

May 2009

**Identification and dissemination
of lessons learned by institutions
participating in the Research
Excellence Framework (REF)
bibliometrics pilot**

**Results of the Round One Consultation – report to HEFCE by
Technopolis**

technopolis_[group]

Table of Contents

1. Summary	1
1.1 Introduction	1
1.2 Research information management systems	1
1.3 Main tasks involved in making the submission	1
1.4 People involved in the preparation of the submission	2
1.5 Scope of submissions	3
1.6 Staff submitted	3
1.7 Alignment with research information systems for RAE	3
1.8 Alignment with internal research information systems	3
1.9 Challenges faced in preparing submissions	4
1.10 Insights gained from checking additional records	4
1.11 Person days involved in the collection and verification of submissions	5
1.12 Implications for development of research information management	5
1.13 Challenges and opportunities of aligning with REF	6
1.14 People and functions involved in research information management	6
1.15 Internal capability to interrogate bibliometric data	7
1.16 Implications for faculty and other researchers	7
1.17 Other implications so far for researchers	7
1.18 Other preparatory work	7
1.19 Practicable suggestions for the design of the future REF	8
1.20 Alignment of REF requirements and internal research information needs	8
1.21 Other issues arising	9
<hr/>	
2. Introduction	9
2.1 Background to the REF bibliometrics pilot	9
2.2 The lessons learned consultation	11
<hr/>	
3. Findings arising from the consultation	14
3.1 Research information management systems	14
3.2 Main tasks involved in making the submission	14
3.3 People involved in the preparation of the submission	17
3.4 Scope of submissions	19
3.5 Staff submitted	20
3.6 Alignment with research information systems for RAE	21
3.7 Alignment with internal research information systems	22
3.8 Challenges faced in preparing submissions	23

3.9 Insights gained from checking additional records	26
3.10 Person days involved in the collection and verification of submissions	28
3.11 Implications for development of research information management	30
3.12 Challenges and opportunities of aligning with REF	32
3.13 People and functions involved in research information management	33
3.14 Internal capability to interrogate bibliometric data	35
3.15 Implications for faculty and other researchers	35
3.16 Other implications so far for researchers	37
3.17 Other preparatory work	37
3.18 Practicable suggestions for the design of the future REF	38
3.19 Alignment of REF requirements and internal research information needs	39
3.20 Other issues arising	40
<hr/>	
4. List of abbreviations	41

Table of Figures

Figure 1 Pilot institutions	10
Figure 2 Units of Assessment (UoAs) encompassed by REF bibliometrics pilot	10
Figure 3 Submission and validation process encompassed by round one consultation	13
Figure 4 Person days involved in the collection and verification of submissions	30

Lessons learned by institutions participating in the Research Excellence Framework (REF) bibliometrics pilot

Results of the Round One Consultation – report to HEFCE by Technopolis

1. Summary

1.1 Introduction

This paper presents the results of a consultation with the 22 universities and colleges involved in the Research Excellence Framework (REF) bibliometrics pilot, and concerns their experience of the data collection and validation phase of the pilot. The feedback exercise also invited the pilot institutions to offer any practicable advice they might have, in light of the pilot, as to how HEFCE might best develop the full REF bibliometrics exercise.

1.2 Research information management systems

In order to help us to understand potential clustering of insights and lessons learned across the 22 pilots, we asked each pilot institution to describe its research information management systems in the period immediately prior to the start of the REF bibliometrics pilot.

Our orientation interviews with pilots and non-pilots had alerted us to a broad spread of systems, in terms of their scope and sophistication, which the consultation confirmed.

There was a small minority of pilot institutions, which had centralised research management systems with comprehensive repositories/data warehouses and bibliographic databases covering most staff (not research students, not certain contractors, not all honorary or visiting posts) and a good level of process automation to support data entry and verification. These institutions also had relatively good levels of compatibility – in terms of data requirements – between other related information systems on grants, HR and so on.

In stark contrast, there was another small group that had no central data management and operated a distributed arrangement with a mixture of digital and paper-based systems to collect publications and log bibliographic data the scope, structure and content of which was often particular to a given department or faculty.

The majority sat somewhere in the middle of these two extremes, with some level of central management, possibly an institutional repository, and a long tail of decentralised systems to hold publications and minimum bibliographic data.

1.3 Main tasks involved in making the submission

HEFCE and its consultants, through the formulation of their information request, defined the main, high-level tasks involved in making a submission, which essentially comprised four steps:

- Identifying all relevant staff, and listing all of those named individuals in a standard format with detail additional information as described in Table 1 of the HEFCE specification

- Identifying publication data for selected staff, and listing all of those named outputs (publication titles) in a standard format as described in Table 2 of the HEFCE specification
- Linking authors to publications, for all identified staff and research papers as described in Table 3 of the HEFCE specification
- Checking lists of additional publications provided to the pilot institutions by the bibliometricians, and identified from a 'top-down' address-based search for each pilot, using the Web of Science database

The detail procedures differ somewhat from pilot to pilot, primarily because of the differences in the state of development of their respective information management systems and the (related) decision on the scope of the submission, in terms of the proportion of all staff and all publications.

The first group comprised just two universities, both of which were able to generate a new linked database specifically for the REF pilot, which involved the planning team expanding the scope of the return beyond the more selected set of staff and four publications compiled for the 2008 Research Assessment Exercise (RAE2008).

The second and largest group (12) involved universities in the extraction of data from the university's central publications database of staff and research outputs, which was often referred to as the RAE database.

A third group prepared the three HEFCE tables by way of a rather more open and variable process. The identification of staff was performed mainly using the HR system, then this list was used to structure a wide-ranging search for related publications in both internal and external databases, from Web of Science to Google, and lastly, people undertook a resource-intensive process of cleaning and consolidating the output data identified in order to construct a final version of Table 3.

In a minority of cases, institutions used the exercise to identify, acquire and deposit missing publications into their central databases.

1.4 People involved in the preparation of the submission

In around 10 cases, the preparation of the submission was overseen by the pro-vice chancellor research, although the degree to which he or she was closely involved did vary. In 12 cases, one or other senior managers from the university's central departments, most often the director of planning or director of research, directed the work. In one case, the heads of the library and planning managed the project jointly.

In every case, the work itself involved a small team centred on either the planning or research office. In a great majority of cases (18), preparing the submission was reported to have required the specialist input of other functions from IT to the library to HR. There does appear to be a pattern of sorts here, wherein those institutions with the more powerful and comprehensive pre-existing research information systems were more likely to be able to run this as a more narrowly based project within planning or research services, with only nominal external input required. Elsewhere, pilots elected to appoint task-and-finish project teams, often with members drawn from across central services.

In a minority of cases, pilot institutions elected to create a REF pilot working group to oversee the institution's engagement with the pilot, offering strategic direction on the one hand and a reassuring learning/dissemination function on the other. In most cases, these included senior university officers, heads of functional services and senior academics from faculties and schools.

On the subject of academic involvement, it appears that this was a supervisory input for the most part, small in extent, and involving for example heads of department or other senior academics. Faculties were often involved in checking and verification of the additional submissions presented by Evidence, however this tended to be managed

by departmental administrators with some input from senior academics. There are however instances where individual researchers were invited to check records.

1.5 Scope of submissions

The ambition of the pilot had been to gather information on staff and research outputs in each of the 35 units of assessment targeted, and which were relevant to the respective institution, however in recognition of resource pressures at the individual institutions, this ambition was expressed as a strong preference rather than mandatory requirement.

In the event, the great majority of pilot institutions (18) made a return for each of their respective subject areas that coincided with the 35 units of assessment that had been selected for inclusion within the pilot.

Just three of the 22 elected to make a partial submission. In each case, the decision to submit staff and papers for a *selection* of relevant UOAs, rather than all that applied, was a function of them trying to strike a balance between available administrative resources and the workload implied by the submission.

1.6 Staff submitted

The great majority of pilot institutions (19) sought to include details of all staff that might have produced publications, whether they fell into an RAE-eligible category or not. In the first instance, institutions sought to compile a more comprehensive return than had been managed for RAE2008, but still working with the HEFCE/Higher Education Statistics Agency (HESA) categories of research-active staff (A to D). The main additional category of staff was junior researchers involved in work that was not linked to grants held by more senior staff at the university, which were ineligible by RAE definitions.

1.7 Alignment with research information systems for RAE

We asked all pilot institutions to judge the extent to which the information requirements set out in the REF pilot specification aligned with the nature and extent of the research information gathered for RAE2008. The intention was to determine how far the REF bibliometrics pilot demanded similar or very different types of data.

10 pilots stated that the types of data asked for by the pilot coincided reasonably closely to the information they collect for making submissions to the RAE. Six pilots stated that the data requirements were a partial match with their RAE procedures and information systems. Another six stated that the requirements were a poor match, somewhere between hardly and not at all, with the current internal data collection procedures.

In answers to later questions, a significant minority stated that the REF bibliometrics requirements were essentially additional to the RAE data. While the broad classes of information were similar, the scope, level of granularity and timeframe were all markedly more onerous for the REF pilot as compared with RAE2008.

1.8 Alignment with internal research information systems

We also asked pilots to comment on the extent to which the information required in the REF bibliometrics pilot coincided with the kinds of information they would expect to gather and report on as a matter of course for internal purposes.

The answers to this question clustered in three groups, which are broadly a group for which the requirements were reasonably well aligned with internal needs, a second group where there was some fit and a third group where the requirements go far beyond that which they gather and report on as a matter of course.

1.9 Challenges faced in preparing submissions

The most widespread challenge reported was the timing of the data collection across the summer months, which meant pilots had to carry out the work at a point in time when most staff and academics would be taking breaks for holidays.

The experiences reported are almost bimodal, with the two pilots with the most advanced information management systems stating that the data collection process had been straightforward, while most of the rest stated that the process of compiling a complete record of staff and outputs for the full period under review had been challenging and frustrating in equal measure. The smaller specialist research institutes had found the process relatively straightforward too, although the time period had posed challenges. The challenge was most keenly felt by the largest pilots with the least developed information systems.

The period under review exacerbated an already demanding information management brief for the pilots, as their historical records and 'legacy' information systems tend to be weaker than current systems on almost every measure.

For the great majority, the biggest challenge revolved around the verification process, which was a novel, externally defined exercise that had to be conducted within four to six weeks in the run up to Christmas.

The short period of time available for the verification process does appear to have had a significant and widespread impact on behaviour with a majority of pilots indicating that they were simply not able to check most of the surprisingly large number of additional records provided to them by Evidence and Symplectic.¹

1.10 Insights gained from checking additional records

The majority of pilots appear to have gained a deeper understanding of the universality of the challenge of maintaining complete, accurate and consistent bibliographic records, whether one is a smaller university or a world-leading information services company such as Thomson-Reuters.

There are numerous comments on tactics and challenges for improved 'disambiguation', the act of resolving any uncertainty around exactly which author or publication is in the list, which range from tactics for dealing with foreign names to the sufficiency of the matching criteria used on this occasion.

The responses suggest that most pilot institutions will have concluded that automated checking can be helpful to maintaining the completeness and integrity of their internal records. However, it has also left the pilots with a view that a top-down, automated process of author-publication matching is likely to have a very high error rate. A significant minority remarked on the inefficiency of the search strategy used on this occasion, which it was felt did not exploit fully the capabilities of the matching tool, with for example, records being matched using initials rather than first names and not being cross-checked by discipline.

There was a general view expressed that this kind of matching exercise would always require a final check by people with good knowledge of the academics in question, arguably carried out at individual faculty level.

For two commentators, the experience had underlined the importance of getting the data right at the time of entry, in an observation that has echoes of the debate in manufacturing between advocates of total quality management – right first time – and more traditional approaches to quality control (inspection, rejection, rework).

¹ Evidence Limited and Symplectic Limited are the contractors appointed by HEFCE to jointly run the bibliometrics work in the REF bibliometrics pilot.

There was a presumption that the verification process would help pilots to gauge the quality of their information systems and procedures. In practice, only a minority of pilots commented directly about the performance of their research information systems, as a result of the verification process, which is to say whether their internal records are more, or less, complete and robust than they had believed to be the case.

The great majority commented on the laborious nature of the task, and perhaps unsurprisingly, the larger research universities felt this most acutely. There was no readily discernible pattern evident in these experiences relating to the differing levels of sophistication of research information systems.

