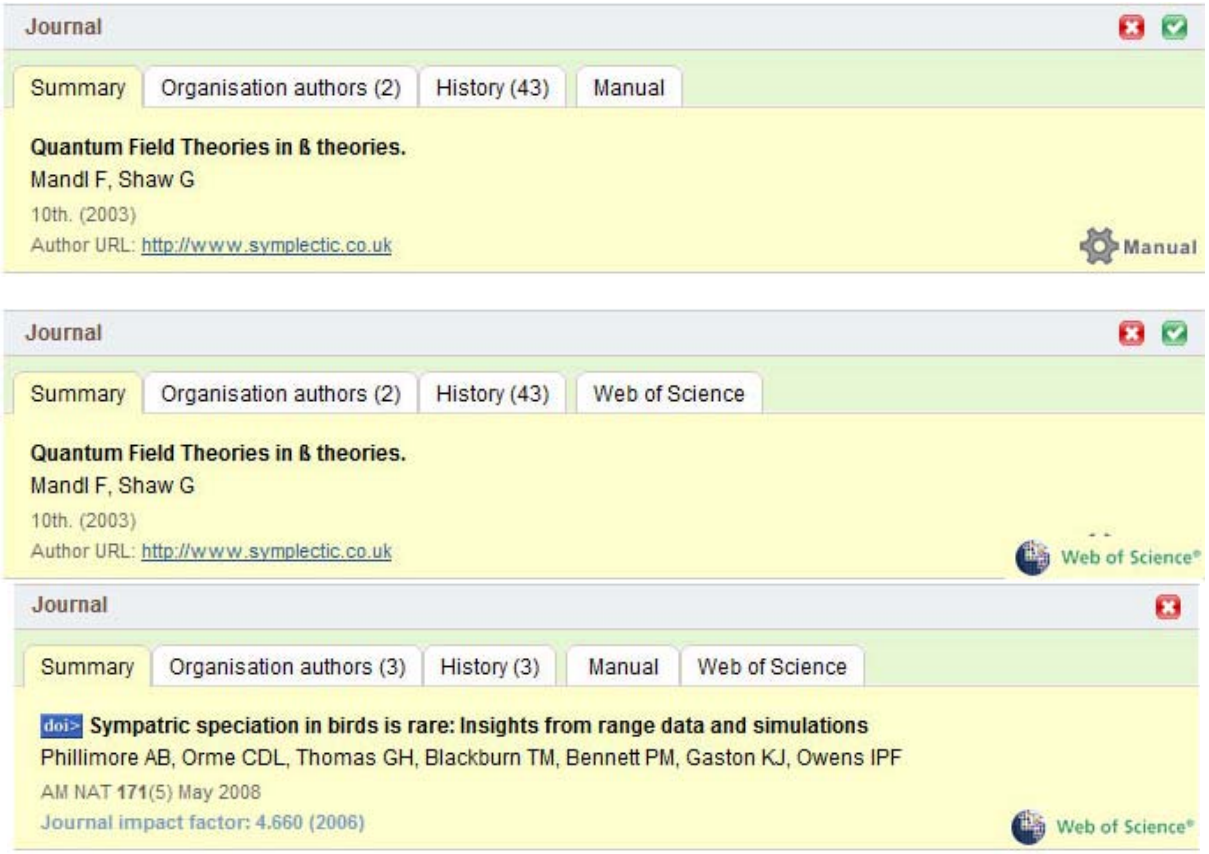


Figure I9 Different views of an article, depending on the source of the bibliographic data



