Modelling national research assessments in CERIF

Stephen Grace, Richard Gartner
Centre for e-Research, King's College London

1 Periodic research assessment

Assessment of research activity in universities is of growing interest across Europe, and internationally. In this paper the term universities will be used as shorthand for all higher education institutions. The global league tables of the Times QS World University Ranking and Shanghai Jiao Tong Academic Ranking of World Universities are studied, debated and publicised annually. At national levels, assessment exercises are used in different ways. Some aim to help entrants choose universities based on notions of popularity and employability. Others may be designed to help improve university research activity and define the performance targets. Still others are formal exercises which determine the allocation of funding.

A recent report from the Expert Group on Assessment of University-based Research (2010) places this growing trend in the European context of a drive to modernise universities in the knowledge economy (Lisbon Strategy). National schemes devised for local reasons, and taking account of local contexts, will likely be supplemented by future European-level assessments which allow for comparison of research groupings across national boundaries. The Expert Group suggests a methodology through its multidimensional matrix for assessing the full range of research activity that reflects the various uses and users.

Whether this is developed into a Europe-wide system or not, there remains the issue of exchanging data in consistent formats to allow aggregation and comparison. CERIF is well placed to become the means by which such data can be collated. The Readiness for REF (R4R) project in the United Kingdom aims to create a method by which British universities can collate and submit their data to the forthcoming national research assessment.

2 Research assessment in the United Kingdom

Research assessment has been conducted in Britain since 1986 in six cycles of assessment under the heading Research Assessment Exercise (RAE). The method of assessment became increasingly formalised, and at the heart is an expert review by disciplinary peer review of the quality of research activity and outputs. The latest (RAE 2008) collected data in five areas, often referred to by the forms which detailed the information submitted by universities

- staff details (forms RA0 and RA1)
- research outputs (form RA2)
- research students and studentships (forms RA3a and RA3b)
- research income (form RA4)
- research environment and esteem (forms RA5a, RA5b and RA5c)

Universities chose which evidence to submit in up to 67 subject groupings called Units of Assessment (UoAs). Universities could submit up to four outputs per staff member – typically research publications like journal articles and monographs, but also patents, software products, performances, artefacts and exhibitions. Measures of esteem could include editorships, membership of academies or governmental panels, and awards. Overall, RAE collected and assessed quantitative and qualitative measures of research quality, and the burden to the sector was seen as substantial.

3 Research Excellence Framework

Following a government review in 2006, a new assessment exercise was initiated called Research Excellence Framework (REF). This aimed to provide benchmarking of quality against international standards, and to provide the basis for distributing funding, but also to "reduce significantly the administrative burden on institutions in comparison to the RAE" (Higher Education Funding Council for England, 2007). Initially it was thought bibliometrics would provide a reduced burden, by offering a proxy for qualitative assessment in citations and journal rankings. A pilot exercise showed that bibliometrics were "not sufficiently robust at this stage to be used formulaically or to replace expert review in the REF" (Higher Education Funding Council for England, 2009). REF will now be composed of three elements

- outputs (research publications and other outputs)
- impact (demonstrable benefits to the economy, society, public policy, culture or quality of life)
- environment (the institutional research environment and infrastructure)

The current proposal is that these elements will be weighted in the proportions 60%: 25%: 15% and submissions will be made in 2012 for outcomes to be published the following year.

4 CERIF4REF: a bespoke CERIF schema

In an attempt to simplify the process of presenting data in a format suitable for inclusion in the REF, and to obviate the need for institutions to learn the CERIF standard in depth in order to make use of its many advantages, it was decided to design a new schema, CERIF4REF (C4R), which would act as a mediator between the REF and CERIF standards. This new schema would form an easy-to-learn interchange format in which any data for the REF could easily be encoded, and from which instances in the REF and CERIF formats could be generated automatically using XSLT stylesheets.

Unfortunately for this project, the requirements of the REF schema have not yet published, and are not due to be made public until the final quarter of 2010, well into the latter stages of the R4R project. It has, therefore, been necessary to design the schema using the RAE2008 schema as its model: evidently, there will be some changes necessary when the REF schema itself is published (most notably the introduction of data on impact which is a new feature of the assessment exercise), but it to be expected that the majority of the information required will be similar and so that the CERIF4REF schema designed with the old standard as its model can be revised when the REF standard itself is finally published.