1.11 Person days involved in the collection and verification of submissions

We asked each pilot institution to provide us with an estimate of how many person days had been involved in the whole data collection and verification process, including checking additional records.

The average was around 65 person days, with a median of around 50 person days. However, the individual estimates varied widely and ranged from 10 person days, in two cases, to more than 200 person days in one instance.

The data-collection process was reported to have taken around 50 days on average or 40 days typically, but with an equally large range from around eight person days to more than 130. The verification process required an additional 15 person days on average, with an even greater spread, ranging from two person days to 70 person days. The verification process amounted to around 20% of the total effort, on average. However, there is a broad spectrum of effort, ranging from 5% to 66%.

The differences appear to reflect the combined effects of two loosely related factors, which are the sophistication of the pre-existing research information system and the degree to which the pilots sought to create the most complete and robust list of authors and publications.

1.12 Implications for development of research information management

The majority of pilots indicated that they are planning to make several changes to their information systems and procedures, as a result of their participation in the REF bibliometrics pilot. There is a wide range of changes planned or foreseen, with the great majority involving incremental improvements to current research information management arrangements rather than major investment projects. The one exception is institutional repositories where four pilots plan to implement major new systems and two others have concluded this is going to be necessary at some point in the near future.

The incremental changes include various means by which to improve data quality (complete, consistent, accurate, timely data input, etc), extend and deepen existing databases and integrate key databases and systems. More specifically, the anticipated changes include but were not limited to:

- Implement institutional repositories or other central databases
- Introduce new procedures to update/improve quality of departmental databases
- Improve integration and inter-operability across information systems
- Extend and deepen existing staff and publications databases
- Subscribe to Symplectic or similar
- Introduce regular reporting of publication data
- Awareness raising

- Increase amount of information we capture about joiners and leavers
- Develop some kind of modelling capability
- Introduce a bonus system to encourage academics to keep records up to date
- Implement a requirement to deposit all publications in the repository
- Identify and plug key gaps in existing information

Not all pilots plan to make changes. Nine of the 22 pilot institutions stated that they expected to make few or no changes to their respective research information management systems as a result of their participation in the pilot so far, and that they were waiting to hear the detailed design of the final bibliometrics arrangements.

A minority stated that the exercise had confirmed that their present system was deficient in various ways, and that they clearly had to do something to improve the situation significantly in the medium term. In two of these cases, people spoke of a dilemma in that the improvements needed were far reaching, in terms of systems and procedures, and would cost a great deal, a proposition that would be tough to sell while the key design parameters for the REF bibliometrics remain open.

1.13 Challenges and opportunities of aligning with REF

On challenges, there was a reasonably consistent view expressed in various ways by the great majority of respondents, wherein pilots foresee a need to develop their research information systems and procedures in order to meet the anticipated REF *requirement* to present data on all staff and all publications. This will be a demanding programme of technological and organisational development, which will have substantial implications for university resources and culture/internal relationships. Some view this systems requirement as essentially a one-off cost, albeit a significant one, to make the transition to the new arrangements.

Technical challenges are not seen as being unduly problematic, as compared with the organisational and financial, although the issues around system integration and interoperability do recur.

As far as opportunities are concerned, there were just two cited and these were closely related: the first is the opportunity to add a report on publication outputs (numbers in the first instance) to complement existing management reports on research inputs, which presently focus on research bids (pipeline) and research income; the second is the anticipation of an improved internal capacity (research or planning offices) to model future income and inform strategy deliberations with more and better analyses of past performance.

1.14 People and functions involved in research information management

16 of the 22 pilot institutions replied saying that they expect to see an increase in the number and range of people involved in the research information management in the future. There was a consistent picture as to which groups would be doing more in the future, although the increase is quite likely to be small in proportionate terms:

- In all cases, the library is expected to play a fuller role in future
- In the great majority of cases, academics themselves are expected to have to do a little more to ensure all publications are deposited and metadata is captured
- IT and HR are both widely expected to become more involved in periods of system development and wider integration
- In several cases, people expect the research and planning offices to be doing rather more, producing periodical reports on performance and ad hoc analyses as an input to senior management's oversight and regular strategic planning

1.15 Internal capability to interrogate bibliometric data

We also asked pilots whether they envisage creating an internal capability to interrogate bibliometric data (i.e. run citation analyses).

15 of 22 stated that they would expect to need to be able to interrogate bibliometric data in future. As regards the route by which such analytical services would be secured, the answers were more wide ranging, with several stating that they would expect to manage this internally while others stated that they had or would enter into commercial contracts with specialist service providers and toolmakers. Most however are undecided, and awaiting more information on the REF bibliometric requirements and any national efforts HEFCE might make to facilitate community-wide access to services and data.

1.16 Implications for faculty and other researchers

The experience of the pilot suggests that everyone will have a role to play in future research assessment, including individual academics and this means paying closer attention to publications records. There is a reasonably clear set of expectations emerging, wherein academics will need to be more aware and accepting of the:

- Importance of individuals' publications data to future institutional research income
- Primacy of central systems within an institution's research information management arrangements
- Importance of keeping one's own publications records up to date
- Importance of data accuracy and consistency
- Importance of becoming more familiar with bibliometrics and citation analyses, and any implications therein

Several pilot institutions make the point that the academic input must be kept simple and light, supported by a fair amount of process automation/software tools, to ensure higher levels of intrinsic data integrity and to avoid disincentives.

1.17 Other implications so far for researchers

We asked pilot institutions whether they foresee any other implications so far for researchers. The majority of respondents foresee no obvious additional implications for researchers, although a significant minority did flag a concern about the wider impact of using bibliometrics on publishing behaviour.

1.18 Other preparatory work

To bring feedback on experiences to a close, we asked pilot institutions, whether, on the basis of their experience to date, they were considering undertaking any other work on their research information management systems to prepare for the REF. The great majority said that no other REF-specific preparatory work was in hand.

A minority did cite some examples of other preparatory work:

- Further development of training materials and training courses
- The upgrading of related information systems, and in particular the HR systems
- The creation of a cross-departmental REF strategy group to track developments and deliberate implications and responses
- The creation of an integrated management information unit, which will pick up REF requirements as one of its responsibilities
- Benchmarking other indicators, on for example the volume and sources of research income

Others simply noted that there was ongoing development on the research information systems, which would be pursued independently of REF. Another pilot reported that it had begun a round of individual consultations between the Vice Principal for Research and all researchers, to discuss publication history and strategies.

1.19 Practicable suggestions for the design of the future REF

We asked pilot institutions what practicable suggestions they had for the design of the future REF in terms of making the data requirements and data collection processes more manageable for their institution.

There was a reasonably consistent set of related recommendations offered, which hinged on the clarity of the final specification, the timeliness of the final detail design and the length of time given over to the preparation of submissions:

- Fix the design parameters as soon as is reasonably practicable
- The final design must be accompanied by a specification for the data requirements and planned analyses that is crystal clear and comprehensive
- Give longer lead times for the preparation of the first full submission, as the timescale for the pilot proved problematic for many
- Give more consideration to the timing of the period in which submissions are to be made

Other recommendations, which were not necessarily additive or complementary, included a request for HEFCE to bear in mind that the bibliometrics is only one part of the proposed, new-look REF and that institutions might very well need to implement several development programmes in parallel. This has implications for what is practicable, for reasons of resources (people and cash) and priorities.

Two institutions recommended the development of common, web-based tools to facilitate consistent and robust uploads of institutions' data to HEFCE, or its contractors running the bibliometrics element of the REF. One institution recommended HEFCE revisit the issue of author identification.

There was also concern expressed regarding the plan to press ahead with a developmental REF bibliometrics exercise in 2010, which it was believed would risk rushed decision-making and implementation.

1.20 Alignment of REF requirements and internal research information needs

The great majority stated that to a large extent their internal requirements evolve to reflect external needs, whether that is HESA, HEFCE or the research councils. In light of this, few felt they were in a position to specify what more HEFCE might do to improve alignment, and simply restated their previous recommendations regarding the timing of the communication of the final design and absolute clarity and transparency around that detail specification. Some suggestions were made however, and these are presented below:

- Five institutions suggested that there would be value to the community overall if HEFCE were to give further consideration to the compatibility between the various reporting cycles and data requirements made by other research funding bodies and statistical agencies (e.g. HESA)
- One institution suggested that there might be value for both parties, HEIs and funding bodies, if there were to be a debate around the more strategic question of what they are collectively thinking and wanting to do with regard to the development of their internal research management systems

- Another institution suggested it would be good if HEFCE were able to find a way to keep the data requirements under review, with some kind of ongoing consultation, in order to optimise the metrics and minimise wasted effort
- Another institution suggested that HEFCE should explore the possibility of making a community-wide deal with one or more of the information services businesses that maintain the key databases, Web Of Science and SCOPUS

1.21 Other issues arising

We asked pilot institutions whether there were any other issues arising from the pilot so far. Half of the pilots said they had covered the lessons learned and matters arising in their replies to earlier questions and made no further input.

Six institutions took the opportunity to cite their more fundamental concerns about bibliometrics and in particular its value for money.

Several institutions took the opportunity to restate the one or two key points they had made earlier, and in particular the critical need for HEFCE to reach a final decision as soon as possible and to be crystal clear about its requirements.

Two institutions asked HEFCE to remember to consider and make a decision on the treatment of a range of other issues, ranging from the treatment of non-standard publication types to issues of equality and diversity.

2. Introduction

This paper presents the results of a consultation with the 22 universities and colleges involved in the Research Excellence Framework (REF) bibliometrics pilot, and concerns their experience of the data collection and validation phase of the pilot.

2.1 Background to the REF bibliometrics pilot

Following consultation on a proposed Research Excellence Framework (REF), the Higher Education Funding Council for England (HEFCE) is implementing a bibliometrics pilot project and other development work to test and develop relevant databases and appropriate analyses.

The pilot aims to develop and test a number of issues:

- Which disciplines should bibliometrics be applied to?
- Questions of scope – universal or selective coverage?
- Are papers credited to the researcher or the institution?
- How to collect data, and the implications for institutions
- Web of Science or SCOPUS?
- Refining the methods of analysis (defining normalisation fields, handling multidisciplinary journals, self citation, etc)
- The choice of indicators and thresholds for excellence
- Interpretation and use by expert panels

HEFCE selected a cross-section of 22 institutions from a long list of more than 50 volunteers, with the intention of constructing a sample that was broadly representative of the total HEI population with respect to size, specialism and geography. Figure 1 presents the list of pilots, in alphabetical order.

Figure 1 Pilot institutions

Bangor University	London Sch of Hygiene and Trop Medicine
Bath, University of	Nottingham, University of
Birmingham, University of	Plymouth, University of
Bournemouth University	Portsmouth, University of
Cambridge, University of	Queens University Belfast
Durham, University of	Robert Gordon University
East Anglia, University of	Royal Veterinary College
Glasgow, University of	Southampton, University of
Imperial College London	Stirling, University of
Institute of Cancer Research	Sussex, University of
Leeds, University of	University College London

Figure 2 lists the 35 units of assessment (UoAs) selected, on the basis of their coverage within the citation databases, from the 64 involved in the 2008 Research Assessment Exercise (RAE2008), which span the natural, physical and social sciences, including several subjects where the suitability of bibliometrics remains undecided. No arts and humanities subjects have been included in the pilot.

Figure 2 Units of Assessment (UoAs) encompassed by REF bibliometrics pilot

UoA	UoA name	UoA	UoA name
1	Cardiovascular Medicine	19	Physics
2	Cancer Studies	20	Pure Mathematics
3	Infection and Immunology	21	Applied Mathematics
4	Other Hospital Based Clinical Subjects	22	Statistics and Operational Research
5	Other Laboratory Based Clinical Subjects	23	Computer Science and Informatics
6	Epidemiology and Public Health	24	Electrical and Electronic Engineering
7	Health Services Research	25	General Engineering and Mineral & Mining Engineering
8	Primary Care and Other Community Based Clinical Subjects	26	Chemical Engineering
9	Psychiatry, Neuroscience and Clinical Psychology	27	Civil Engineering
10	Dentistry	28	Mechanical, Aeronautical and Manufacturing Engineering

11	Nursing and Midwifery	29	Metallurgy and Materials
12	Allied Health Professions and Studies	32	Geography and Environmental Studies
13	Pharmacy	34	Economics and Econometrics
14	Biological Sciences	40	Social Work and Social Policy & Administration
15	Pre-clinical and Human Biological Sciences	43	Development Studies
16	Agriculture, Veterinary and Food Sciences	44	Psychology
17	Earth Systems and Environmental Sciences	46	Sports-Related Studies
18	Chemistry		

The ideal data requirement for the pilot study was understood to be onerous and challenging. This is because the scope of the general data required for policy context and the analysis of methodological options goes beyond the scope of the specific data thought likely by HEFCE to form the target of the future REF (i.e. a set of outputs associated with research-active staff). Institutions will also bear in mind the responses to HEFCE's prior consultation, and the extent to which these responses from the research community have posed questions about the bibliometric function and methodology which HEFCE now seeks to answer. The key point is that the specification for the pilot was deliberately comprehensive including a wide range of staff and outputs so that we could test different models using subsets of the data. It was made clear from the outset that the 'real REF' requirements would be less comprehensive.

