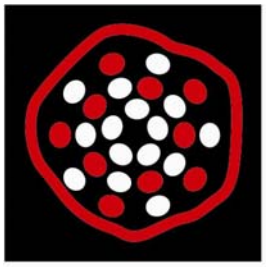
	HERA Humanities in the European Research Area **HERA website: *www.heranet.info***
Deliverable number	HERA D4.2.1
Title	FEASIBILITY STUDY: THE EVALUATION AND BENCHMARKING OF HUMANITIES RESEARCH IN EUROPE
Work Package	WP4
Actual submission date (contractual date)	August 27 2007 (draft); January 21 2008 (final version)
Organisation name(s) of lead contractor for this deliverable	UK Arts and Humanities Research Council (AHRC)
Author(s)	Carl Dolan
With the help of	
Nature	Report
Status	Final version
Dissemination level	public
Abstract	

Contract no: ERAC-CT-2005-016179

SIXTH FRAMEWORK PROGRAMME



HERA

Humanities in the European Research Area

FEASIBILITY STUDY:
THE EVALUATION AND BENCHMARKING OF HUMANITIES RESEARCH IN EUROPE

final version, January 21 2008

Arts and Humanities Research Council (AHRC)

work package 4

D4.2.1

1. INTRODUCTION

This report examines the feasibility of establishing a common approach to evaluating the outputs and outcomes of humanities research in Europe, including the possibility of defining robust benchmarks for cross-national comparison.

The need to address this issue is made all the more urgent by two related research policy developments. The first is the increased emphasis that national governments are placing on a transparent method of performance management in the higher education sector to focus strategies, improve results, and ensure accountability of public funds spent on research. (Some of these trends are reported upon in *Section 3*). The second is the substantial funds that are now available for humanities research through EU funding programmes and agencies such as the Framework Programme (FP) and the European Research Council (ERC). Therefore at both national and supranational level, it is necessary to ensure that the outputs and outcomes funded by these means are assessed with due regard to the distinctiveness of humanities research, to avoid future allocations of funding between disciplines or thematic areas being driven by a flawed evidence-base.

The main conclusion of the report is that due to the current data deficit that exists in tracking the outputs and outcomes of humanities research at national level, reliable cross-national comparisons or benchmarks are not available. This is an acute problem in the most comprehensive databases traditionally used to benchmark national systems of research – the bibliometric and citation databases of commercial companies. This is likely to remain the case in the short-term without a concerted initiative on the part of national and or supranational funders.

However, the report has not stopped short at that negative outcome, and has attempted to formulate common principles for the assessment of humanities research which could be accepted by all national and supranational funders. This common framework would be the basis of more specific initiatives to be undertaken and would represent the first step toward a common approach if not agreement on the precise quantitative benchmarks to be used.

The framework consists of a series of recommendations that open the report in *Section 2*. These recommendations were discussed and approved by representatives of the HERA network at 2 workshops on impact and quality assessment issues that took place in March 2006 and January 2007.

Section 3 of the report looks at the state of research assessment in the humanities in Europe and further abroad today. This section builds on the results of the HERA impact and quality assessment survey conducted in February 2006 and which forms the basis of the HERA report on impact and quality assessment practices published in May 2006. It extends that analysis with a further survey of national systems of assessment which supplements the previous survey and provides a more comprehensive overview. It provides an indication of the type and level of data that can serve as a platform for further initiatives.

Section 4 examines the arguments for and against the use of bibliometrics in the humanities and looks at promising developments in that field for the development of tools adequate to the assessment of humanities research. The conclusion is that despite promising aspects to these developments, there is no immediate solution to the data deficit without the mobilisation of significant resources in the higher education sector as a whole.

The importance of peer review is one of the central recommendations of *Section 2. Section 5* looks at some possibilities for international benchmarking which avoid the problems associated with a bibliometric approach and retain a central function for peer review and direct academic input. The section focuses on possibilities for systematic peer review of entire programmes and sectors that does not require the direct peer review of each output (which has been the central feature of the UK Research Assessment Exercise). This, in the view of report, is the only way that such methods would be feasible.

1.1 HERA: Background

HERA (Humanities in the European Research Area) is a network and partnership between national funding agencies for the Humanities. The Consortium has 14 full partners and two sponsoring partners from 15 different countries all of which are intertwined with their own national research communities. In addition, the European Science Foundation (ESF) offers a forum of 31 research councils and acts as a pan-European member in HERA.

The HERA Consortium signed a contract with the EU in 2005 to undertake a number of activities designed to enhance large-scale cross-border coordination of research activities within the broad field of the humanities. It has built on the work done by the European Network of Research Councils in the Humanities (ERCH), which was

established in October 2002 as a forum for the chairs of the humanities research councils across Europe.

The network aims to exchange information and best practice on issues such as national and international peer review, programme management, quality and impact assessment, and benchmarking, thereby ensuring the highest excellence in nationally funded research as well as research conducted within the framework of HERA activities. The ultimate objective of this EU-funded project is to coordinate research programmes in a cumulative process leading to the initiation of two joint research-funding initiatives.

An outline of HERA and its objectives can be found on the project website (www.heranet.info) and the project work involved is fully described in the Description of Work. Its main aims are to:

- stimulate transnational research cooperation in the humanities
- overcome the historic fragmentation of humanities research
- ensure that the European Research Area (ERA) and EU Framework Programmes benefit from the relevance and dynamism of humanities research
- advance innovative collaborative research agendas
- improve cooperation between research funding agencies and co-ordinate existing funding programmes

2. RECOMMENDATIONS

The following recommendations were provisionally agreed at HERA workshops on impact and quality assessment in March 2006 (recommendation 8) and January 2007 (recommendations 1-7). The recommendations aim to form the basis of a common approach to the assessment of the outputs and outcomes of humanities research across Europe and further abroad by both national and supranational funding agencies and ministries.

1. Research is a continuum

There is no fundamental difference in the nature of the research enterprise in science, technology, engineering and mathematics (STEM) disciplines on the one hand, and the humanities on the other. Rather, these disciplines represent a continuum of research endeavour, along which methods and resource requirements vary in ways that do not map easily onto the current subject divisions. The demand for research inputs ranges along the spectrum from resource- intensive disciplines, like chemistry, archaeology to non-resource-intensive disciplines such as mathematics and philosophy. The disciplines that make up the humanities are distinctive in their approaches and concerns but should not be considered exceptional.

2. Disciplinary variation

This distinctiveness is also apparent at the level of discipline. Although it should be possible to devise a broad framework of assessment that applies to all disciplines, the nature and scope of the elements of that framework should be sensitive to the distinctive characteristics of each discipline such as the size of the community, its demand for inputs, the inputs available to it, its publication patterns and the nature and organisation of the research process.

3. Assessment of disciplines

The most appropriate level for international comparison is at the level of cognate disciplines or large research groupings. However, institutional data and other publicly available data should be used to inform the assessment wherever possible.

4. A holistic approach

The holistic approach involves the combination of a number of elements to gain a more accurate picture of research quality and performance. There is consequently no one indicator that is appropriate for measuring research quality. There is a danger that, in focusing solely on the quality of outputs, valuable research activities and collaborations that reflect on the quality of a research group or research environment are neglected as they are not well served by that focus.

5. The need for peer judgement

It is likely that credible quantitative methods for the assessment will emerge in the medium to long term (e.g. improvements in the coverage and quality of citations databases). However, at this point in time, it will be necessary to retain the application of human judgment through peer review processes to gain an accurate picture of the quality of humanities research.