4.1 The CERIF4REF schema

The CERIF4REF schema itself encodes five main groups of data, which correspond broadly to the subdivisions of the RAE2008 schema:

- research groups
- research staff and students
- research outputs
- funding
- research environment and esteem

Research groups includes basic information on the institution's research units, including their names and the units of assessment (the subject-based committees who will assess their outputs for the REF exercise) to which they have been assigned. The *research staff and students* section details the personnel whose research outputs are to be submitted for consideration: in addition to basic biographical information, including names, dates in post and research group memberships, this section allows linkages between individuals to be made to indicate, for example, research student/supervisor and co-authorship relations. The research outputs themselves are detailed in the next section, which primarily contains bibliographic information necessary for their unambiguous identification, also allows the flagging of sensitive, interdisciplinary and sensitive material. The *funding* section encodes information on research income, including research studentships and funding from external sources. Finally the *research environment and esteem* section allows prose statements of the institution's overall research environment and an analysis of the esteem with which its output is held in the academic community.

The conversion to the RAE2008 and CERIF schemas is handled by two stylesheets written in XSLT. Much of the information required for RAE2008 relies on aggregations of data held in the CERIF4REF file: for instance, overall head counts and full-time equivalent (FTE) figures are generated by counting occurrences in the *research staff and students* section. When data of this kind needs to be broken up on a year-by-year basis, this is easily done by selecting on the basis of service start and end dates recorded for each staff member or student. The RAE2008 stylesheet also creates automatically any identifiers (such as sequential A-Z identifiers for each research group) which are required by the RAE2008 schema.

Although the CERIF4REF document produces a single (usually much smaller) XML file in the RAE2008 format, the same information requires nineteen CERIF files for it to be adequately encoded. These are generated by a single XSLT stylesheet which makes use of the <result-document> feature present for the first time in XSL 2.0: this allows output to be redirected to files specified in the stylesheet itself rather than in the processing instructions which invoke it. The following CERIF files are produced by this process:-

- cfOrgUnit-CORE
- cfPers-CORE
- cfFundProg-2ND
- cfPersName-ADD
- cfOrgUnitId_OrgUnit-LINK
- cfPersName-OrgUnit-LINK
- cfOrgUnit_ResPubl-LINK
- cfPers_ResPubl-LINK
- cfResPubl_Class-LINK
- cfOrgUnit_FundProg-LINK
- cfResPublBiblNote-LANG
- cfResPublAbstr-LANG
- cfClassTerm-LANG
- cfOrgUnitResAct-LANG
- cfOrgUnitName-LANG
- cfPers_Pers-LINK
- cfPers_ExpSkills-LINK
- cfExpSkillsDescr-LANG
- cfResPubl-RES

As can be seen, almost half of these are LINK files used to join elements found in the CERIF information environment to each other: for instance, cfPers_ResPubl-LINK is used to link persons to their research outputs. CERIF's linking mechanism requires the detailed use of identifiers for almost every component: these are readily generated automatically by the XSLT stylesheet, so obviating the need for the user to bother with this (potentially error-prone) process.

It is envisaged that this schema will simplify considerably the process of generating and formatting the information required for the sometime onerous REF exercise. In particular, it offers a relatively simple XML schema to which information already held in CRISs and other internal databases of institutional research outputs can readily be mapped: in this respect, it is easier to use than either the RAE2008 (which requires extensive calculations and the pre-processing of information) or CERIF formats. By designing output to map to this simple format, institutions will then have their information ready for submission to REF and for exposure to the wider information community via CERIF after running two simple stylesheets.

It may also be feasible to encode information directly into the CERIF4REF schema, although this could potentially be cumbersome as the file for any given institution is likely to be large, including, as it does, biographical information on all of its research staff and bibliographic details of every output. Its usage is therefore envisaged very much as a half-way house between the two schemas themselves, a method of synthesising information held in a wide variety of disparate pre-existing systems via a process which should be as simple and non-technical as possible. Future systems may well attempt to implement a similar interchange format as part of their make-up, but until then this represents a usable approach to collating data from a messy information environment.