Compliance, insofar as is feasible, remains essential if all the policy issues raised by the introduction of a metrics-based assessment system are to be fully addressed. The guiding principle is therefore the best possible data-set should be gathered where available. If there is a great disparity between the ideal and what is actually provided by institutions, then institutions and individual researchers might question the outcomes of the derived analyses. Furthermore, if Evidence and Symplectic were to provide and disambiguate the bulk of the data required for the required bibliometric analyses without substantial involvement by higher education institutions (HEIs) then this would not address any part of a long-term and systematic implementation plan nor assess the true challenge to institutions of complying with a national system.

Although the principle therefore remains unchanged, the practice must be varied to accommodate what institutions can reasonably accomplish in the time available. The development and supply of data from institutions must also be completed sufficiently early for the contractors to be able to carry out their tasks to produce the final project database by the end of October. The project partners have therefore agreed that there should be a hierarchy of data requirements, from the comprehensive ideal to a mandatory minimum. This revised specification is set out below.

2.2 The lessons learned consultation

The collection of feedback on lessons learned is being carried out in two rounds, with this the first-round report detailing experiences and insights associated with the data collection, submission and validation phase. A second round consultation is planned for May 2009, several weeks after the 22 pilots will have received the results of the bibliometrics pilot.

Figure 3 details the process the pilot institutions have gone through in preparing and validating their submissions, in order for HEFCE to arrive at a comprehensive

validated database on which to conduct analyses. It is these elements of data collection, submission and validation that form the subject of the round one consultation.

Figure 3 Submission and validation process encompassed by round one consultation

Pilots	Project partners	Timetable
<p>Institutions submit initial databases in an agreed format via the HEFCE extranet.</p> <p>1. Table 1 – Data on research staff – these data should be as complete as possible at this stage</p> <p>2. Table 2 – Currently available data on publications (including RAE2008 submitted outputs, and any other data available at this stage)</p>	<p>Evidence checks staff records (Table 1) for staff who joined after Jan 2001, identifies additional output records in Thomson (authored by those staff before they joined the HEI)</p> <p>Evidence matches initial publications to Thomson data</p>	August 2008
<p>Institutions submit extended databases:</p> <p>1. Additional output records (Table 2)</p> <p>2. Links between outputs and staff (Table 3)</p>	<p>Evidence reviews data and supplements Table 2 with additional – indicative – Thomson records (based on HEI address matching) where necessary</p>	September 2009
<p>HEI administrators (or individual staff) check the additional indicative outputs in Table 3 by accessing Symplectic/Evidence central system</p>	<p>Evidence/Symplectic continue to develop links, and iterate with institutions for validation</p>	December 2009
<p>Validated and author disambiguated database of Thomson records covering all the pilot institutions</p>		March 2009

All 22 pilot institutions provided written feedback on a series of questions relating to their research information management systems, the scope of their submission to the pilot and any notable points arising regarding the process by which they had prepared and made that submission.

3. Findings arising from the consultation

3.1 Research information management systems

In order to help us to understand potential clustering of insights and lessons learned across the 22 pilots, we asked each pilot institution to describe its research information management systems in the period immediately prior to the start of the REF bibliometrics pilot. Our orientation interviews with pilots and non-pilots had alerted us to the broad spread of systems, in terms of their scope and sophistication, and that this might very well make a big difference to both the experiences of individual pilots and the relevance of those lessons to segments within the much larger non-pilot population.

In short, the broad spread was confirmed. There was a small minority of pilot institutions, which had centralised research management systems with comprehensive repositories/data warehouses and bibliographic databases covering most staff (not research students, not certain contractors, not all honorary or visiting posts) and a good level of process automation to support data entry and verification. These institutions also had relatively good levels of compatibility – in terms of data requirements – between other related information systems on grants, HR and so on.

In stark contrast, there was another small group that had no central data management and operated a distributed arrangement with a mixture of digital and paper-based systems to collect publications and log bibliographic data the scope, structure and content of which was often particular to a given department or faculty.

The majority sat somewhere in the middle of these two extremes, with some level of central management, possibly an institutional repository, and a long tail of decentralised systems to hold publications and minimum bibliographic data. However, most institutions have been developing these systems in the recent past, in large part to meet the very specific requirements of RAE2008: as such, most systems have significant gaps in coverage, which become much more pronounced for earlier periods, and significant problems with duplications and other quality-related issues.

The spread of system capabilities did not relate in any obvious way to institutional size or specialisation or research strength.

3.2 Main tasks involved in making the submission

HEFCE and its consultants, through the formulation of their information request, defined the main, high-level tasks involved in making a submission, which essentially comprised four steps:

- Identifying all relevant staff, and listing all of those named individuals in a standard format with detailed additional information as described in Table 1 of the HEFCE specification
- Identifying publication data for selected staff, and listing all of those named outputs (publication titles) in a standard format as described in Table 2 of the HEFCE specification
- Linking authors to publications, for all identified staff and research papers as described in Table 3 of the HEFCE specification
- Verification of lists of additional publications identified by the bibliometricians using an address-based search in the Web of Science database

The detail procedures differ somewhat from pilot to pilot, primarily because of the differences in the state of development of their respective information management

systems and the (related) decision on the scope of the submission, in terms of the proportion of all staff and all publications.

The first group comprised just two universities, both of which were able to generate a new linked database specifically for the REF pilot, which involved the planning team expanding the scope of the return beyond the more selected set of staff and four publications compiled for RAE2008. It also involved expanding the fields of information presented on both staff and outputs. This was an automated process by and large, with new standard reports having been defined such that the information could be re-produced quickly and efficiently in future, should that prove necessary.

The system was more or less capable of providing the data required unchanged. We additionally selected those staff who had joined the University since the end of the RAE 2008 and staff who had been eligible for RAE2008, but not submitted. The RAE/REF System was modified to export all category D outputs, rather than just the best four.

The data for the REF pilot was sourced exclusively from the RAE2008 repository of RA1 (staff) and RA2 (research outputs) information. This meant that the mapping of data and data derivations applied for the RAE were retained. A new database was created to hold the REF return, which generated the relevant Tables 1, 2 and 3. A significant number of data updates were required to the data, in the main due to differences between RAE08 and REF Pilot criteria, and the significant extension to the pool of data required for the REF Pilot, relative to the RAE. The validation and data updates required to non-RAE submitted research outputs were the most significant tasks in compiling the REF pilot return. This involved completing data fields (for example via validation against the CrossRef reference system) to maximise uniformity within fields and to minimise the number of duplicate (i.e. instances of multi-authored) outputs in Table 2.

The second and largest group (12) involved universities in the extraction of data from the university's central publications database of staff and research outputs, which was often referred to as the RAE database. In some cases this was used as the starting point for producing a submission that expanded upon the selected RAE publications. Several pilots noted that these central databases required a significant amount of work to capture a greater proportion of all REF pilot fields, of which there were several that had not been required for the RAE. In two cases, before beginning to compile a return, institutions asked staff to update the university's central publications database with relevant research outputs produced in the period since 2001, prioritising articles published in ISI Web of Knowledge² journals in the first instance.

Staff information and associated publication data was extracted from the University's RAE database. The staff data set was checked to ensure the accuracy of eligibility and RAE category fields. Prior institution information was added to our data set where it existed in our Management Information Systems.

Given that we have a central database, providing publications data was relatively straightforward. The list of current and former staff was also straightforward, and was extracted from our HR and payroll system. The publications database includes the staff ID number used in the HR system, which enable[s] the two to be linked.

Some work was needed to extract the data in the required format, but the main issues for the School were: (i) we do not have a central record of

² The Institute for Scientific Information (ISI) was founded by Eugene Garfield in 1960. It was acquired by Thomson Scientific & Healthcare in 1992, and became known as Thomson ISI and now as Thomson Scientific. The ISI provides a wide range of bibliographic database services. Its speciality is citation indexing and analysis, a field pioneered by Garfield. It maintains citation databases covering thousands of academic journals, including a continuation of its long-time print-based indexing service the Science Citation Index (SCI), as well as the Social Sciences Citation Index (SSCI), and the Arts and Humanities Citation Index (AHCI). All of these are available via ISI's Web of Knowledge database service.

Category C and D staff; (ii) we do not record the previous/destination institution of starters and leavers on the HR system. The only way of finding this information is from CVs in paper files, and we did not have the time or resources to do this for the pilot; (iii) providing the initial datasets was relatively straightforward because HEFCE/Evidence were so inclusive (they asked for all publications for all staff). This last point was a problem for those HEIs without publication databases, but made the task easier for the us.

We found this to be a relatively simple process as we imported data from our Access database directly into the template provided. There were difficulties regarding ensuring that staff data was accurate, particularly in regard to staff who did not have straightforward contracts or who had multiple contracts over the time period some of which were not RAE eligible. Neither our Access database nor our HR system accurately records CAT C staff or prior or destination HEI. We did not provide a 'Table 3' as our data already linked each publication to a member of staff.

A third group prepared the three HEFCE tables by way of a rather more open and variable process. The identification of staff was performed mainly using the HR system, then this list was used to structure a wide-ranging search for related publications in both internal and external databases, from Web of Science (WOS) to Google, and lastly, people undertook a resource-intensive process of cleaning and consolidating the output data identified in order to construct a final version of Table 3.

The main tasks involved in providing data to the REF pilot were: identifying relevant staff (using RAE eligibility), sourcing publication data (the university's internal author-deposited bibliographic database, WOS, Scopus, Google, staff CVs, etc), and obtaining personnel data from HR Department. The most time-consuming task was then crosschecking and linking the data in REF tables and verifying publication information via online sources. The main task involved in checking the additional publication records supplied by Evidence was checking with the online source or using WOS, Scopus, Google, etc.

Table 1 was compiled with HR data, this was not complete to the level required for the pilot, and many fields such as 'early-career researcher' and prior and destination institutions were blank. This information was not available therefore we could not complete these fields, even manually. The departmental Reference Manager files were collected centrally and compiled to complete Table 2. However, departments had been collecting different levels of data and not all matched the requirements of the REF pilot. This involved some manual input in order to ensure all the necessary information was entered into the return. This was time-consuming, but achievable in the timeframe (I'm sure only due to the fact that we are a relatively small institution). We have a Web of Science subscription, but not a SCOPUS subscription, therefore in order to complete the Table 2 column 'Indexed by Thomson-Reuters' we had to search Web of Science for all of our publications and merge these with what was provided by departments. This also helped in completing some of the fields left blank by departments as the information came from Web of Science. Table 3 involved a lot of manual manipulation of the information we had for the previous tables, again it was only possible in the timeframe, with limited staff input due to our size.

Downloading records from Web of Science, merging with RAE data and cross-checking (manually) with staff personal webpages. Manual staff allocation to UoAs.

The first task was to assemble publications information for the Departments into a single format. Subsequently, each item of information requested within Tables 1, 2 and 3 involved us in major effort, not least because of our lack of an integrated electronic data management system. For example, the new HR system which we needed to enter start/end dates only holds information post 2007 so we had to re-open the old system, which uses a different staff numbering system, and manually switch between the two to get the information required. Table 3 (link author – publication) was

particularly difficult due to no obvious way of collecting the data electronically.

In a minority of cases, institutions used the exercise to identify, acquire and deposit missing publications into their central databases.