6. Proxies for peer judgement are available

The indicators chosen as part of the assessment framework should reflect the multiplicity of peer-review systems which are already in place and are an integral feature of academic life, e.g. peer-review of books and journal articles other outputs, peer-review of project-based and infrastructure grant applications, some measures of esteem. It would be important for a system of quality benchmarking to use these current practices and other peer review systems in order to be feasible and to avoid overburdening peer reviewers.

7. The importance of quantifiable evidence

Research indicators have an important role to play in research quality assessment, particularly in providing the evidence to inform the judgements of reviewers. Evidence of research quality in the humanities is gained through the following elements:

i. Research outputs
ii. Spend on research infrastructure and environment
iii. Wider social, cultural and economic significance of the research process
iv. PhD completions
v. Peer-reviewed research income

vi. Esteem indicators

These elements should be used together to create appropriate indicators for each discipline that allow comparisons to be made across countries and regions. The evidence would allow international panels of reviewers should make final judgements on quality of the research according to a standardised scale.

8. The importance of measuring wider social, cultural and economic impact

A framework for the measurement and benchmarking of would be desirable in the long term and work on the comparability of case-studies and workable indicators would be welcome. The humanities research community could benefit greatly from an evaluation framework that looked at its wider impact, adjusted appropriately to the context of its research process and the specific social, cultural and economic domains where the impact takes place. However, work on this has just started in Europe. A HERA survey indicates some progress being made on this issue in the Netherlands and UK but little substantial work being conducted elsewhere. In is beyond the scope of this report to provide an exhaustive survey of this growing field, but a good overview of recent developments in this area is available at www.eric-project.nl Our conclusion is that in the medium term indicators for international benchmarking of impact will not be feasible.

9. The importance of better data collection

There are two elements to this recommendation. The first concerns harmonization of data-gathering and reporting activities to ensure better comparability. It is clear that research funding agencies already gather large amounts of data through the monitoring regimes that are standard practice. This information is gathered through mid-term and end-of-award reports and other regular reviews and evaluations of programmes and funding instruments. Standardisation of these forms - with a view to robust international comparisons and all agencies' data requirements in the short- to medium-term - would be a relatively straightforward task.

The second element concerns the need for data on humanities research to be routinely collected by national statistics agencies. OECD derives the data it uses for international comparisons from these sources, but basic information on humanities research is not available from all countries. Even the countries which

do submit data on humanities research do not do so consistently over time. This makes systematic comparisons of trends in humanities research across countries virtually impossible. There should be concerted lobbying of national statistics agencies for a more consistent approach to collecting data on humanities research

3. STATE OF THE ART

3.1 Evaluation by humanities funding agencies

As part of the work package on impact and quality assessment, HERA officers conducted a survey of impact and quality assessment practices in all HERA partner agencies and other major humanities funders in Europe and further abroad. The results of this survey were published on the HERA website in June 2006. (http://www.heranet.info/Default.aspx?ID=106)

The survey revealed a variety of overlapping and complementary practices. Apart from the centrality of peer review, there was no core method or approach to evaluation which was shared by all or by a majority of agencies.

There were a number of general conclusions that emerged from the survey:

(i) There is at least a common element of practice at the level of data-gathering. All agencies require an end-of-award report from their award holders. Agencies differ in how the report is processed. For the majority of agencies this is simply a matter of noting the completion of the project and its outcomes. The report may or may not be seen by an academic panel. Only 7 agencies in 6 countries insist that the award-holder's self-assessment is peer reviewed to give the award a quality rating.

(ii) The widespread practice of collecting end-of-awards does, however, indicate that there is a large amount of raw data on the outputs and outcomes of humanities research which at a very fine level of detail (see Annex 3 of the Survey Report). It was recommended at the HERA workshop on impact and quality assessment that this routine information collecting should be standardized in such a way to provide more easily comparable data-sets. Such data sets should also be stored in such a way as to be readily retrievable and manipulated.

(iii) There are 3 main levels of evaluation. Firstly, there is the project level evaluation of end-of-award reports to which reference has already been made. Secondly, there is programme-level evaluation which is conducted by 6 HERA partners. This represents the largest area of

common agreement and best practice, since virtually all those agencies who have run large-scale programmes in the humanities have conducted evaluations of this type. While all such evaluations involve – at a minimum – panels of peer reviewers looking at end-of-award reports, there is no widespread agreement on the other main elements of such an evaluation. There is disagreement on the need for site visits and whether panels should be given bibliometric information, for example. There is also no consensus of whether such panels should have international membership. This report recognizes that there is a core of best practice here and tries to enlarge the consensus about what is seen as best practice in such exercises and apply it to international benchmarking.

(iv) The third level is discipline-level evaluation, which is only carried out by 2 HERA partners. This differs from the second level in scope, but in terms of how the evaluation is carried out it is structurally similar to programme-level evaluation.

3.2 System-wide evaluation

This section looks at ex-post research evaluation systems in Germany, Netherlands, Finland, Poland, Australia, Belgium, Slovenia and the UK. These represent the main types of research assessment of the HE sector in countries which operate a dual-support system for publicly-funded university research. This section looks at one side of dual-support only: the evaluation of research in the university sector for the purposes of the distribution of 'block funding' (also called operational or core funding) or improving universities' research strategies (formative evaluation). It does not look at the ex-post evaluation of project-funding distributed by funding agencies.

The first distinction to be made is between countries where evaluation is not linked to funding allocations (Netherlands, Germany) and those where it is (all the rest). In the former countries, the evaluation serves a more formative purpose, with the aim of informing research managers' decisions and improving institutions' research strategies. Unlike the Netherlands, where the specific differences between institutions are recognized and they are evaluated along four different dimensions, a further aim of the

German funding ranking is to provide an explicit comparative measure of the performance of German institutions.

Secondly we can distinguish between those where the entire basis of the evaluation is quantitative indicators (Germany, Poland, pre-2004 Australia, Belgium) and those where peer-review still plays a large role (Slovenia, Finland, Netherlands, UK).

Position of the Arts and Humanities

In none of the countries surveyed have separate indicators for the arts, humanities and social sciences been developed. In this they are typical of research evaluation systems more generally. Perhaps the only example of special treatment of the arts and humanities is in Slovenia where a *lower* weight is placed on journal articles published in the Thomson-ISI Arts and Humanities Citations Index (AHCI). It should be noted, however, this weighting takes place in a wider evaluation context which includes an element of peer review. This safeguard ensures there is no systematic bias against humanities research.

Most countries are aware of the limitations of citations indices for the humanities and social sciences, and this is why they have been excluded from Belgium's (Flanders) experiment with bibliometrics as a method of distributing funding. Germany's 'Funding Ranking' also comes with a number of health warnings regarding the use of indicators for these fields. While there is explicit recognition that the humanities (and indeed the social sciences) is badly served by indicators which have been developed with an eye to the communication and research practices of the natural sciences, there has been a lack of sustained initiatives to develop indicators specific to these fields (the major exception being the ESF/HERA-sponsored European Reference Index for the Humanities (ERIH)) The consequence is that humanities disciplines either receive a derogation from quantitative evaluation (as is the case in Flanders and the Netherlands) or these measures are applied uniformly to all disciplines. This results in the failure to capture the full productivity of humanities research, thereby undervaluing its achievements.