4.2 Recording Impact

REF is composed of three assessment areas, outputs, impact and environment. Impact is a new concept in REF; the funding council intends that significant additional recognition will be given where researchers have built on excellent research to deliver demonstrable benefit to the economy, society, public policy, culture or quality of life. This approach has not found universal favour, especially in the arts and humanities or by those conducting blue-sky research. Impact measures will likely be more qualitative than those for outputs and environment, comprising an impact statement and one or more case studies. R4R has prepared a separate XML schema for modelling impact based on the guidance available to date, ready to incorporate into CERIF4REF when the requirements are finalised in late 2010.

5 Using institutional repositories

The rationale of the R4R project is that institutional repositories (IRs) can be used to collect data needed for REF submissions. This raises questions about the relationship between repositories and CRISs. Such relationships are of widespread interest currently, with recent work by Knowledge Exchange’s CRIS/OAR Interoperability Project (https://infoshare.dtv.dk/twiki/bin/view/KeCrisOar/WebHome) and euroCRIS establishing the CERIF-IR Task Force.

IRs exist at almost all universities in the United Kingdom, with established mechanisms for working at the institutional level – though these mechanisms vary widely. JISC has supported the establishment or enhancement of institutional repositories in various ways; direct co-funding of new repositories, enhancing the services such repositories offer, development of tools to aid functions such as deposit and preservation, training and programmatic support through the Repositories Support Project (http://www.rsp.ac.uk/) and Repositories Research Team (http://www.ukoln.ac.uk/repositories/digirep/index/Repositories_Research). University repositories follow different models, on a spectrum from Open Access (providing free access to the full text of research articles in line with the 2003 Berlin Declaration on Open Access) to a publications database recording all outputs of staff whether or not peer-reviewed or available as full text. Most rely on voluntary author deposit but a few have institutional mandates requiring staff to deposit full texts, usually on acceptance for publication.

Some institutions have CERIF-compliant CRISs as well as IRs, though in the UK such CRISs are rare at present. Others have built CRISs or research databases inhouse as a means to gather information for RAE/REF which are not built on the CERIF model. Some institutions have established repositories but do not have institutional-level research information systems. Some, especially smaller institutions, have neither. To cope with various system infrastructures, HEFCE as the administrators of RAE provided several different means of submitting data. It offered bulk upload options for XML, Excel, Access, RTF, tab-delimited, EndNote™ and Reference Manager™ formats, as well as web services for programmatic uploading in the relevant format. To support those who did not have institutional systems holding the relevant data, HEFCE also created an online system for direct data entry.

It is easy to foresee different scenarios, and progress paths, given this wide range of institutional approach. These will be explained in turn, with consideration given to their current and future engagement with CERIF.

- repository plus inhouse research database
- repository and CRIS
- repository but no CRIS

5.1 Repository plus inhouse research database

Universities with established IRs sometimes also have an inhouse research database, which typically covers research grants and, less frequently, research studentships. Such databases are built on enterprise platforms like Oracle when seen as corporate assets, or using workgroup databases like MS Access (or even spreadsheets) when owned and maintained by a single department such as a research office. These databases clearly have some characteristics of a CRIS, but do not use CERIF; in fact, many tend to have ad hoc structures not conforming to external standards, or to follow the RAE data model if primarily designed to meet that need. IRs and research databases are not connected, so publications as the results of research are not linked to information about the projects from which they arose.

CERIF4REF offers a common means of data exchange between repositories and research databases, so that publication metadata could be enhanced with information on the project funding and staffing. As staff become familiar with CERIF though mapping internal structures to C4R, institutions might be encouraged to formalise their research database as a CRIS using CERIF: this has been the approach at King’s College London, which is currently testing a CERIF-enabled version of its database.

5.2 Repository and CRIS in concert

Some universities with established repositories also have CRIS systems. Only a handful have CERIF-compliant CRISs but others are procuring such systems. For good CRIS-IR interoperability there needs to be authority control of persons in the repository, so that authors of publications are identified as members of the organisation. This is by no means common, and may require data reconciliation within the IR – perhaps by recourse to the human resources system or through manual intervention. In addition, publications are not typically associated with a project in repository metadata. Much work is to be done to enhance the data in repositories if they are to relate to research information entities in CERIF-compliant CRISs. CERIF as a data exchange format between the two systems will be of use here, whereby either system may inherit data from the other to enhance understanding of research information.