3.2.1 Checking additional staff-output records supplied by Evidence/Symplectic

Checking the data was laborious and time-consuming. Evidence had problems matching staff to publications (despite both tables having staff ID numbers), and this meant that we initially had around 10,000 pending outputs. Evidence were very helpful in trying to correct this, and we eventually ended up with only 1,000 or so to check. But by this time we had very little time to complete the task. Checking was made more difficult by the fact that Evidence/Symplectic had removed our publication ID numbers, which could have been used to cross-check pending items against the database.

The process for checking records provided back by Evidence was time consuming and extensive. In the time available, the College took an automated approach based on co-author expertise to approve/reject the huge database of proposed records. Significant numbers of these records were already included in the College's data submission – the identification of surplus and duplicated records was very time consuming and significantly reduced the time available to validate genuinely 'new' outputs.

Additional records supplied by Evidence were disseminated to REF contacts within departments for verification.

Checking additional records was reasonably straightforward, however it involved a double effort for approximately 10% of the outputs (since we had provided a complete Table 3!), in those cases where the same publications were presented for verifying again (especially in cases of common surnames).

3.3 People involved in the preparation of the submission

All pilot institutions were asked who was involved in the preparation of their submission to the REF bibliometric pilot, in order to obtain a better understanding of both the scale of the endeavour and perhaps the breadth of functions involved in the submission.

In around 10 cases, the preparation of the submission was overseen by the pro-vice chancellor research, although the degree to which he or she was closely involved did vary. In 12 cases, one or other senior managers from the university's central departments, most often the director of planning or director of research, directed, managed and supervised the work. In one case, the heads of the library and planning managed the project jointly.

A group chaired by the Pro-Vice-Chancellor for Research and involving representatives from Information Systems Services (ISS), the Library and Research Support was convened to oversee our participation in the pilot and to take high level decisions, such as the range of data to be provided and the wording of the communication to staff about the pilot. In practice the majority of the work in collecting, providing and checking the submissions fell on ISS.

Our approach to the REF pilot was planned by the Pro-Vice-Chancellor (Research) and Registrar in liaison with the RAE Support Officer and IT Consultant who had been fully involved in RAE2008 preparations. The collection, provision and checking of the data was carried out by the RAE Support Officer and IT Consultant. REF contacts within academic departments were responsible for verifying additional publications provided by Evidence Limited.

The University's Planning Officer, in consultation with the Pro-Vice-Chancellor Research, co-ordinated a team comprising staff from the Library

(RAE/REF administrator), IT services, student helpers and members of professional services at Departmental level. Academic members of staff checked and updated their research portfolios

In every case, the work itself involved a small team centred on either the planning or research office. In a great majority of cases (18), preparing the submission was reported to have required the specialist input of other functions from IT to the library to HR. There does appear to be a pattern of sorts here, wherein those institutions with the more powerful and comprehensive pre-existing research information systems were more likely to be able to run this as a more narrowly based project within planning or research services, with only nominal external input required. Elsewhere, pilots elected to appoint task-and-finish project teams, often with members drawn from across central services.

A University REF Data Collection Team was created comprising members of Research & Enterprise, the University Library, HR, the Data Protection/FoI Office and Faculty contacts. This group was responsible for managing the central administration of the process. Table 1 data was collated using the University's central Research and HR Systems and then sent to key Departmental contacts for checking and amendment/confirmation. For Table 2 data, where available, local publications data was collected from departmental contacts, checked, converted (e.g. from Word, Web, Excel format) and loaded into an Access database. This data was supplemented by the Library via searches of bibliographic databases. Table 2 data also needed Staff Identifiers for authors to be added. For some units, the Library had to put databases together from scratch. In the Medical Faculty, individual members of staff were asked to check their publications lists personally.

Although the Pro-Vice-Chancellor (Research) was keen that the University volunteer for the Pilot Exercise, and had an overview of the University's preparations, planning and preparation of the submission was undertaken by those responsible for administering and co-ordinating the University's 2008 RAE submission. Officers in HR were responsible for collecting the staffing data but their involvement was limited to report preparation. In hindsight, this was far too small a team for the amount of work involved within a very limited timescale, but there was little recognition in the University of the significance of the REF Pilot Exercise (mainly because of a focus on RAE outcomes).

The main planning and collection of data was done by Deputy Head of Research alongside specialised additional support from within one of the University's larger Schools (this School already had a large volume of data available due to its existing publication database and could provide specific support to the large number of staff and outputs that were submitted from this School). Support was also available from staff within the HR Department (provision and checking of personnel data) and within the library (provision and checking of output data). The pilot information was submitted centrally and was largely verified centrally but some local checking was done within the larger School as mentioned.

In a minority of cases, pilot institutions elected to create a REF pilot working group to oversee the institution's engagement with the pilot, offering strategic direction on the one hand and a reassuring learning/dissemination function on the other (feedback to academics on pilot design parameters and likely implications for funding allocation and work processes). In most cases, these included senior university officers, heads of functional services and senior academics from faculties and schools.

A REF Pilot Working Group has been set up with a member of each of the College's Faculties included in the Pilot.

We have set up a REF pilot Project Group, chaired by the Director of Research Innovation Services, to manage the REF pilot return centrally, which has the remit of planning, setting up strategic and operational direction and decision making for the delivery of the REF pilot.

Director of Research, Pro Vice-Chancellor (Research), Associate Deans of Research, Research Manager and administrator heavily involved in administrating the RAE.

On the subject of academic involvement, it appears that this was a supervisory input for the most part, small in extent, and involving for example heads of department or other senior academics. Faculties were often involved in checking and verification of the additional submissions presented by Evidence, however this tended to be managed by departmental administrators with some input from senior academics. There are however instances where individual researchers were invited to check the additional records.

In the Medical Faculty, individual members of staff were asked to check their publications lists personally.

Our research office (CREDO) and our three Faculty research co-ordinators, with some Library support towards the end of the project.

Academic members of staff checked and updated their research portfolios

3.4 Scope of submissions

The ambition of the pilot had been to gather information on staff and research outputs in each of the 35 units of assessment targeted, and which were relevant to the respective institution, however in recognition of resource pressures at the individual institutions, this ambition was expressed as a strong preference rather than mandatory requirement.

In the event, the great majority of pilot institutions (18) made a return for each of their respective subject areas that coincided with the 35 units of assessment that had been selected for inclusion within the pilot.

Of course, the scope of the submissions differed across the pilots, in line with the size and specialisation of the institutions concerned. However, there was clearly a recognition of the need for a submission of sufficient breadth/diversity to enable the institution to identify and understand any issues arising from internal variance in research information systems across schools or subjects.

We made a return for all relevant UoAs. We had bibliographic information on outputs over and above those submitted in the 2008 RAE for all units, but the extent of coverage varied across the disciplines. We were unable to submit a record of all known research publications in any UoA, and we could not guarantee the quality of the bibliographic data for any non-RAE outputs. We felt it was important to participate consistently across all disciplines in order to gain maximum insight to the variations in citation rates and trends.

The subject coverage of the REF pilot mapped broadly to our disciplines of Science, Technology and Medicine. Therefore, we took an inclusive approach to the REF Pilot, including all staff and research outputs held within our systems. This approach was taken to enable a true test of bibliometric data coverage provided by the REF Pilot outcomes. The limitations of bibliometric data are journal specific and so it was decided that, given staff publish across the boundaries of journal subject categories, we did not wish to limit the pool of staff, and hence outputs, based on organisational unit or discipline.

Just three of the 22 elected to make a partial submission. In each case, the decision to submit staff and papers for a *selection* of relevant UOAs, rather than all that applied, was a function of them trying to strike a balance between available administrative resources and the workload implied by the submission, which was itself a function of university's breadth of interest and state of its research information systems.

Given the resources available, it was impossible for us to submit to all the UoAs concerned. In view of the inadequacies in local publications storage, we volunteered for only those UoAs where we were reasonably positive that the Departments concerned had some store of electronic data that we could work with. These comprised some 14% of the University's RAE submission.

In addition, the HEFCE asked that we collect data in two further UoAs in order to ensure that they had sufficient data.

Information was returned to five UoAs for the pilot, which was as many relevant UoAs as possible within the required timescale. Information on publications was readily available or was more easily accessed for these subjects and the outputs were either already verified for RAE purposes or could easily be verified retrospectively. Other UoAs not included were either not accessible within the timescale or included outputs that did not fulfil the criteria specified for the pilot. Effort was therefore concentrated on those UoAs that the University could provide information for as fully as possible.

We only put in a submission for UoAs covered (mainly) by our Faculty of Health. This was because this Faculty did have a single publications database and was the furthest along in terms of developing links with the institutional repository. The state of systems and other resourcing priorities precluded going forward with any other UoAs/Faculties. As far as the Pilot goes, it is therefore fair to say this was only a test of our most well prepared Schools and systems; it is therefore not necessarily representative of the university as a whole; other UoAs will be worse to a lesser or greater degree.

3.5 Staff submitted

The great majority of pilot institutions (19) sought to include details of all staff that might have produced publications, whether they fell into an RAE-eligible category or not. In the first instance, institutions sought to compile a more comprehensive return than had been managed for RAE2008, but still working with the HEFCE/HESA categories of research-active staff (A to D). The main additional category of staff was junior researchers involved in work that was not linked to grants held by more senior staff at the university, which were ineligible by RAE definitions.

The institution took an inclusive approach to the selection of staff for submission to the REF Pilot. The submission included all academic and honorary staff, beyond those eligible for submission into the RAE, i.e. including junior researchers who would not have been eligible for submission into the RAE. The 'Eligible for RAE' flag in Table 1 was completed accordingly. Whether or not the institution held a record of publications against staff was not a criterion for the inclusion of staff in the REF Pilot. Therefore staff were submitted in the REF Pilot without a corresponding entry in Table 3. The institution has a fairly comprehensive database of outputs for academic staff, but only limited dataset for junior researchers and honorary staff. The REF Pilot was seen as an opportunity to quantify the gaps in the database, and so the full list of staff was included.

We made a return of 83% of staff to RAE 2008. As far as possible we submitted information on those as well as the other 17% of staff. However it was impossible to get additional information on staff that had left the University during the period. Some of this information surfaced with the additional outputs received from Evidence.

On the RAE criteria specifically, there are clearly some issues of interpretation as to what amounts to an eligible, research-active individual, and the extent to which a principal investigator might be used as a proxy for an 'independent' researcher.

We included all staff potentially eligible to have been included in RAE depending on whether they satisfied the criteria for Category A. This included a significant number of postdoctoral researchers who, whilst holding contracts listing research as their main activity, may or may not have been eligible on the basis of being principal investigators on grants or holding major fellowships. It was impossible to distinguish using information available in the HR database. There was no time to do further digging given the numbers involved and the length of time provided for the task. We again felt there was no sense in being selective especially given Evidence's insistence that they wanted as wide a pool of outputs to cut from as possible.

All institutions expressed a desire to be fully inclusive, however a small minority stated that they had elected to submit only RAE-eligible staff as they believed there was a much higher risk that information would be incomplete or in some way compromised and that this risked distorting the results. Several others noted the practical difficulties of compiling this additional material, and concluded that it would not have been possible within the timescale defined by the REF pilot. However, it is clear that the requirement had been understood and that the notion of being more inclusive was generally well received.

We returned information for eligible staff. We thought that including information from other colleagues was likely to be unhelpful in that it was unlikely to be complete and might distort results.

Our aim was to gather publications for all RAE-eligible staff including all Cat C whether or not they had been returned. Ultimately, we opted not to include all staff and students associated with research outputs, research assistants, etc due to the restrictions of the timetable and the availability of data.

It was not possible within the timescale to include all staff as opposed to RAE eligible staff. Prior to the pilot, output information for non-RAE eligible staff was not held in entirety and, due to the time of year and the impending deadline, it was not feasible to collect this information. As such, only RAE eligible staff for the relevant UoAs were included for the pilot.

We submitted over 90% of our HEFCE-funded staff in the University's RAE submission, so adding in permanent staff not returned was fairly easy. What was far more problematic, and time consuming, was tracking down information of staff who had left and had not been included as Categories B or D. In the case of contract staff, we collected information only for those who would have qualified under the University's equal opportunities eligibility criteria, which had been applied in the RAE2008. Students (in case of publishing members of staff) were not submitted because of time and resource constraints.

The great majority of pilots noted the challenge of compiling comprehensive and complete information on non-eligible staff, whether that relates to contractors or post-doctoral researchers or research students. Most also commented on the challenge of compiling reliable information on joiners and leavers.