Main indicators

The main indicators used can be classified as measures of research inputs (e.g. third-party income, staff numbers) or measures of research productivity (e.g. publication counts, number of graduate students). Measures of research quality are more rarely used and are confined to controversial measures such as the use of citations (Belgium, Poland), esteem measures such as invited lectures (Netherlands) or data on the internationalization of the research base, such as evidence of researcher/student mobility or the degree to which publications are in international publishing houses/journals (Germany, Slovenia, Finland). A summary list is provided in the table below:

Research inputs	
Operating income	Germany, Netherlands,
Staff numbers	Germany, Netherlands, Poland, UK
Research income	
Third-party income (excluding RC income)	Germany, Poland, Australia, UK
Income from international sources (EU etc)	Germany, Poland, Australia, UK
RC income	Germany, Poland, Australia, UK
Internationalisation of research	
Co-operation in RC-funded national networks	Germany, Finland
Visiting lecturers/ incoming researchers	Germany, Poland
Incoming graduate students	Germany
Numbers of researchers in international networks	Finland, Poland, Slovenia
Researcher productivity	
Numbers of monographs	Netherlands, Finland, Poland, Australia, Slovenia, UK
Numbers of journal articles	Netherlands, Finland, Poland, Australia, Slovenia, UK
Publications in leading international journals	Germany, Slovenia, UK
Indexed in international bibliographic	Slovenia

database	
Book chapters	Australia, UK
Published conference proceedings (refereed)	Australia, UK
Bibliometric analyses (e.g. citations)	Netherlands, Belgium
Patents	Netherlands, Finland, Poland, Slovenia
Development of databases etc	Slovenia
Organisation of significant national or international conferences	Poland
PhD completion rates	Netherlands, Finland, Poland, Australia, UK
Masters degrees	Finland, Poland, Australia
Measures of research esteem	
Number of RC reviewers	Germany
Membership of national evaluation bodies	Poland
Invited lectures/keynote speeches	Netherlands
Awards and prizes	Poland
KT Measures	
Integration of research into teaching	Slovenia
KT/relevance to industry	Finland, Poland, Slovenia
Expert reports commissioned	Poland, Slovenia
Wider dissemination of research findings	Poland

3.3 Country Files

Germany

Although certain Länder - such as Lower Saxony - have developed performance-related measures of research quality, there is no Federal–wide assessment of university research performance *that is related to the distribution of federal funds.*

That said, the German Research Foundation (DFG) produces a periodic report –'Funding Ranking' - on the performance of German universities which ranks them according to a variety of criteria. This

ranking does not carry any consequences for the allocation of funds.

The data used for this ranking all comes from outside the universities themselves, which in the opinion of the authors makes the data more robust. The data can be divided into two main groups a) third-party funding and b) the degree of internationalization of German research. The sources are as follows:

(i) The Federal Statistics Office provides data on expenditure (third-party income, administrative income, regular core income) and Full Time Equivalent (FTE) staff. The data does not explicitly distinguish third-party or core income for research and teaching.
(ii) DFG approvals: research projects approved by DFG broken down by funding scheme and discipline
(iii) Numbers of DFG reviewers
(iv) Co-operation in DFG-funded co-coordinated programmes (e.g. research training groups, collaborative research centres) broken down by type of programme and discipline. Institutions are ranked in terms of their 'centrality' in these funded networks.
(v) Data from the Alexander von Humboldt foundation on visiting researchers, as well as information on each institutions AvH fellows and prize winners.
(vi) Reports from the German Academic Exchange Service (DAAD) on international scientists, students and graduates in Germany broken down by research area, country of origin and institution.
(vii) Data on EU funding
(viii) Bibliometric data: publications in international journals gleaned from the Centre for Scientific and Technology Studies (CEST) in Switzerland.

The 'Funding Ranking' is a series of tables comparing all German institutions that receive more than Euro 500K of funding from DFG under each of these headings. However there are also summary

tables which compare universities performance in all of these categories (unweighted).

The most striking thing about these summary tables is the degree to which universities' success in gaining DFG approvals is correlated with their ranking in the other indicators. When DFG approvals are normalized according to the number of professors in each university, the correlation is not so strong, but is still high (see Funding Ranking pp.131-33 http://www.dfg.de/en/ranking/download/dfg_funding_ranking_2003.pdf)

The other striking conclusion is that the significance of third-party income (as measured by the Federal statistics office) - either in absolute terms or relative to numbers of researchers – varies greatly between disciplines and is not a good indicator of high research activity, especially given the fundamentally secondary role of this stream of funding in the humanities and social sciences.

Perhaps the most innovative aspect of the ranking indicators is (iv) which, along with a network cluster analysis, provides evidence that it is not only inputs such as funding and staff at institutional level which are important but (as one might guess) regional scientific structure, and the structure of opportunities provided by collaborations with neighbouring universities and non-university research institutions.

Netherlands

Ex-post evaluation (Quality Assessment of Research) in the Netherlands is carried out by the Association of Netherlands Universities (VSNU[1]). It uses a method of informed peer review similar to the UK Research Assessment Exercise (RAE). Evaluation is carried out not for the purpose of allocating funds but for formative, strategic purposes.

[1] Veriniging van Univeritataen (VSNU) - http://www.vsnu.nl/web/show/id=26111/langid=42

The spectrum of academic research in the universities under evaluation (13 out of the 14 Dutch universities) is divided into 27 disciplines, which are further subdivided into research programmes that are carried out at departmental level. Each of the 27 disciplines is evaluated by a separate Review Committee (RC). Excluding the chair, the committee members are all external to the Dutch HE system. Disciplines were not evaluated simultaneously, but were evaluated on a rolling basis over the course of 4 years (up to 2003). Since then, the VSNU has aimed for a lighter touch with self-evaluation by institutions every 3 years, followed by external peer review evaluation every 6. The system also allows institutions to ask the RC for targeted, confidential advice.

The RC takes into account the following information provided by institutions on research performance over a 5 year period:

- an overview of academic staff
- a summary of the programme mission and research plan;
- content of the programme and its main results;
- a list of all publications;
- list of five selected key publications from the programme;
- other indicators of quality and reputation (such as patents, invited lectures, etc.)

Supplementary information is also gleaned from

- interviews with Research Programme leaders and site visits
- Bibliometric analyses

The RC assesses the research performance of each research programme along four different dimensions (unlike the uni-dimensional rating for the RAE). These are:

1. Scientific quality: In the assessment of quality, attention is focused on quality
measures, such as originality of ideas and methodology, the importance of research output for the performance of a discipline,

the scientific impact of the research activity and the international prominence of the research group.

2. Scientific productivity: This aspect relates the inputs to the outputs of research activities. Number of staff and the size of the monetary resources allocated to research are considered to be the measures of input. Important indicators for outputs are the number and nature of publications in refereed and non-refereed scientific journals and books, the number of dissertations, patents and invited lectures.

3. Scientific relevance: For this aspect, the research is assessed in terms of its relevance to the advancement of knowledge in the discipline and science in general, and the possible impact and application for future technology as well. In addition, the benefits to society are also considered.

4. Long-term viability: This aspect is assessed based on the submitted plans and ideas for future research. In addition, the publication policy of the research group, the coherence of the programme and the continuity of research lines are also assessed.

Each of these dimensions is assessed according to a 5-point rating system (from 'poor' to 'excellent'. The peer reviewers are asked to use the international excellence of the research, gauging whether the research output is better or worse than the 'world-average'. It is unclear whether there is any benchmark for this 'world average' other than the reviewer's expert knowledge of the field. Section 5 will look at more systematic methods of establishing such benchmarks through peer review.