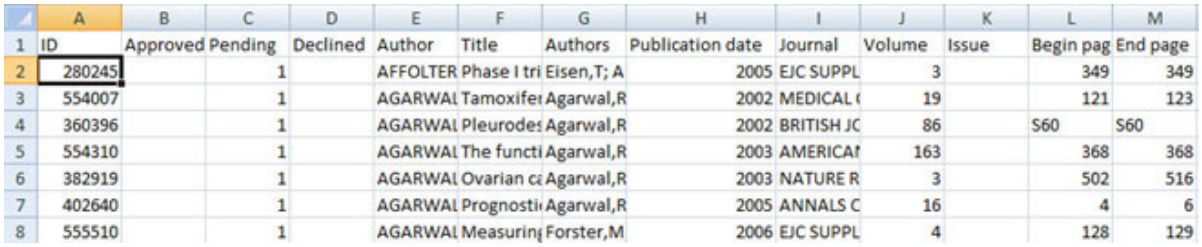
The first article record has only institutional data associated with it (hence only the 'Manual' tab is visible). It can be seen that there are two authors associated with the work – this corresponds to two authors across the whole pilot dataset (these authors may be the same person listed from different institutions or they may be different authors from the same or different institutions). Not all authors of an article may be contained in the dataset since they may not have been in a cohort associated with a pilot institution.

The second article shows the appearance in the Symplectic system of an output with an *Evidence*/Thomson Reuters data record but no institutionally supplied data.

The third picture represents an article having both institutional and *Evidence*/Thomson Reuters data.

On the top bar in each case it can be seen that the articles are journal articles (this may include sub-categorisations such as review, letter etc), and the two buttons on the right side allow the user to decide whether the publication is associated with them or not.

Figure I10 Spreadsheet with information for pilot HEIs, allowing them to approve or decline a suggested link by moving the '1' into the appropriate column.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ID	Approved	Pending	Declined	Author	Title	Authors	Publication date	Journal	Volume	Issue	Begin pag	End page
2	280245		1		AFFOLTER	Phase I tri	Eisen,T; A	2005	EJC SUPPL	3		349	349
3	554007		1		AGARWAL	Tamoxifer	Agarwal,R	2002	MEDICAL (	19		121	123
4	360396		1		AGARWAL	Pleurodes	Agarwal,R	2002	BRITISH JC	86		S60	S60
5	554310		1		AGARWAL	The functi	Agarwal,R	2003	AMERICAN	163		368	368
6	382919		1		AGARWAL	Ovarian ca	Agarwal,R	2003	NATURE R	3		502	516
7	402640		1		AGARWAL	Prognostic	Agarwal,R	2005	ANNALS C	16		4	6
8	555510		1		AGARWAL	Measuring	Forster,M	2006	EJC SUPPL	4		128	129

Table I11 Comparison between the count of links provided by pilot HEIs and those suggested from the Symplectic automated process

Pilot HEI	Links provided in pilot HEIs Table 3	Additional links suggested by Symplectic	Suggested links approved by pilot HEI	% of suggested links approved
Bangor	3021	2387	1890	79.2
Bath	7211	1025	770	75.1
Birmingham	22781	7331	1402	19.1
Bournemouth	1112	97	93	95.9
Cambridge	21283	3411	2208	64.7
Durham	7554	2455	2213	90.1
UEA	2649	364	119	32.7
Glasgow	18276	7171	3355	46.8
Imperial	54054	63078	23790	37.7
ICR	4554	2302	2159	93.8
Leeds	10307	9878	5675	57.5
LSHTM	9967	1063	679	63.9
Nottingham	28705	8362	3477	41.6
Plymouth	6658	521	258	49.5
Portsmouth	430	1401	980	70.0
Queens	8153	3672	3110	84.7
Robert Gordon	1275	316	294	93.0
Royal Vet	423	1475	1466	99.4
Southampton	25254	18489	4808	26.0
Stirling	3933	343	121	35.3
Sussex	5210	2342	2209	94.3
UCL	49926	38071	17229	45.3

The above table shows the number of links identified using the automated algorithm. Extremely large numbers of suggested links were developed for some of the pilot HEIs, and the contractors worked with those institutions to develop and employ strategies to bulk-approve or bulk-decline articles.