Just one of the 22 pilots indicated that they had sought to retain the 'independent researcher (i.e. principal investigator)' concept as a selection criterion, however as a specialist research centre this amounted to a listing that included almost all staff.

We have publications data for all academic staff, so were able to include everyone. The issue of how HEFCE defines RAE eligible staff has obviously led to a lot of debate between institutions, and those in the pilot have inevitably adopted different interpretations of the guidance. But, rightly or wrongly, we adopted a fairly simple approach. As we are so research intensive, there is a clear expectation that all members of academic staff will be research active (even if some do more teaching than others). We do not have anyone on a teaching only contract. The key question for us was therefore, "when does someone become an independent researcher?" For both the RAE and REF pilot, we took the view that anyone employed on the lecturer, senior lecturer, reader or professorial pay-scale was independent and therefore eligible, whereas anyone employed as a research assistant or research fellow is likely to be working on someone else's grant, and isn't a PI in their own right, and is therefore ineligible even if they have good publications. There was only a small number of eligible staff who were not returned to the RAE, however we used the same threshold to decide eligibility for the REF pilot.

3.6 Alignment with research information systems for RAE

We asked all pilot institutions to judge the extent to which the information requirements set out in the REF pilot specification aligned with the nature and extent

of the research information gathered for RAE2008. The intention was to determine how far the REF bibliometrics pilot demanded similar or very different types of data.

10 pilots stated that the types of data asked for by the pilot coincided reasonably closely to the information they collect for making submissions to the RAE.

We collect information on research outputs irrespective of whether we are taking part in an external exercise such as the RAE or REF pilot. There was thus a close correlation between the RAE, the REF and our normal operating procedures.

Totally, as we have used the same systems but in a lesser extent.

The REF Pilot submission was compiled exclusively from the repository of data collected for the 2008 RAE. However, only outputs that were submitted in the RAE (i.e. a very small proportion of the total pool) had been thoroughly validated by the Library. The data on research outputs submitted for the REF Pilot was far more extensive than that submitted to the RAE.

Hardly at all – that was a process devolved to UoA coordinators, entered onto the RAE software by Faculty/School administrators, and only covered four publications of submitted staff.

Six pilots stated that the data requirements were a partial match with their RAE procedures and information systems.

Personnel and outputs information submitted to the recent RAE was stored on a central database and could easily be accessed for REF pilot purposes. There was a significant amount of data required for the pilot, which was not available from the RAE database, however, the data provided a strong starting point from which to begin the data collection process for the pilot.

Another six stated that the requirements were a poor match, somewhere between hardly and not at all, with the current internal data collection procedures.

They did not coincide with collection of information on research outputs for RAE, the information on each publication required by the REF pilot far exceeded that required for RAE2008. Also, because the scope was wider (pilot data was to include all staff, not just RAE eligible), the RAE2008 material was not really used.

The purpose and criteria of the RAE informed the data collection carried out for that particular exercise (i.e. we focused on quality rather than quantity and the expectation was that the database would contain at least four publications per academic). Conversely, we aimed to collect complete research portfolios for the REF pilot (for the period 2001-2007).

Our collection of outputs for RAE2008 was not consistent across the Institution; we had complete records for a small number of departments only and had all RAE2008 submitted outputs for all submitted departments. Post RAE2008 we had begun a process of collecting all outputs from 2000 onwards for all departments and this process was hastened once we were included in the pilot.

3.7 Alignment with internal research information systems

We also asked pilots to comment on the extent to which the information required in the REF bibliometrics pilot coincided with the kinds of information they would expect to gather and report on as a matter of course for internal purposes.

The answers to this question clustered in three groups, which are broadly a group for which the requirements were reasonably well aligned with internal needs, a second group where there was some fit and a third group where the requirements go far beyond that which they gather and report on as a matter of course.

Five institutions stated that the REF information requirements were similar to those operated for internal research management, albeit the pilot had sought more detailed information on certain classes of data.

We use the same systems for internal management purposes, however we need to manipulate data to cater to different needs.

The arrangements are similar to those for internal research management.

The University's RAE submission database was originally set up to feed staff profiles on the web, but was enhanced to support RAE2008 planning and submission. Before RAE planning began in earnest many academic departments had individual systems for recording research outputs, but information has since been transferred to the central system.

Nine institutions stated that the REF information requirements had some similarities to those operated for internal research management purposes, and that this had provided a useful starting point for their submission.

Some output data was already being collected by the library to populate the university's Institutional Repository (IR) and also locally within some Schools, so the pilot was a useful tool to aid in this process and actually resulted in a much more unified system for collecting and managing research outputs. The pilot did not really coincide with any other internal management purposes.

Our publications policy requires academics to provide copies of papers prior to submission and again when accepted for publication. In practice compliance with this policy is patchy and although information on publications we had collected in the run up to the RAE was fairly comprehensive, some work was required to complete this information for staff who were not submitted (although eligible). As a result of the REF pilot we have refined our management practices for collection of publications.

Eight institutions stated that the REF information requirements were a poor fit with present arrangements, however a significant minority indicated that the pilot had been used to help to develop their central systems and research information management.

The work mapped reasonably well onto the implementation phase of the University's digital repository and enabled us to populate our new system with a significant volume of metadata records in a way that would not otherwise have been possible. The university would not have imposed such strict deadlines on departments at that time of year without having an urgent external requirement to do so. It was convenient for the university as a whole that the work fell when it did, although the work would not have taken place at that time, or so quickly, were it not for the REF pilot. Having said that, much of the additional work fell to a single individual, the research publications librarian, without whose hard work, long hours and endless goodwill the exercise would not have been manageable.

We had already begun a process of collecting research outputs for all staff for internal management purposes; our involvement in the pilot hastened the process. However, we did not at that point have annual monitoring of research outputs.

3.8 Challenges faced in preparing submissions

Overall, it seems the principal challenges related to any aspect where the REF bibliometric pilot specification sought information that went beyond the information maintained for the RAE, in terms of classes of staff, classes of output and the time period under review.

The most widespread challenge reported was the timing of the data collection across the summer months, which meant pilots had to carry out the work at a point in time when most staff and academics would be taking breaks for holidays. Concerns over timing and timescales for the most part reflect the challenge on the ground of working with research information management systems that are still somewhat rudimentary in terms of their inter-connectedness, scope, completeness and internal consistency and inter-operability.

The proposed timetable for data collection was the first obstacle as the activity at Departmental level in the summer months is significantly reduced. This was particularly challenging due to our intention to submit complete portfolios for

academics of the selected UoAs, considering that the only person that knows when a set of publications in any given period is complete is the academic him/herself.

Our challenges also included the multiplicity of 'data silos' and the inconsistency of the data stored in each one of them. This required the careful searching of publishers' databases and available catalogue records to ascertain which version of the data actually corresponded to the formal publication details. The unreliability of data made the verification of records both time-consuming and difficult. Inevitably, given the extremely tight schedule for data collection for the pilot, many of the records included in our submission were partial or contained some level of inaccuracy.

i) Significant gaps in metadata in internal systems following migration to new HR system around 2004;

ii) Vast gaps in metadata for research outputs in internal systems, wrongly attributed outputs, vast duplication, generally very unclean data due to poor data management;

iii) Too many outputs provided by Evidence/Symplectic with wild assumptions of connection – e.g. vast numbers of High Energy Physics publications with large numbers of authors and ambiguous names linked to researchers clearly not in that field.

The most challenging issue was in manually copy and pasting outputs data into the Excel spreadsheet from the IR and other online sources and then in providing the linkages in Table 3 (again, a manual, time-consuming process). Ensuring that the data was in the correct format as required for the pilot and that as much data as possible was included also took time to verify and complete. Some time was also spent in collecting staff CVs where outputs data were not already available. The collection of up-to-date staff CVs was quite challenging due to the time of year as many academic staff were on leave.

De-duplication of records, manual creation of the link table, searching out prior institutions, Excel automatic formatting converting Vol(No) into digit strings that did not revert on 'undo'.

The experiences reported are almost bimodal, with the two pilots with the most advanced information management systems stating that the data collection process had been straightforward, while most of the rest stated that the process of compiling a complete record of staff and outputs for the full period under review had been challenging and frustrating in equal measure. The smaller specialist research institutes had found the process relatively straightforward too, although the time period had posed challenges in both cases, as it required a higher degree of control over information on joiners and leavers than it had been customary to maintain. The challenge was most keenly felt by the largest pilots with the proportionately least developed information systems.

The period under review exacerbated an already demanding information management brief for the pilots, as their historical records and 'legacy' information systems tend to be weaker than current systems on almost every measure: fragmented, paper and digital, completeness of records, duplication of records, very limited metadata, consistency and accuracy of data entries and formatting, etc. This database problem was evident with regard to both research outputs and staff, with the great majority of pilot institutions citing difficulties with controlling for and capturing publications for both joiners and leavers.

The overall sense from the comments is that this challenge will gradually lessen in future, as institutions upgrade their research information management systems, in terms of its architecture and inter-operability and the completeness, accuracy and consistency of its content. However, legacy systems will continue to cause problems, and it is not clear to what extent institutions will underwrite the cost of digitising, infilling and integrating this archive of mixed and partial systems within their current systems.

The shortcomings evident in many information management systems had many knock-on effects with pilots reporting difficulties with creating complete records of outputs, with staff movements and older local databases posing particular problems, which only the smaller and more specialised institutions appear to have managed to cope with.

Time and timescales were arguably too ambitious given the information requirements and the state of the information management systems in the majority of pilot institutions, which had implications for the effort involved in constructing complete records, the accuracy/control over which was sorely tested by the requirement to cross-reference the two lists (people, papers) in Table 3. It seems likely that the data collection timetable led to pilots making submissions that were a little less complete and a little less internally consistent than people aspired to, however the great majority managed to compile and submit what they regarded as a reasonably complete list of people and outputs.

For the great majority, the biggest challenge revolved around the verification process, which was a novel, externally defined exercise that had to be conducted within four to six weeks in the run up to Christmas.

The short period of time available for the verification process does appear to have had a significant and widespread impact on behaviour with a majority of pilots indicating that they were simply not able to check most of the surprisingly large number of additional records provided to them by Evidence and Symplectic. Where people did look more closely at the additional records, there were concerns expressed regarding the criteria used to generate these 'missing' records, which were reported to be different to the criteria used to construct the original tables, and some concerns regarding the full use of the power of the Symplectic software to resolve certain problems automatically.

Since we did not seek additional publications, the initial submission was relatively straightforward. However, verifying the additional data provided by HEFCE's own data trawl threw up many issues, particularly as the search criteria used were substantially different to those used in our normal trawls. It was particularly noticeable that some of the capabilities of the underlying Symplectic software did not appear to have been used to resolve multiple references to the same publication. It was also quite unhelpful for the trawl to include individuals who were not RAE category A as no straightforward means of reconciliation was available. With 11,000 references to resolve and only 8,000 of these easy to resolve without massive additional resource it was decided that the remaining publications would remain unresolved, pending more information on how the data is to be analysed.

The data checking has been the most challenging part of the process. The main issues for the School were:

- *the fact that the data we were sent back did not include our own publication IDs, which made reconciliation more difficult*
- *the limited time available, which was partly due to the initial problems with matching, noted above. We therefore had to focus on the pending publications. Around a third of these were rejected (usually because they were for people from other institutions with similar names, or for incoming/outgoing staff who had some items credited to the School and others to another institution). We did not have time to check for false positives in the "approved" column. Larger institutions with more UOAs or less specialised areas of research would have found this a much bigger problem*
- *it wasn't entirely clear how we should deal with the incorrect data. For example, there were a number of papers listed twice – once as approved and once as pending. In these cases we rejected the duplicate entry. But different institutions may have taken different decisions. If we had all been doing the REF for real, we would need much clearer guidance to ensure consistency*

- *checking our data was time consuming. Having DOIs³ was helpful, but quite a number of these turned out to be incorrect. Even when they worked, it was still extremely laborious to cut and paste 100s of URLs⁴ into a browser and then try to work out institutional affiliations (which aren't always listed in the abstract the DOI took you to). Again, if we were doing this for real, it wouldn't be practical.*

3.9 Insights gained from checking additional records

The REF bibliometrics pilot adopted a hybrid approach to the data collection and verification process for several reasons, which were firstly a more complete and robust final database, secondly an opportunity to explore the relative merits of a top-down versus bottom-up data collection strategy and thirdly to permit individual pilots, and the wider community, to benchmark their research information management systems.