RC's also have the discretion to assess (3) and (4) by reference to a research group's own mission, recognizing that a single standard for each university won't apply. Given the relatively small size of the higher education sector, the Dutch have elected to pay more attention to the specific characteristics of disciplines, research programmes and institutions. It has been acknowledged from a humanities and social science perspective that uniform use of quantitative publication indicators is problematic. There has also been more attention paid to the international context for purposes of comparison.

The use of quantitative indicators such as bibliometrics to measure scientific productivity has been more notable in recent years, and there have been indications that this will be extended to AH/SS disciplines, subject to the "development of new tools".

Finland

There is no comprehensive ex-post evaluation of university outputs in the Finnish higher education system that is linked to funding allocations. Every 3 years, the universities negotiate their block grant with the Ministry of Education, and a small proportion of this (3%) is performance related. The performance is measured through agreed indicators such as (i) number of international centres of excellence (ii) Finnish Academy (AKA) funding (iii) international collaboration and funding (e.g. EU grants) (iv) graduate placements and (v) success of the universities in achieving their stated strategic aims. All this information is available through a national database (KOTA) which the universities are responsible for updating. The database also contains information on publication patterns (what is published and where it is published), and PhD completion rates.

There is a formula used for the remainder of the block grant that includes target numbers of masters and doctoral degrees. In a 2006 document entitled 'Government Resolution of the Structural Development of the Public Research System', there is a commitment to increasing the performance-related share of the block grant. No precise figure is indicated, but a 1997 Ministry of Education report ('Management by Results') suggested the figure should be as high as 35%.

The Finland Higher Education Evaluation Council (FINHEEC) also carries out broad formative institutional evaluation, including evaluation and validation of courses and teaching quality, but not for the purposes of funding or ranking. The method consists of peer review of a university self-evaluation.

The Academy of Finland (AKA) carries out evaluations of research fields at the level of programme or project group to assess Finland's international standing, but this does not affect budget allocations. Once the research field has been delimited and the research groups identified, the following criteria are applied to evaluate them:

1. The mission, vision and goals;
2. The supply of resources and the efficiency of their use;
3. The scientific competence and the degree of innovation;
4. The technological competence and the co-operative activities with other research groups, industry and users of research results;
5. The national and international importance of the research group and of their research results for the scientific community and for the further qualification of researchers;
6. The relevance of the research group and their research results for industry.

The methods used are self-evaluation by questionnaire, peer review of the questionnaires and site visits, all of which lead to final report. Research groups are offered the opportunity to comment on the report.

AKA also provides incremental funding for international centres of excellence. In selecting these, the following main criteria are used:

1. The national and international position of researchers;
2. The scientific significance, innovativeness and effectiveness of research;
3. The quality and quantity of scientific production and where published (especially work published in internationally respected scientific series);
4. Societal relevance and effectiveness of the research (incl. patents);
5. The national and international mobility of researchers;
6. Systematic international cooperation of the unit (incl. cooperation with business companies);

7. The success in training researchers (incl. numbers of graduates and supervisors)

The criteria are adjusted to accommodate the differences between disciplines. There are 8 centres of excellence in the humanities supported by the Research Council for Culture and Society (humanities section of AKA).

Poland

Since 1998, statutory research funding (as distinct from operational funding based on numbers of students) has been allocated using a parametric method based entirely on quantitative methods. It consists of the sum of the points received for **performance R(P)** and for so-called **general results R(G)** divided by the number of staff, giving an indicator of effectiveness (E).

R (P) consists of 6 indicators:

1. the number of publications in refereed journals;
2. publication of books (monographs);
3. scientific degrees awarded to academic personnel in the unit;
4. number of patents;
5. implementation of research results; and
6. a right (licence) to carry out quality evaluation or accreditation of national laboratories.

The following indicators are taken into account when calculating R (G):

- various research projects (grants);
- research commissioned at the unit;
- research projects financed from abroad;
- international co-operation agreements;
- numbers of long-term scientific visitors from abroad;
- numbers of citations;
- awards for scientific or practical achievements;
- expert reports commissioned from the unit;

- the right to award academic degrees by the unit;
- dissemination of knowledge among the lay people (e.g. presentations in popular journals);
- existence of doctoral studies organised in the unit;
- organisation of international and national conferences;

Australia

Core funding for Australian universities was distributed until 2000 using a funding formula entitled the 'Relative Funding Model'. Funding streams for teaching and research were distinguished, and the research component allocated on the basis of the 'Research Quantum'. Initially based on universities' success in gaining competitive research grant, the formula was made increasingly more complex. The Composite Index was introduced in 1995 and is composed of *research input measures* and *research output measures,* viz:

1. Research input measures (funding):

- the amount of each university's funding from Commonwealth competitive grants;
- other public sector research funding;
- industry and other research funding.

2. Research output measures:

- numbers of research and scholarly publications produced by staff and students;
- numbers of higher degrees completed (Masters and PhD).

The weightings for these changed from year to year as a more refined balance was sought. In 1999 the weightings were as follows:

Funding Source Weighting

Category 1 - National Competitive Research Grants	
Source	**Weight**
Commonwealth schemes (including a share of DISR funding to Co-operative Research Centres)	2
Non-Commonwealth Schemes	2

Category 2 - Other Public Sector Research Funding	
Source	**Weight**
Local Government (competitive and non-competitive)	1
State Government (competitive and non-competitive)	1
Commonwealth Government (other than those listed above)	1

Category 3 - Industry and Other Research Funding	
Source	**Weight**
Australian contracts	1
Australian grants	1
Australian donations, bequests, and foundations	1
Australian syndicated research development	1
International Funding	1

Publication Category* Weighting

Publication type	Weight
Authored book - research	5
Book chapters	1
Article in scholarly journal	1
Conference publication – full written paper, refereed proceedings	1

*****The value of joint publication is shared equally between the authors

Degree Completion Category Weighting

Degree level	Weight
Doctoral degree by research	3
Master degree by research	1

Furthermore, each of these categories - grants, publications, degree completion - is weighted in a ratio of 8:1:1 respectively. Each university's share of each of the categories is calculated, and then the appropriate weightings are applied to determine the universities share of overall funds available.

The process relied on accurate data being submitted by the universities (signed off by the head of the institution) and procedures being established to handle this.

The Research Quantum has been praised for (i) rewarding institutions with a strong research performance and (ii) avoiding the transaction costs associated with other selective institutional grants.

However, it has also been criticized for the following reasons:

(i) Research grant income is not a measure of performance, but of research input.
(ii) Research grant income as a performance measure has a low validity, and is conflated with a number of other factors unrelated to research performance
(iii) Monetary value of grant income is a poor measure as the cost of a project is not correlated with its scientific merit. Number of research grants may be a better measure
(iv) The Quantum captures the volume of research undertaken, but does not capture quality.

In 2000, a new framework allowed for the distribution of block funding via the Institutional Grants Scheme (IGS) and the Research Training Scheme (RTS). The IGS absorbed funding previously distributed under the Research Quantum and the Australian Research Council (ARC) Small Grants scheme. The formula for

funding allocation depended on success in attracting a range of research grants (60%), success in attracting research students (30%) and the quality and output of their research publications, assessed through a revised publications measure (10%).