A key question will be the relationship between the two systems. Does the CRIS rely on the repository for publication-related data, or does the repository need information on the wider context of individual publications? Often the two systems are managed and overseen by different functions within universities, with the library managing a repository and the CRIS coming under the control of research support offices or other central functions. Interoperating between the two systems can thus be as much about organisational practices as technological ones. Guidance is needed on ways to make these links straightforward, and CERIF4REF may be useful here.

5.3 Repository in the absence of a CRIS

Smaller and less research-intensive universities may not have research databases of any kind, and would not see the value of a CRIS for the organisation. They do have repositories, though, and several are interested in the prospect of enhancing or extending the capabilities of repositories to handle other research-related information. R4R will develop plug-ins for ePrints, DSpace and Fedora Commons platforms to support such institutions. The content or data model will be extended to handle summary information of the core CERIF entities of Person, Project and Organisational Unit, as it is required for the national research exercise, and the CERIF4REF schema will be able to generate data for exchange.

Repositories are embedded institutionally, with staffing and technology support for a sustainable service. Repository managers have developed awareness and training programmes so that the repository is seen as part of the institutional landscape, and thus able to take on the enhanced role of research information management. The task will be to populate such enhanced repositories with core entity data, either by importing from other systems or by extending the deposit function. Here, authors would associate person, project and organisational unit information with their publications at the point of submitting them to the repository. Managers will then have the task of verifying the reliability and authenticity of such information, but this is a process universities are familiar with as a result of preparing submissions for research evaluation exercises. Enhanced repositories thus become a sustainable source of research information, and act as a *de facto* CRIS. The ability to generate CERIF data allows for future migration to a CRIS, should that become an institutional requirement.

5.4 Developing relations between CRIS and repository systems

CRIS and IR systems are accepted parts of university life, albeit with different development histories, technologies and remits. They form part of the infrastructure needed for research assessment, can work in tandem to manage information flows and assets within institutions, and beyond them to other stakeholders. Data can be linked to that held by funders, for instance, so that universities can reduce the need for recreating information on grants. At the same time, the funder can track the outputs funded by its programmes through the semantic association of publications with grants – either by harvesting such data or searching for it in university systems.

Universities will continue to develop their repositories as a showcase of research. As public-facing systems, repositories can be used as windows on other research-related information especially where CRISs are seen as internal management systems. CRISs can receive or link to enhanced information about publications, research datasets and other forms of research output from repositories, and thus provide a fuller map of research activity for internal and external audiences.

6 Applicability beyond the United Kingdom

The R4R project has been focussed on the British context of a renewed research evaluation exercise. The CERIF4REF schema offers a useful model for other countries to map their assessment systems to CERIF. By using CERIF as a common standard it would be possible over time to compare across jurisdictions – using the British system to benchmark German universities, for instance, or comparing universities across the Norwegian and Danish systems. This presupposes that the data required in each assessment is available across national boundaries, and it does not perform the review process itself. Nevertheless, being able to compare peers across borders would be useful at the European level – to strengthen the European Research Area, for instance. Countries could compare the insights from other assessment methods by expressing existing data in other countries' formats.

On a more local level, universities could profile themselves against their partners either as a challenge or to show the aggregate strength of the network. King's College London has existing partnerships with Humboldt-Universität in Berlin and Sciences Po in Paris, and the three institutions could profile their collective strengths against different assessment regimes. The Expert Group report offers a framework for research assessment which could provide a useful way forward for such transnational comparisons, as an alternative to the national regimes.

References

Expert Group on Assessment of University-Based Research (2010): Assessing Europe’s university-based research. Brussels: European Union.

Higher Education Funding Council for England (2007): Future framework for research assessment and funding (Circular letter 06/2007). Bristol: HEFCE.

Higher Education Funding Council for England (2009): Report on the pilot exercise to develop bibliometrics indicators for the Research Excellence Framework (Report 2009/39). Bristol: HEFCE.