The first two points are a matter for HEFCE and its contractors, while the third issue, insights and lessons learned, is a matter for this feedback exercise, and all pilot institutions were asked what they had learned from the process of receiving and checking the additional records provided to them by the HEFCE contractors.

As with other elements of the data collection phase of the pilot, several pilots remarked upon the openness of the task specification, and its rationale, for the process of constructing and checking the additional records. This has influenced the lessons learned with most pilots choosing to offer advice to HEFCE on how this procedure might work better in future rather than commenting on their own insight. One respondent stated that the exercise might better be thought of as a sharing of experiences with HEFCE and its contractors.

Turning to lessons learned, the majority of pilots appear to have gained a deeper understanding of the universality of the challenge of maintaining complete, accurate and consistent bibliographic records, whether one is a smaller university or a world-leading information services company such as Thomson-Reuters.

There are numerous comments on tactics and challenges for improved 'disambiguation', the act of resolving any uncertainty around exactly which author or publication is in the list, which range from tactics for dealing with foreign names to the sufficiency of the matching criteria used on this occasion. This narrative suggests the verification exercise might very well have improved people's command of the art of disambiguation, at least amongst the staff directly involved in the verification exercise.

When linking publications to authors with similar names we found it was quite easy to do for people with very common names like e.g. Jones because these authors tend to publish with two initials. For staff with fairly common names like e.g. Johnson we found it more difficult because we still had to distinguish between several people but they often published with only one initial. The Symplectic system for checking needed an exact match on the title and so a few publications that had the word letter or review on the end of the title, since that is how they were stored in the repository, had to be checked as pending publications.

Trying to verify data based on any form of name mapping produces too high an error rate to count as statistical noise. Even with improvements in the technical mapping this approach is still questionable. Manual checks are resource intensive and in some difficult cases with common surnames may require actual author involvement. Unique staff identifiers are important and really the problem will not be completely solved until each researcher has an internationally recognised unique transferable researcher identifier.

³ A digital object identifier (DOI) is a permanent digital identifier given to an object. In this case, the DOI is used to give a scholarly paper or article a unique identifying number that anyone can use to obtain information about the publication's location on a digital network.

⁴ In computing, a Uniform Resource Locator (URL) specifies where an identified resource is available and the mechanism for retrieving it. In popular usage, a URL is often used to refer to a Web address.

We discovered the importance of knowing the former employer of the staff member, and although this was not always possible, it was beneficial if we did. It indicated that the system used to retrieve data needs to take into account the first name of the individual not just their initials, and if they looked at first names it may be possible to easily differentiate between the sexes. If this had been done it would have cut down the number of erroneous records sent to us for checking.

The additional outputs provided by Evidence only amounted to about five per cent of the University's total pilot submission, and most of these were very easily accepted or rejected via the online Symplectic database. Most staff submitted only had one or two publication queries, if any at all; two staff had a significant number of queries but that was because only their RAE submitted publication data was available prior to the submission as they had left the University. The lesson from this is that the more output data which is available on the IR, the more complete and accurate the data and the less onerous the data collection task.

The responses suggest that most pilot institutions will have concluded that automated checking, using a tool similar to Symplectic, can be helpful to maintaining the completeness and integrity of their internal records. However, it has also left the pilots with a view that a top-down, automated process of author-publication matching, as could be one potential option for the REF proper, is likely to have a very high error rate, with numerous duplications and misallocations. A significant minority argued that such a procedure needs to be more closely specified to avoid creating unnecessary work while doing little to increase confidence in the completeness of the final record. Several people remarked on the inefficiency of the search strategy used on this occasion, which it was felt did not exploit fully the capabilities of the matching tool, with for example, records being matched using initials rather than first names and not being cross-checked by discipline.

The process used to identify authors was overly inclusive. Around 70% of the authors HEFCE's contractor identified had no connection with the university.

We are rather concerned that a significant volume of work done to submit a full set of publication records to the pilot was unnecessary – as outputs not recorded in the Thompson database were not used in the analysis.

It also helps to be very close to the data and to know your staff. If we hadn't been working with this information for a long time it would have taken much longer to check the additional outputs.

We identified 11 papers where the author had been incorrectly identified as being one of our academics. Had we not had a very good idea of all the areas our academics publish in these 'incorrect' papers would have been difficult to identify.

A more sophisticated process is required to associate staff to outputs. Given the scale of the data and inaccuracies returned, significant disambiguation of the data is required. Staff surname to co-author surname is not a sufficient matching criterion for an exercise on this scale. Individual co-author to institution association is required (for example for collaborative work) as part of the matching criteria. In addition, local academic knowledge is the only reliable method to validate the association of staff to research outputs. This was not possible in the limited time available.

There was a general view expressed that this kind of matching exercise would always require a final check by people with good knowledge of the academics in question, arguably carried out at individual faculty level. Distributed human checking of a central software verification process was thought to be unavoidable, but ultimately undesirable. For two commentators, the experience had underlined the importance of getting the data right at the time of entry, in an observation that has echoes of the debate in manufacturing between advocates of total quality management – right first time – and more traditional approaches to quality control (inspection, rejection, rework).

Data cleansing is a real issue. The process was totally reliant on local knowledge and staff memory to identify spurious links and deal with synonyms.

Disambiguation is a big problem for researchers with common surnames; confirmation that our internal output data is not very robust but equally neither is a lot of the Thomson data.

We need to have more incentives and checks in place to ensure that our data collection is comprehensive. We need access to the raw data of any essential proprietary datasets used by HEFCE with any access arranged as part of existing licence structures.

The odd misattributions generated by the automated process for identifying additional outputs were corrected manually by our REF administrator. This was only possible because the number of outputs returned to us was comparatively small. We are aware that this would not be possible for a larger dataset and find this a source of concern. We found that the information available in Web of Science was not all of the quality we would have expected. We believe that accepting all proposed outputs as 'ours' would have an undesirable (distorted) effect on individuals' profiles and could have a negative effect on our academic's staff perception of the processes associated with the pilot.

There was a presumption that the verification process would help pilots to gauge the quality of their information systems and procedures. In practice, only a minority of pilots commented directly about the performance of their research information systems, as a result of the verification process, which is to say whether their internal records are more, or less, complete and robust than they had believed to be the case.

We learned that our database is fairly robust, but that there are certain areas, such as computing, where the data held by Evidence is more patchy and will need to be resolved before the REF.

Implementation of the REF and participation in the pilot provides the University with a strong impetus to develop a comprehensive research management system.

However, many of the responses do imply a recognition, as a result of the combined experiences of the data collection and verification exercise, that their respective research information systems are in need of substantial further development, in terms of procedures and content, and integration with other internal systems.

The question posed was rather open and there were only three cases where respondents elected to provide any indication of the attrition rate, which indicated 70-85% redundancy in the additional records provided. These results might not be representative, however acceptance rates for all pilots will be reported separately by *Evidence Limited*.

The great majority commented on the laborious nature of the task, and perhaps unsurprisingly, the larger research universities felt this most acutely. There was no readily discernible pattern evident in these experiences relating to the differing levels of sophistication of research information systems.

3.10 Person days involved in the collection and verification of submissions

We asked each pilot institution to provide us with an estimate of how many person days' work had been involved in the whole data collection process, including checking additional records.

It cannot be underlined strongly enough that the numbers returned were estimates, arrived at in different ways by people at institutions of differing sizes and with different ambition levels for the pilot. As such, they are best viewed as indicative.

The replies also make it clear that the internal costs were tracked to a greater or lesser degree by different institutions, and in particular that the more extensive the team involved, the more uncertain were the estimates.

Figure 4 presents the estimates for each of the 22 institutions, with the data collection costs shown in the bottom data series and the verification costs in the top series. It should be noted that the responses for all institutions were used to interpolate estimates of the split between these elements in three cases.

The average was around 65 person days, with a median of around 50 person days. However, the individual estimates varied widely and ranged from 10 person days, in two cases, to more than 200 person days in one instance.

In total we spent about 10 person days collecting and checking data. It should be stressed however that we did not check the data we sent to HEFCE and we checked only the easiest data which we received from HEFCE. Systematically checking the data would have multiplied this figure many times over, and was not feasible given the very short timescales involved.

Approximately 20-30 person days were involved in the initial collection and submission of the REF Pilot data. A further 20 person days were involved in the feedback to Evidence process. The limited time spent on the feedback process was due entirely to the very limited time available. Therefore a total of 40-50 person days were involved in the whole process. If more time had been available for the second part, then this figure would have increased substantially.

Up until the end of August 08 we had 3 members of staff (2.3 fte [full-time equivalent]) working on the pilot although only 1 fte was full time on this project. From September to the end of December the 1 fte worked for three full weeks on checking the additional outputs.

Approximately 160 person days.

The data-collection process was reported to have taken around 50 days on average or 40 days typically, but with an equally large range from around eight person days to more than 130. The verification process required an additional 15 person days on average, with an even greater spread, ranging from two person days to 70 person days. The verification process amounted to around 20% of the total effort, on average. However, there is a broad spectrum of effort, ranging from 5% to 66%.

The differences appear to reflect the combined effects of two loosely related factors, which are the sophistication of the pre-existing research information system and the degree to which the pilots sought to create the most complete and robust list of authors and publications. The second factor is a question of intent and relates both to generating the initial database and the subsequent checking of the additional records. There is also a scale effect, however that appears to be a weaker influence.

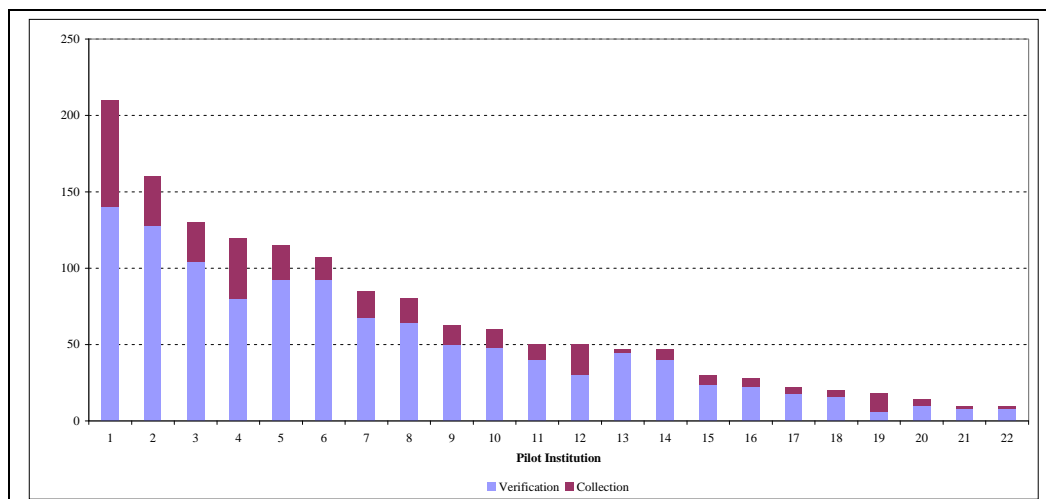
The responses also point to another possible issue for HEFCE when considering the workload implications of any future REF bibliometrics element, which is that the compilation of a database of authors and publications is a task that institutions will increasingly ask of themselves, to inform institutional management and not just to meet the requirements of HEFCE or other funding bodies.

The estimate for the days required for the additional records checking stage is 40 days FTE. This consists of a number of people contributing a day (mostly in their own discipline area) plus the hours for the 1.5 core editorial repository staff. Unlike the earlier stage of the process all of this is additional work. During the data collection period at least some of that activity was work that may have been done eventually anyway, albeit with a longer timescale.

In this sense, the cost to produce such a list should reduce in time, as systems evolve and procedures bed down, and a reasonable proportion of this 'cost' would have to be borne anyway, to support ongoing monitoring. Therefore, the additional cost to the system of a REF bibliometrics exercise ought to be rather less than the costs estimated

here. The verification process however appears to be a function that few of the institutions would carry out as a matter of course, although that might change in future of course. As such these costs, should such a process be run for the REF bibliometrics proper, will be wholly additional, and attributable to the national assessment exercise.

Figure 4 Person days involved in the collection and verification of submissions



3.11 Implications for development of research information management

We asked pilot institutions what if anything the bibliometrics pilot might imply for their planning and development of the university’s research information management, during the next 12-24 months.