The weightings of the revised publications measures are as follows:

Publication type	**Weight**
A scholarly book produced by a commercial publisher	5
A chapter in a scholarly book produced by an international publisher	1
An article in a scholarly refereed journal	1
A peer reviewed paper presented at a conference of national or international significance and published in its proceedings	1

Four weakness have been identified with the publications component

(i) The publications measure rewards quantity rather than quality

(ii) This has lead to unintended changes in behaviour. Studies by Butler (2003; 2004) have found a relationship between the introduction of performance-based block funding and a sharp rise in journal publications in lower impact journals

(iii) The publications element is highly correlated with the other elements of the IGS formula – in particular grant income - and therefore adds little value to the assessment process.

(iv) Certain important categories of publication and research output were omitted, disadvantaging the creative arts and design in particular.

The result of these reforms was that an increasing share of research funding was distributed via funding formulae and this attracted much criticism from the academic community.

Belgium (Flanders)

Until 2003, all university block funding was distributed in Flanders on the basis of student numbers. Since 2003, 50% of funding is distributed on this basis, but the remainder is funded on the basis of bibliometric analysis of the outputs of the 6 universities using Thomson-ISI data. This analysis is carried out by Steunpunt O&O Statistieken (SOO), an agency established specifically for this purpose.

In order to use the data for the purpose of allocating funding, an enormous amount of 'data-cleaning' needed to be undertaken (misspellings, different listings of individuals and affiliations). This was feasible for a small higher education sector such as Flanders, but would be very difficult for a much larger exercise.

Because of the limitations of Thomson's Social Science Citations Index (SSCI) and Arts and Humanities Citations Index (AHCI) bibliometrics are not used for the allocation of funds to these agencies.

Slovenia

The Slovenian Research Agency (ARRS) is the primary organization in Slovenia for the distribution of research funding and evaluation of research. Disciplines are evaluated every five years, research institutes and researchers are evaluated annually, and quantitative indicators are used to assess researchers' suitability to be project leaders.

For the 5-year evaluation of disciplines and sub-disciplines, qualitative methods are used: questionnaire surveys, interviews, site visits, case studies. Experts from the academic sector are also involved to prepare this kind of evaluation. These qualitative methods are used in conjunction with the indicators defined by the Government Regulations Act for the evaluation and financing of research, and which are an integral element of the ranking of research institutes. The Regulation Act defines the following indicators for researchers:

- Indicators of researchers efficiency (for precise weightings see **appendix 1**)
- Citation and research achievements;
- Involvement in EU and other international research programmes and projects;
- R&D co-operation with other research and private and public organizations.

The indicators of research efficiency are divided into three categories as follows:

I. INDICATORS OF RESEARCH EFFICIENCY

1. Scientific Articles indexed in SCI Expanded:
First quarter of journals: 80 points
Second quarter of journals: 60 points
Third quarter of journals: 40 points
Fourth quarter of journals: 20 points

2. Scientific Articles indexed in SSCI:
Above median of corresponding sci. journals: 80 points
Below median of corresponding sci. journals: 40 points

3. Scientific Articles that is indexed in A&HCI: 20 points

4. Scientific Articles that is not indexed in ISI, but it is indexed in international bibliographic data base: 10 points

5. Scientific Articles published in Slovenian research journals: 5 points

6. Short scientific contributions are evaluated with 80% of what sci. articles are getting in corresponding scientific journals.

7. Book published at international scientific published: 100 points

8. Book published at domestic scientific publisher: 50 points

9. Book published at other publishers: 30 points

II. INDICATORS OF DEVELOPMENTAL EFFICIENCY

10. Transfer of knowledge into economy and social sphere

11. Integration of research in university study programmes

12. Publishing of faculty handbooks

13. Patents or selling of patent rights

14. Research and developmental work in support for development of data bases, indicators, dictionaries, glossaries, lexicons etc

15. Development of systematic, normative, programmatic, methodological and organizational solutions, including evaluations, reviews, expert report

16. Published expert works

III. INDICATORS OF MANAGEMENT EFFICIENCY

17. Efficiency and success in previous periods

18. Comparison of research goals with available infrastructural capacities

19. Integration of researchers with domestic business sector and local social networks

20. International integration of researchers

The United Kingdom (UK)

The RAE operates through a process of peer review by experts of high standing covering all subjects. Judgements are made using the professional skills, expertise and experience of the experts; it is not a mechanistic process. All research assessed is allocated to one of 68 'units of assessment' (UoA) which are discipline-based.

For each unit of assessment there is a panel of between nine and 18 experts, mostly from the academic community but with some industrial or commercial members as well.

Every higher education institution in the UK may make a submission to as many of the units of assessment as they choose. Such submissions consist of information about the academic unit being assessed, with details of up to four publications and other research outputs for each member of research-active staff. The assessment panels award a rating on a scale of 1 to 5*, according to how much of the work is judged to reach national or international levels of excellence (see table below).

Units of Assessment

There are 67 units of assessment in the 2008 RAE. Each unit covers a broad subject area. For example, Mechanical, Aeronautical and Manufacturing Engineering are included within one unit; Drama, Dance and Performing Arts are all included in another. The units of assessment have been identified in consultation with the higher education sector and continue to evolve to reflect changes in the pattern of research in institutions.

Assessment Panels

There is a two-tier panel system: 67 sub-panels of experts, one for each UOA, work under the guidance of 15 main panels. Under each main panel are broadly cognate disciplines whose subjects have similar approaches to research. The panel chairs were nominated by members of the 2001 RAE panels and appointed jointly by the four funding bodies. Panel members are nominated by a wide range of organisations, including research associations, learned societies, professional bodies and those representing industrial, business and other users of research. Panel members are then selected by the

funding bodies, on the advice of the panel chair, based on their research experience and standing in the research community, so as to ensure coverage of the subject concerned. The funding bodies also seek to reflect the profile of nominations received in terms of geographical coverage, gender, and type of institution. The chair and members of each panel participate as individuals, rather than as representatives of a particular group or interest. The names of the panel chairs and members are published.

Nearly half of the panels have established subpanels; these often include people who are not members of the main panel. The sub-panels advise on assessment of research in particular sub-areas within the subject. Panels may also draw on the advice of specialists covering specific areas of expertise outside the panel's experience. In addition, all panels consult with advisers based outside the UK to confirm their application of the standard of international excellence which is the benchmark for the exercise.

What Information is Provided by Universities and Colleges?

Each publicly funded university and higher education college in the UK is invited to submit information about their research activity for assessment. The information they supply provides the basis on which judgements are made. Submissions have to be in a standard format, which includes qualitative and quantitative information. Most of the information is provided electronically on specially written software.

The submissions are based around members of staff in each academic unit in which the institution is submitting. It is up to each institution to decide which subjects (and therefore which units of assessment) to submit to, and which members of staff to include in each submission. For each member of research staff, up to four items of research output may be listed. All forms of research output (books, papers, journals, recordings, performances) are treated equally; panels are concerned only with the quality of the research. Similarly, all research (whether applied, basic or strategic) is treated equally. In addition, the HEI must provide information in a number of different categories shown below:

Category Description

Staff information

• summaries of all academic staff

• details of research-active staff

• research support staff and research assistants

Research output

• up to four items of research output for each researcher

Textual description

• information about the research environment, structure and policies

• strategies for research development

• qualitative information on research performance and measures of esteem

Related data

• amounts and sources of research funding

• numbers of research students

• number and sources of research studentships

• numbers of research degrees awarded

• indicators of peer esteem

How do the Panels Make their Judgements?