The majority of pilots indicated that they are planning to make several changes to their information systems and procedures, as a result of their participation in the REF bibliometrics pilot. There is a wide range of changes planned or foreseen, with the great majority involving incremental improvements to current research information management arrangements rather than major investment or development projects. The one exception is institutional repositories where four pilots plan to implement major new systems and two others have concluded this is going to be necessary at some point in the near future.

It is clear that the University will need to have a complete, up to date, accurate and accessible central record of publications for REF and that sufficient staff will be in post and adequately equipped and trained to manage the administrative requirements of the exercise once they are known.

The University has decided to create a central publications database and to look at how that and other research support systems can interoperate. Although the design and requirements of the bibliometric component of the REF remain to be seen, it is likely that we will need to develop our staff data systems to allow us more easily to identify eligible staff. Specific developments will be required should the REF require us to gather data on prior/destination institutions, alternative surnames/aliases and Early Career Researcher status.

The incremental changes include various means by which to improve data quality (complete, consistent, accurate, timely data input, etc), extend and deepen existing databases and integrate key databases and systems. More specifically, the anticipated changes include but were not limited to:

- Implement institutional repositories or other central databases (four)

- Introduce new procedures to update and check the completeness/quality of departmental databases (four)
- Improve integration and inter-operability across information systems operated by different departments and central functions, such as research, HR, library, etc (three)
- Extend and deepen existing staff and publications databases (three)
- Subscribe to Symplectic or similar (two)
- Introduce regular reporting of publication data, as part of the current process of reporting on research income and bids to senior management (one)
- Research & Enterprise Office is looking to produce recommendations to departments and HR in order to ensure better research information management in the future (one)
- Awareness raising to bring the repository to the attention of all staff and to underline the importance of data quality to the university (one)
- Increase amount of information we capture about joiners and leavers (one)
- Develop some kind of modelling capability (one)
- Introduce a bonus system of sorts to encourage academics to keep records up to date (one)
- Implement a university-wide requirement to deposit all publications in the repository (one)
- Identify and plug key gaps in existing information and begin process of enrichment of the archive, with more metadata and more granularity (one)
- Provide training on institutional repositories and other aspects of the research information systems, for administrative and academic staff (one)

Not all pilots plan to make changes. Nine of the 22 pilot institutions stated that they expected to make few or no changes to their respective research information management systems as a result of their participation in the pilot so far, and that they were waiting to hear the detail design of the final bibliometrics arrangements.

Unfortunately, our experience of the pilot has not changed our post RAE 2008 plans regarding research information management. We had hoped that by being involved in the pilot we would gain a clear understanding of the system requirements for the REF. As no clear guidelines have emerged and the possibility of a REF exercise in 2010 we feel we have no choice but to continue with our development of our existing system in the hope that it is compatible with any REF system that is introduced. It has encouraged the introduction of annual monitoring of outputs.

The experience of the REF pilot demonstrates the need to install a centralised repository according to the specifications for REF. However, this will involve obtaining resources at a time of considerable funding constraints and also overcoming considerable local resistance from those who see no need for a common system. We are extremely concerned, given the low base form which we are starting, that REF parameters have yet to be published particularly given the timeframe for the 2010 bibliometrics exercise.

A minority stated that the exercise had confirmed that their present system was deficient in various ways, and that they clearly had to do something to improve the situation significantly in the medium term. In two of these cases, people spoke of a dilemma in that the improvements needed were far reaching, in terms of systems and procedures, and would cost a great deal, a proposition that would be tough to sell while the key design parameters for the REF bibliometrics remain open.

3.12 Challenges and opportunities of aligning with REF

We asked all pilots what they foresee as the challenges and opportunities arising from the need to align their internal research information management system with major external requirements, such as those that will be defined by the REF.

This was a somewhat speculative question, and most people expressed a high degree of uncertainty given the fact that the REF data requirements have yet to be finalised and specified in detail. Notwithstanding this fact, most people did offer a view on possible future challenges and opportunities.

We preface the discussion with the observation that the great majority of pilot institutions stated, directly or indirectly, that they expect their research information management systems to align with the data requirements defined by REF. One commentator noted that the evolution in their research information management system tracks the evolution in the data requirements of the RAE, and that they would expect that to continue with the REF in the future.

We are currently reviewing our research management systems to ensure they are aligned and hold all the required information. The problem is of course we still do not know what will be required for REF. However, what little we do know will be incorporated into our systems during the next 12 months. We are also looking at ways of incorporating citation factors into our research output database.

Our current systems were developed with regard to RAE2008 and the extent of the challenge in aligning our internal systems with the requirements of the REF depends heavily on the scope of the REF proper in terms of census dates, staff eligibility, research income, PGR students etc. Implementation of the REF and participation in the pilot provides the University with a strong impetus to develop a comprehensive research management system.

On challenges, there was a reasonably consistent view expressed in various ways by the great majority of respondents, wherein pilots foresee a need to develop their research information systems and procedures in order to meet the anticipated REF requirement to present data on all staff and all publications. This will be a demanding programme of technological and organisational development, which will have substantial implications for university resources and culture/internal relationships. Some view this systems requirement as essentially a one-off cost, albeit a significant one, to make the transition to the new arrangements. Others fear there might be a step change in the ongoing costs of research information management, centrally and across academic departments, whether that is staff time or licences or ongoing capital investment. In several cases, people expressed concern over the affordability of such a transformation at a time when finances are expected to worsen and central budgets are already tightening.

The opportunities are to centralise data on all publications our academics publish and make these available to external stakeholders. We can also better monitor activity in different sections of the institution, publishing this information internally and creating more internal competition for quantity and quality of publications.

The challenges are to develop a system that is sustainable with the resources available – this will partly depend on HEFCE's decisions over the distribution of 'R' funding – investment in new research management systems will come out of our Research Support Office budget which is highly dependent on QR [quality-related] funding.

Several people anticipate the need for a conscious effort to embed the notion of future income being determined in part at least by the performance of all staff and their full publication repertoire.

The main challenge will be embedding the concept of recording publication details for all staff in the university. This will be something that even individual researchers will have to take part in (making sure they report to departmental administrators, who will in turn report to the central Research

& Enterprise Office). Also, it is important to embed this change as soon as possible, even if we don't have any further information on the future of research assessment from HEFCE. The main opportunity that we will gain from this is the ability to collate research assessment data in the future with minimal effort.

Several people noted the different levels of aggregation, units of assessment, between the outline proposals for the REF and the university's current internal monitoring and reporting (schools, research groups). This might lead to changes in the granularity of internal reporting over time, but, in the first instance, is likely to mean additional work – post-processing – will be needed in order to meet the reporting requirements of the REF.

Technical challenges are not seen as being unduly problematic, as compared with the organisational and financial, although the issues around system integration and interoperability do recur.

Challenge will be to ensure that systems, e.g. for HR and publications data, are interoperable. Key issue here is to provide a single data entry point for staff for internal and external research reporting and web page content.

As far as opportunities are concerned, there were just two cited and these were closely related: the first is the opportunity to add a report on publication outputs (numbers in the first instance) to complement existing management reports on research inputs, which presently focus on research bids (pipeline) and research income; the second is the anticipation of an improved internal capacity (research or planning offices) to model future income and inform strategy deliberations with more and better analyses of past performance.

3.13 People and functions involved in research information management

In response to advice from our earlier orientation interviews, we asked every pilot institution whether or not the REF bibliometrics pilot had led them to expect to expand the number and range of people/functions involved in research information management.

This was another prospective question concerning what people expect to happen in the future, and in this case the answers were pretty similar across the group. The views expressed here tally with the answers regarding future challenges and the widespread expectation of a more extensive research information system being slightly more demanding in resource terms than the present arrangements.

16 of the 22 pilot institutions replied saying that they expect to see an increase in the number and range of people involved in the research information management in the future. Just two said the work was likely to be done by the same people and functions. The other four replies were unclear as regards any likely increase.

We would expect our existing processes for the annual review of research performance (which involve staff at University [PVC], College [Head of College/College Director of Research] and School [Head of School] level) to take more cognisance of bibliometrics data, should this be built into REF, and which would mean we would need some capability to monitor citations performance. We would also anticipate the need for staff in Library Services to become more involved in this area and in general there will be a requirement to educate staff about bibliometric and citation issues.

There is a clear requirement for individual researchers to take more responsibility in managing their own research information and profiles and in monitoring bibliometric information. This is going to require a large culture change. Library Services is to extend its existing responsibility for bibliographic data into bibliometrics. Other central divisions such as Academic Services, Finance, HR, Registry, Development Office, office of the Vice-Provost (Research), in addition to faculty research managers/school research coordinators, will continue to have input into these areas.

There was a pretty consistent picture as to which groups or functions would be doing rather more in the future, although the increase is quite likely to be small in proportionate terms:

- In all cases, the library is expected to play a fuller role in future, with an increased workload associated with maintaining more comprehensive and up to date repositories and bibliographic databases
- In the great majority of cases, academics themselves are expected to have to do a little more to ensure all publications are deposited and metadata is captured
- We imagine there might be a substantial task to fill the gaps in the archive, whether that is whole records or additional data on existing records
- IT and HR are both widely expected to become more involved, however this is likely to be sporadic and concentrated on periods of system development and wider integration
- In several cases, people expect the research and planning offices to be doing rather more with these new facilities, producing periodical reports on performance and ad hoc analyses of strengths and weaknesses and trends, as an input to senior management's oversight and regular strategic planning

3.14 Internal capability to interrogate bibliometric data

We also asked pilots whether they envisage creating an internal capability to interrogate bibliometric data (i.e. run citation analyses), or if they might secure that capability through other routes, whether through the market or through some user-group or other mechanism.

15 of 22 stated that they would expect to need to be able to interrogate bibliometric data in future.

We would definitely need this capability.

It will be necessary to do this once the rules are made available. The choice of mechanism will depend on the options available at the time.

We would like to have such a capability.

As regards the route by which such analytical services would be secured, the answers were more wide ranging, with several stating that they would expect to manage this internally while others stated that they had or would enter into commercial contracts with specialist service providers and toolmakers. Most however are undecided, and awaiting more information on the REF bibliometric requirements and any national efforts HEFCE might make to facilitate community-wide access to services and data.

[We] would not want to duplicate any efforts that HEFCE and its contractors might be making as part of the full REF. There is also the issue of copyright to consider.

Given that every HEI will be trying to develop the same systems and same capabilities at the same time, the scope for collaboration needs to be considered seriously. But previous attempts at large-scale collaboration across the sector have not been successful, and the fact that institutions all have different systems in place (or will want to implement new ones that reflect their own internal structures, coding systems etc) would be a major stumbling block. There needs to be some form of national licensing agreement with the selected supplier (Thomson or Elsevier).

Yes, the sector will want to do this and support in doing this will be needed. If this support were not provided, every HEI would need contracts with Symplectic, Evidence/Thompson Scientific, Leiden, etc to provide such data and analyses. Note that this is an increased cost; would this cost more than running an RAE every five to seven years?

Five of the remaining seven stated that they were undecided, at this point in time, as to whether such a capability would be needed and if so what it would look like; the need and solution were contingent on the final specification for the REF bibliometrics element and its choice of citation analyses.

It is too early to decide this. We would only take this step once it is certain that HEFCE are including citation analysis in the REF, once it is rolled out in 2013-14.

We have not made a decision on this, but it would seem likely. In common with other pilot organisations we see a role for the funding council in negotiating access to this data on behalf of the sector.

The last two respondents simply stated that they did not envisage a need for such a capability at this time.

3.15 Implications for faculty and other researchers

We asked pilots what if anything the pilot implies for the university's faculty and other researchers, in the future, in terms of research information management and associated procedures.

While bibliometrics is likely to be one of several types of ‘intelligence’ used by the actual REF peer review panels in order to render judgements on an institution’s research quality in a given field, most pilot institutions appear to expect it to be a significant influence, for science, technology, engineering and mathematics (STEM) subjects at least. Moreover, bibliometrics is likely to become more prominent over time, with successive iterations of the REF, as confidence builds in the robustness of the underlying data (which are improving all the time) and calibration and interpretation improves.

As such there is a pretty well universal commitment across the 22 pilot institutions that whatever the detail design of the actual REF, they must continue to bring information on research outputs under greater control, and that means for all staff and for all research publications.