The panels use their professional judgement to form a view of the overall quality of the research in each submission within their unit of assessment, using all the evidence presented in the submission.

To assess submissions fairly and consistently within each UoA, each panel draws up a statement describing its working methods and assessment criteria. These are published in advance of submissions being made. This statement shows which aspects of the submission the panel regards as most important, and areas that it wants institutions to comment on in their submissions. The differences in working methods and criteria between panels are a reflection of the need to recognise differences in the way research is conducted and published in the various disciplines.

Panels review all submissions, and read selectively from the research outputs cited. Because the panels are concerned with quality, not quantity, information on the total number of publications produced is not requested. Panels do not visit institutions as part of their work.

What are the Ratings?

The subject panels use a standard scale to award a rating for each submission. Ratings range from 1 to 5*, according to how much of the work is judged to reach national or international levels of excellence. The table below shows the definition of each rating in the 2008 exercise.

Quality Level	Description
4 star	Quality that is world-leading in terms of originality, significance and rigour
3 star	Quality that is internationally excellent in terms of originality, significance and rigour but which nonetheless falls short of the highest standards of excellence
2 star	Quality that is recognised internationally in terms of originality, significance and rigour
1 star	Quality that is recognised nationally in terms of originality, significance and rigour
unclassified	Quality that falls below the standard of nationally recognised work. Or work which does not meet the published definition of research for the purposes of this assessment.

The five quality levels from 4* to Unclassified apply to all UOAs. Some panel criteria statements include a descriptive account of the quality level definitions, to inform their subject communities on how they will apply each level in judging quality. These descriptive accounts should be read alongside, but do not replace, the standard definitions. The attached MS Excel spreadsheet sets out

the additional subject specific explanations added to the standard quality level definitions in use for RAE2008.

4. BIBLIOMETRICS AND CITATIONS IN THE HUMANITIES: AN OVERVIEW

4.1 Introduction

The bibliometric approach is now widely accepted as useful method to measure some key aspects of research performance and is increasingly used by government ministries as a method of providing cross-country comparisons of research performance[2]. However, the consensus amongst bibliometricians[3] and other experts is that bibliometric data only provides a valid impression of scientific performance where research publication and citations sufficiently reflect the state and dynamics of research in the corresponding research areas. This cannot be said to be the case in the humanities for reasons that are given in full below. At the current time, the accepted methods developed for the use of bibliometrics in evaluations of research performance in the natural sciences cannot be applied with confidence to the field of the humanities.

4.2 The use of bibliometrics in the arts and humanities

The following discussion concentrates on citation analyses as the most common bibliometric technique used in the evaluation of research. The data for citation analyses are obtained by counting the citations achieved by a researcher or a department. This data is typically aggregated and analysed at the level of HEI or discipline in order to assess the impact or quality of research output (Analyses at lower levels of aggregation are generally considered to be unreliable indicators of impact or quality). The database most commonly used is the Science Citation Index (SCI) of the Institute of Scientific Information (ISI), a US-based commercial operation. Other commonly used citation databases include Elsevier's Scopus, the publicly available Citeseer, and Google Scholar (a freely-accessible web search engine that indexes the full-text of scholarly literature across an array of publishing formats and disciplines).

It should be noted before discussing the limitations of citations that there a central contention (widespread amongst bibliometricians) that citations do not

[2] The *PSA Target Metrics for the UK Research Base* prepared by Evidence Ltd uses the Thomson ISI database to compare UK research performance in 5 broad fields (including the humanities). http://www.dti.gov.uk/files/file27330.pdf

[3] See for example *Use of bibliometrics in evaluations of social science research: a review*, Nederhof and Tjissen, and *The Use of bibliometrics in the Social Sciences and Humanities,* Science-Metrix Report.

measure the inherent quality of publications at all, that they merely demonstrate the impact, influence or, at best, significance of a piece of work[4]. There is a widespread and informed view that citation or other bibliometric analyses should only be used as part of a set of indicators intended to measure quality, or as a way of informing or challenging peer-reviewed assessment. However, many policy-makers and researchers recognize citation analysis as an objective indicator of research quality for many fields. Furthermore, citations have certain advantages when evaluating research practices:

- They contribute to the objectivity and transparency of the research process
- They provide a 'bigger picture', revealing macro-patterns in the communication process that cannot be seen from the perspective of the individual researcher.

There are a number of features of citations which makes them problematic more generally: these can be divided into generic difficulties with all publication measures (which in turn are reflected in citation analyses), and more specific issues relating to citations. Thirdly, there are a number of specific objections to their use in measuring the output of humanities research. These are listed as (a) generic publication count issues (b) generic citations issues and (c) humanities-specific issues in turn:

(a) Generic publication count issues:

1) The most fundamental objection to simply counting the number of publications to assess research performance is that this is merely reflects (at best) the productivity a researcher or research group, and has no bearing on the quality of the research produced.

2) The mobility of staff may alter in a significant way the output of a department, consequently different ways of ascribing the output of a researcher to a department or other research grouping– *e.g.* to the one where he or she was based or to the current one – may have an important impact on the output indicator

[4] *Setting the scene: a review of current Australian and international practice in measuring the quality and impact of publicly-funded HASS research*, Donovan, C., (Research School of Social Sciences , ANU, 2005) – p.17.

3) Determination of the number of staff in a department depends on who is classed as a research member of the department. Different accounting procedures for post-doctoral students, Ph.D. students, visiting staff, etc. result in significant variations in the *per capita* figures

4) Particularly in medicine and the natural sciences it is common practice to have a large number of co-authors, hence the publications can be counted either on a whole or on a fractional basis, giving rise to different output indicators.

5) The use of publication counts as an indicator of research performance is strongly limited by the fact that the variations in the level of resources (inputs) explains much of the observed difference in publication activity across departments

6) Biases favouring the publication of established authors may exist in the publishing process, distorting the significance of the indicator.

(b) Generic citations issues:

1) The SCI tends to have a bias in favour of publications in the English language and especially towards North American sources.

2) The SCI reports only the first author; moreover it is not uncommon to find programming errors both in the author's name and in the journal citation

3) Citations are made not only to works considered to be of high quality, they can also be used in a negative or derogatory way, but citation counts cannot distinguish between the two. However, citation impact studies at the macro-level of entire countries and research fields will tend to smooth out such differences, yielding citation frequency data and derivative measures that are amenable for cross-country comparisons of impact and visibility.

4) Different citation windows (how many years are considered after the publication) may give rise to variations in the indicator measurement.

5) Self-citation, citation to co-authored papers, citations to different journals, all require the development of weighting schemes that at present cannot be applied in an objective way

6) Seminal or radical works may be difficult to understand or, after their acceptance, become common knowledge, and then may not receive the number of citations that they deserve.

7) Citation counts can be distorted by the inappropriate use of the citations such as in the case of a citation circle (researchers unduly citing each others' work) or citations for more personal reasons (e.g. junior staff citing senior researchers). The usual response to these concerns is threefold: if data is aggregated at a high enough level, then this won't be a problem; that this behaviour would be randomly distributed across the sample when aggregated highly enough, so it is not statistically significant; and that this behaviour is a part of academic life, and that because bibliometrics detect it, it is not a weakness of bibliometrics.

8) In all research fields, review articles are much more frequently cited than research articles. Impact factors, therefore, tend to overemphasise journals that give more attention to review articles than research journals.

(c) Humanities-specific issues

1) *Poor coverage by citations indices:* The SCI has a much better coverage of the natural sciences (particularly medicine) than of social science, and arts and humanities journals (Social Sciences Citations Index (SSCI) and Arts and Humanities Citation Index (AHCI) respectively), where commercial demand for their services is much weaker.

2) *National or regional orientation:* Much humanities research is by its nature concerned with a wide variety of specific cultural phenomena.

The most appropriate place to publish research outcomes may be in non-Anglophone journals which are based outside the US/UK. AHCI coverage of such journals is very poor. Furthermore, research that has a strongly regional or national orientation (for example, Law) is less likely to be cited by the major US or UK journals covered in the AHCI database.

3) *Wider audience:* To a much greater degree than is the case in science disciplines, humanities authors will also target the general public, through the medium of books and book-chapters. There is also a strong tradition of writing for the non-scholarly press.

4) *Different publication patterns and characteristics:* It is claimed that the typical citation window (2-3 years) is too short for the humanities. Much excellent work may take time to be recognized. The humanities differ from the natural sciences in such publication characteristics as a larger half-life of publications and a higher citation rate of older literature. In a similar vein, the life-span of influential work in the arts and humanities is thought to be longer than in other disciplines. For example one study (Glanzel and Schoepflin) calculated that the mean reference age for the history and philosophy of science was 39 years (compared to 7-8 years in the biomedical sciences)[5]. A study in the field of psychology by the same authors found that, over a 14 year period, articles took 8 years to reach 50% of their citations compared to 4.5-6.5 for physics articles. Moreover, while up to three-quarters of physics articles were estimated not to receive any citations after 14 years, this fraction was less than a quarter for psychology articles. It should be noted, however, that these variations may not be uniform across all arts and humanities disciplines; although as psychology is a discipline that in many ways most resembles the patterns and publication characteristics of the natural sciences, the picture in other humanities disciplines is likely to be if anything even more anomalous.

5) *Different publication channels:* The primary mode of communication between researchers in many humanities fields is not primarily through journal articles, but through book chapters and monographs. For

[5] *A bibliometric study of reference literature in the sciences and social sciences,* Glanzel, W. and Schoepflin, U.

example, Dr Henry Small (ISI) has calculated that 61% of references in the field of the history and philosophy of science in a selected data-set were to non-journal publications[6]. Compare that to a study by Small and Crane[7] which found that only 1% of cited items in high-energy physics referred to books. Similarly, there may be a significant body of scholarly literature which is accessed by researchers in the course of their work in the form of grey literature or other non-standard publications that is neglected by journal citation databases.

The dominance of books and book-chapters as the primary publication channel also has an effect of distorting the citation counts. It has been shown that citation peaks for non-journal material tend to occur relatively late[8]. Therefore a five year citation window would be a minimum requirement for meaningful analysis.

6) *Non-text-based outputs*: (1) - (5) above addresses issues related to text-based outputs in the humanities broadly conceived. A significant proportion of research outputs in this domain are not text-based – e.g. musical compositions or exhibitions - and, obviously, it would not be possible for the impact or quality of these outputs to be detected through standard bibliometric measures.

7) *Differences in academic culture:* It might be claimed that criticisms (1) – (5) are also *in principle* remediable if the deficiencies of existing databases were rectified either through the creation of a bespoke database or the extension of existing ones, and modifying the other parameters (e.g. citation windows) of your chosen method of measuring impact. However, there may be residual *cultural* differences in academic practice which would distort any attempt to benchmark quality in the humanities through bibliometric methods. For example, there is a lot of anecdotal evidence that humanities scholars in France do not cite as extensively as British or Scandinavian colleagues for example. To our knowledge, there has been no study which compares

[6] Paper presented to Royal Academy for Sciences and Arts in Brussels, 26 January 2005.
[7] Small, H.G. and Crane, D., *Specialties and disciplines in science and social science*
[8] *A bibliometric analysis of six economics research groups: a comparison with peer review*, Nederhof et al

such national or cultural citation practices, and to what extent this might be statistically significant. However, it should be borne in mind that the problems in comparing or benchmarking research outputs may not have technical solutions.

4.3 A bibliometrics fit for the humanities? Some developments

The reasons set out above provide a powerful set of arguments for not using standard bibliometric tools to assess research performance in humanities fields. However, it might be argued that these objections to the use of bibliometrics in the humanities reflect the current inadequacies of citations indices, bibliographical databases and measurement techniques rather than differences in scholarly practice between the natural sciences and humanities. In the main – apart from areas of practitioner- or practice-led research – humanities researchers publish outputs which cite and acknowledge the work of others and – if captured accurately in the range of publication channels used by humanities scholars - this would provide one measure of the impact of that work. This section looks at some of the ways that bibliometric tools can be designed or improved to provide a more comprehensive view of scholarly impact in the humanities.

4.3.1 Overcoming the data deficit in the humanities: mining the Web of Science

In their paper *Extending citation analysis to non-source items*, Linda Butler and Martijn Visser of ANU's Research Evaluation and Policy Project (REPP) examine the possibility of mining standard citation indices such as the Web of Science for references to publications outside of the indexed journal literature.

The paper reports the first results of the extension of citations analysis to 'non-source' items and their use in the assessment of university departments. 'Non-source' items are defined as those publications not indexed in the Thomson ISI database, but are visible within the database nonetheless. That is, the journals covered by the database not only cite other journals, but also books and book chapters, conference proceedings and other scholarly literature. Both source and 'non-source' citations were then mapped from the Thomson database to lists of publications provided by all Australian universities, and university departments were ranked according to the resulting citation per publication (cpp). Six fields (disciplines) were analysed in detail including History and Law. It was found that in some fields, notably History, the inclusion of non-source citations significantly

changed the ranking of the university departments as compared to the ranking based on source citations. In Law, however, the rankings were largely unaffected by the inclusion of this new material.

The extension of any analysis to non-source items inevitably results in more publications and citations being identified, but the effects are not uniform across fields. For example, the number of publications identified in the field of languages increased by 700% and in History by over 500%. The figure for the increase in citations is very roughly proportional in these fields, approximately 700% and 200% respectively. However, there is a broad group of subjects (including Law, Arts and Architecture) where the citation counts at least double, but publications increase by a much larger factor, which has the effect of depressing citations per publication. A third group (including Philosophy and Communication Studies) sees a large increase in publications, but a relatively modest increase in citations.

The study confirms the traditional criticism of the coverage of the Thomson ISI indices, with the number of citations to non-source items in the humanities outnumbering the citations to source items.

There are a number of caveats which should be noted in connection with this study:

- The study mapped citations from journals to other non-source material, but that still ignores other channels of communication and citation, for example citations from books to books. As the study puts it: “we are still bound by the conversation that takes place within the ISI world”
- It nearly all cases the changes in rankings affected those institutions at the lower end of the scale.
- The project proved that this mining of the Thomson database for non-source data is a feasible exercise for the ranking of a small number of university departments in Australia (based on DEST publication returns that Australian universities submit annually). The feasibility of the process for a higher education sector on a European scale, where universities do not currently return information on all publications is untested. A more restricted project which looked at the publications which resulted from European Research Councils or Funding Agency projects would still need to overcome issues concerning (i) the lack of

comprehensive publication data (ii) scale and (iii) comparability, due to doubts whether outputs from Research Council funding are representative of national quality as a whole.