The experience of the pilot suggests that everyone will have a role to play in future research assessment, including individual academics and this means paying closer attention to publications records.

The onus on maintaining accurate research information (publications) is already on researchers – but the pilot has highlighted the need to integrate this task as part of our academic colleagues’ standard routine. Systematic and regular updates will only take place if we as an institution provide the right tools and the right incentives – we are working on this.

The ability of the institution to store and analyse information about all aspects of research activity needs to be more robust. Systems need to be fully integrated. Data needs to be updated regularly as part of a culturally embedded process. Data needs to be re-used for different purposes more efficiently.

There is a reasonably clear set of expectations emerging, wherein academics will need to more aware and accepting of the:

- Importance of individuals’ publications data to future institutional research income
- Primacy of central systems within an institution’s research information management arrangements
- Importance of keeping one’s own publications records up to date
- Importance of data accuracy and consistency
- Importance of becoming more familiar with bibliometrics and citation analyses, and any implications therein

Several pilot institutions make the point that the academic input must be kept simple and light, supported by a fair amount of process automation/software tools, to ensure higher levels of intrinsic data integrity and to avoid disincentives. One pilot institution suggests that there is a case for HEFCE to take a more proactive role in developing community-wide solutions, as it believes the data management tools used for the pilot worked badly and would be both a disincentive and source of additional costs and data quality problems.

The REF Pilot confirms the need to have accurate and complete bibliographic records. It also confirms the need for automated processes, which are supplemented by academic and other inputs.

It will be important for researchers to interact with the central publications database to ensure that its currency is maintained. However we intend to minimise the work that individual researchers will have to do. It’s difficult to speculate without knowing the shape of the REF and whether it will, in fact, differ significantly from the RAE. However, if it turns out to be peer review but with more citation analysis (i.e. RAE+) then the burden could potentially be greater for institutions and for individual researchers.

The ultimate impact will depend on the final shape of the REF. But experience to date indicates that the bibliometric analyses cannot simply be done automatically (as was originally envisaged by the Treasury). Producing accurate data which is acceptable to the sector is likely to involve significant investment across the sector in new or improved systems; there could also be significant licence costs to access citations data; and the issue of data quality would also impose additional costs. If metrics are to drive funding, then it might be possible to justify this cost. But this may not be the case if they are being used to provide one indicator, which panels then interpret alongside others as part of a much broader peer review process.

In causing the pilot institutions to look much more closely at the completeness and integrity of their own databases, and indeed those of the major information service providers, in particular Thomson-Reuters WOS, the REF pilot has arguably increased anxiety levels about the good sense of using bibliometrics for determining future research income. Because their systems have gaps, duplications and inaccuracies, there is anxiety that without very significant additional work their institutional publication listings will be compromised and as such the results of any subsequent citation analyses must be arbitrarily incorrect, even if the denominator is sufficiently aggregate to be robust.

It is complicated and will not be an entirely 'machine-driven' process. We have very serious doubts about the potential for an accurate return without pouring very significant resources into the exercise, certainly more than we do to run a 'normal' RAE. If, as has been indicated in some quarters, the citation analyses are now being seen as merely an adjunct to peer review, we cannot see any resourcing 'discount', just greater demands overall.

Together the 22 responses imply a pretty much universal ambition to get to a much better place with research information management, where an institution will be able to submit a return to HEFCE that is as close as is reasonably practicable to being 100% complete and accurate (this viewpoint has some resonance with the environmental protection principle, Best Available Techniques Not Entailing Excessive Cost). In addition, the REF pilot has led a minority to conclude that the evolution of their systems and related behaviour is a major undertaking that cannot be justified if citation analyses are going to be downgraded in the actual REF to be a simple adjunct of peer review.

3.16 Other implications so far for researchers

We asked pilot institutions whether they foresee any other implications so far for researchers.

The majority of respondents foresee no obvious additional implications for researchers, although a significant minority did flag a concern about the wider impact of using bibliometrics on publishing behaviour (e.g. what and where to publish). This however is a subject that we will return to in the second round consultation, when reviewing the implications arising from the results of the pilot's citation analyses.

3.17 Other preparatory work

To bring feedback on experiences to a close, we asked pilot institutions, whether, on the basis of their experience of the pilot to date, they were considering undertaking any other work on their research information management systems to prepare for the REF.

The great majority said that no other REF-specific preparatory work was in hand, and that they did not envisage doing more until the detail design of the REF was confirmed.

A minority did cite some examples of other preparatory work:

- Further development of training materials and training courses for staff, on using the institutional repository and more general awareness raising on bibliometrics and citation analyses
- The upgrading of related information systems, and in particular the HR systems to ensure all staff, contractors, research students were captured and that the treatment of joiners and leavers was recorded more systematically and comprehensively
- The creation of a cross-departmental REF strategy group to track developments and deliberate implications and responses
- The creation of an integrated management information unit, which will pick up REF requirements as one of its responsibilities
- Benchmarking other indicators, on for example the volume and sources of research income

Others simply noted that there was ongoing development on the research information systems, which would be pursued independently of the decisions on the final design for the REF: this included adding in more external data sources and further automation of the process by which output data are imported into the publications system, reducing the reliance on and burden for academics. One of the specialist research institutions reported that it had begun a round of individual consultations between the Vice Principal for Research and all researchers, to discuss publication history and strategies.

3.18 Practicable suggestions for the design of the future REF

We asked pilot institutions what practicable suggestions they had for the design of the future REF in terms of making the data requirements and data collection processes more manageable for their institution.

A reasonably consistent set of related recommendations were offered, which hinged on the clarity of the final specification, the timeliness of the final detailed design and the length of time given over to the preparation of submissions:

- Fix the design parameters as soon as is reasonably practicable, giving institutions as much notice as possible to give time to determine how best to develop their systems to fit, with due acknowledgement of the fact that for the majority implementation will take years and not months
- The final design must be accompanied by a specification for the data requirements and planned analyses that is crystal clear and comprehensive, with good definitions and case examples
- Give longer lead times for the preparation of the first full submission, as the timescale for the pilot proved problematic for many and meant that many made less complete and weaker (in terms of data quality) submissions than they would have wished to and in several cases institutions elected to make narrower submissions than would be necessary in a full REF bibliometrics exercise
- Give more consideration to the timing of the period in which submissions are to be made, as the pilot exercise involved data preparation across the summer months, when staff holidays caused difficulties, and the verification process was conducted in the run up to Christmas, which was problematic for reasons both to do with the absolute duration available for the task and clashes with the timing of parallel submissions to other funding bodies

Other recommendations, which were not necessarily additive or complementary, included a request for HEFCE to bear in mind that bibliometrics is only one part of the proposed REF and that institutions might very well need to implement several development programmes in parallel. This has implications for what is practicable, for both reasons of finite resources (people and cash), and for the balance of priorities.

Two institutions recommended the development of common, web-based tools to facilitate consistent and robust uploads of institutions' data to HEFCE, or its contractors running the bibliometrics element of the REF.

One institution recommended HEFCE revisit the issue of author identification, from the perspective of establishing broad use of a standard methodology, and as a platform for improving the criteria/routines by which authors are matched to publications, and the creation of a rating system to indicate confidence in each match.

One institution recommended that HEFCE should provide data to the institutions for them to manage the verification process, given their firm belief that local knowledge – people not software – is essential to ultimate accuracy.

One pilot institution noted that there would need to be confirmation that the funding bodies servicing the devolved administrations (e.g. the Scottish Funding Council) were indeed planning to proceed with the same research assessment system as the rest of the UK.

There was also concern expressed regarding the plan to press ahead with a developmental REF bibliometrics exercise in 2010, which it was believed would risk a dash to evolve systems at the level of individual institutions and the resulting poorly planned, possibly incompatible systems. One respondent also foresees a possible negative impact on value for money, through rushed decision-making and implementation and possibly even inflationary pressures as suppliers struggle to meet the sudden and dramatic expansion in demand.

3.19 Alignment of REF requirements and internal research information needs

We asked pilots to offer any practicable suggestions they might have as to how HEFCE might refine its REF data requirements such that they more closely align with the university's internal RI needs.

The great majority stated that to a large extent their internal requirements evolve to reflect external needs, whether that is HESA, HEFCE or the research councils. In light of this, few felt they were in a position to specify what more HEFCE might do to improve alignment, and simply restated their previous recommendations regarding the timing of the communication of the final design and absolute clarity and transparency around that detail specification.

Some suggestions were made however, and these are presented below.

- Five institutions suggested that there would be value to the community overall if HEFCE were to give further consideration to the compatibility between the various reporting cycles and data requirements made by other research funding bodies and statistical agencies (e.g. HESA)
- One institution suggested that there might be value for all parties, HEIs and funding bodies, if there were to be a blue-skies consultation or debate around the more strategic question of what they are collectively thinking and wanting to do with regard to the development of their internal research management systems
- Another institution suggested that data requirements tend to change in use, when certain data prove to be less helpful than others or even redundant and 'missing' indicators reveal themselves. From a learning and evolutionary perspective, it would be good if HEFCE were able to find a way to keep the data requirements under review, with some kind of ongoing consultation, in order to optimise the metrics and minimise wasted effort
- Another institution suggested that HEFCE should explore the possibility of making a community-wide deal with one or more of the information services businesses that maintain the key databases, WOS and SCOPUS. The purpose of this would be to enable individual institutions to gain access to those services on

more advantageous terms than they might negotiate bilaterally, which would in turn encourage the use of these reference data on an ongoing basis rather than intermittently and with extended periods in between usage. Such an outcome would possibly help to more quickly embed these data and analytical techniques within the mindset of senior research managers and the academics they serve.

3.20 Other issues arising

We asked pilot institutions whether there were any other issues arising from the pilot so far, not already dealt with under previous topics, and which they would like HEFCE to pay more attention to in developing the REF.

Half of the pilots said they had covered the lessons learned and matters arising in their replies to earlier questions and made no further input.

Six institutions took the opportunity to cite their more fundamental concerns about bibliometrics and in particular its value for money when judged against, on the one hand, the time, cost and disruption implied by a presumed need to bring all HEIs up to a similar and high level of data completeness and integrity, and on the other, the expectation that peer review (and other metrics) might continue to be the principal determinant of funding outcomes. One institution invited HEFCE to consider whether it might run some kind of cost benefit analysis of the added value of the bibliometrics, as compared with the results for RAE2008. Five other institutions took the view that the implementation of the REF bibliometrics will amount to a significant additional cost, RAE plus bibliometrics, and was not likely to deliver the promised economies.

Several institutions took the opportunity to restate the one or two key points they had made earlier, and in particular the critical need for HEFCE to reach a final decision as soon as possible and to be crystal clear about its requirements. On the subject of clarity, one respondent said she would welcome clarification of the means by which the different indicators and forms of assessment – peer review and citation analyses – will be reconciled, or added up. Another institution stated that there was a great deal of urgency around finding a workable solution for precise matching of authors to publications, as the pilot arrangements had been inadequate in their opinion and would weaken the credibility of any resulting analyses.

Two other institutions suggested that HEFCE would need to keep in mind its commitment to develop a broad-based assessment system, making use of a range of types of intelligence and data, from bibliometrics to research income to postgraduate researchers to peer review.

Two institutions asked HEFCE to remember to consider and make a decision on the treatment of a range of other issues, ranging from the treatment of non-standard publication types (e.g. policy studies published in the grey literature⁵) to inter-disciplinarity to issues of equality and diversity (part-time workers, people returning from career breaks, etc). Another institution asked for early guidance on what HEFCE planned to do about the fact that a significant proportion of certain universities' outputs, including journal articles, was not well covered by WOS.

⁵ Grey literature is a term used variably by librarians and research professionals to refer to a body of materials that cannot be found easily through conventional channels such as publishers. Examples of grey literature include technical reports from government agencies or scientific research groups, working papers from research groups or committees and numerous other forms of contract research reports.

4. List of abbreviations

fte	full-time equivalent
HEFCE	Higher Education Funding Council for England
HESA	Higher Education Statistics Agency
RAE	Research Assessment Exercise
REF	Research Excellence Framework
UoA	Unit of Assessment
WOS	Web of Science

Technopolis Ltd
3 Pavilion Buildings
Brighton BN1 1EE
UK
T +44 1273 204320
F +44 1273 747299
E info@technopolis-group.com
www.technopolis-group.com