4.3.2 Hirsch's h-index

Physicist Jorge Hirsch has proposed an easily computable single-figure index, *h*, which gives an estimate of the importance, significance and broad impact of a researcher's cumulative research contribution[9]. The definition of *h* is:

As researcher has index h if h of his/her papers published over n years (N_p) have at least h citations each, and the other (N_p – h) papers have no more than h citations each.

For example, if Professor T has an h-index of 15, then Professor T has written 15 papers with at least 15 citations each.

The h-index has a number of advantages over other single-figure criteria used to judge scientific impact, viz:

- It provides a measure of impact rather than just productivity (unlike publication counts)
- It avoids distortions due to a papers having a few 'big hits' or the effects of highly-cited review articles (unlike total or average citation counts)
- It recognizes and rewards the more productive researcher (unlike citations per paper, which may reward low productivity)
- It is less arbitrary than measures such as 'number of significant papers' (Papers with more than x citations), which needs to be adjusted for different disciplines and levels of seniority.
- It is a single figure that allows easy comparability (unlike number of citations to each of the q most cited papers)
- It provides a measure of a researcher's impact over a lifetime, so that citations continue to be recorded outside of the usual citation-windows. This is a particular advantage in the arts and humanities, where older, seminal works can still be heavily cited.

[9] *An index to quantify an individual's scientific research output,* JE Hirsch - http://xxx.arxiv.org/PS_cache/physics/pdf/0508/0508025.pdf

- Self-citation can be dealt with relatively easily. All papers with citations just above *h* are scanned for self-citation. If a paper with *h+n* citations has more than *n* self-citations it is dropped from the h-count, and *h* will drop by one.

The source Hirsch uses for his calculations is the ISI Web of Science, but free-access databases such as Google Scholar can also be used (with caution, as citations can sometimes be dramatically lower than in the ISI databases).

One of the interesting features of the h-index is that two individuals may have the same number of publications and total citations but one may have a much higher h-index than the other. It is argued that the researcher with the higher index is the more accomplished researcher.

The h-index can also be applied to groups of researchers as well as individuals. The *h* for a research group is not simply the sum of, nor is it proportional to, the h-indices for the individuals in that group.

4.3.3 Applying the h-index to Arts and Humanities research

The h-index does not avoid some of the common criticisms of the use of citations in the arts and humanities. Its reliance on databases such as Web of Science and Google Scholar means that issues such as disciplinary coverage and the focus on journal publications arise here also.

The h-index therefore would not assuage the concern that, as far as comparisons between disciplines are concerned, the dice are loaded in favour of the natural sciences. However, the h-index does have an in-built disciplinary sensitivity: there will be differences in typical h-values in different fields, determined in part by the average number of references in a paper in the field, the average number of papers produced by each researchers in the field, and the size of the field. The index could be used as part of an exercise in setting disciplinary benchmarks in citations within the domain of the arts and humanities (as part of a larger set of metrics), given the advantages set out above. This would have the benefit that the distortions or underreporting of citations that result from the known biases of the current indices would apply across the whole discipline (on average). There may, however, be exceptions to this: for example, archaeological and linguistics research may not be uniform in its preference for publishing in book or journal

format. In such cases an h-index based exclusively on journal publication will systematically underestimate the performance of some archeological researchers.

There are additional caveats which apply to the use of the h-index which may have particular ramifications for the arts and humanities:

Volume bias: although the h-index is not simply a measure of productivity, the fact that the upper bound of the index is determined by the total number of publications means that it militates against those whose careers have just started, or those whose careers have been truncated, or those who publish few outputs.

As seen above, it is considered one of the merits of the index that it does not reward low productivity. However in subjects such as history, a low number of outputs should not be equated with low productivity. One large monograph can be the result of a number of years' intense scholarship. The index works well in disciplines where a consistent flow of publications of roughly equal weight is the normal model of communication of research results, but less so in disciplines where outputs are accorded varying weights.

The more general point here is that although the h-index is a reliable indicator of high accomplishment, the converse is not true. Much high-quality, innovative work which is as yet unacknowledged by large sections of the academic community will fail to be rewarded by such a measure.

Nascent or minority interest research: researchers working in non-mainstream areas will not achieve the same very high h-values as the top echelon of those working in highly topical areas. Arguably, this issue is accentuated in the arts and humanities by the very nature of the way research areas evolve.

Collaborative papers: Disciplines with typically large collaborations will typically exhibit large h-values. There are two responses to this. One might argue that in concentrating on disciplinary benchmarks rather than comparing across disciplines, these differences would have a negligible influence on output. However, as mentioned above, there may be considerable intra-disciplinary variation. A researcher with a high *h* achieved mostly through collaboration with other authors would be favoured by the operation of the h-index. Secondly, in cases where there are significant differences in the numbers of co-authors, it

would be useful for the purpose of comparison to normalize *h* by a factor that reflects the average number of co-authors.

4.3.4 Assessing research output according to journal and book weights

This approach entails the development of sets of journal and monograph weights. It is particularly valuable when publication in ISI source journals is low. Lists of journals and publishers are offered to national and international experts to rank, and statistical weightings are calculated based on the rankings. These weights are them applied to research outputs in a subsequent evaluation. This addresses the problem of outputs targeted at a non-scholarly public mentioned above, as these can be accommodated within the system. The system is a more sophisticated version of evaluation via publication count (see criticisms above), but note that it does not measure impact of the publications through citations.

A version of this bibliometric method has been used recently in relation to performance-based budgeting for Norwegian higher education institutions. A model was commissioned by the Norwegian Ministry of Education and Research in 2002 and developed by the Norwegian Association of Higher Education Institutions in 2003-2005.

The model covers about 8,000 scientific and scholarly publications per year in all types of research (from art to astrophysics) at several types of institutions (from traditional universities to specialized or regional university colleges). The publication activity is reported by the institutions in a common documentation system as ordinary bibliographic references. But unlike normal publication lists in CV's or annual reports, the bibliographic references in the documentation system are standardized and analyzable by *publication channel* and *type of publication*, just as in professional bibliometric data sources. *Co-authored publications* can be identified and shared among the participating institutions; they are not double-counted.

A dynamic *authority record* of 16,000 controlled scientific and scholarly publication channels ensures that references to non-scientific publications are not entered into the system. *Publication channels* are defined as ISSN-*titles* (journals, e-journals and series) or *publishers* of ISBN-titles. It is required that all publication channels in the authority record make use of external peer review. They must also publish on a minimum national level, which means that not more

than two thirds of the authors that publish in the channel can be from the same institution.

Publication data from professional bibliographic data sources are *imported* to the documentation system in order to facilitate the registration of publications by the employees. To achieve this, the Norwegian Government has made a special agreement with Thomson ISI and the National Library of Norway. The latter has an indexing service for Norwegian and Scandinavian scientific and scholarly journals which covers most journal articles that are not indexed by ISI. The two data sources together cover 90 per cent of the reported journal articles. But the documentation is not limited to these data sources and to journal articles. It is important, especially from the point of view of the humanities, the social sciences and technology, that all publications that can be defined as scientific and scholarly may be entered into the system and counted. The new system records all publications in a controlled and bibliometrically analyzable manner.

The documentation system provides the delimitation and structuring of the publication data. In the measurement for the funding formula by the end of each year, the publications are *weighted* as they are counted. In one dimension, three main publication types are given different weights: *articles in ISSN-titles, articles in books (ISBN)* and *books (ISBN)*. In another dimension, *publication channels are divided into two levels* in order to avoid an incentive to productivity only. The highest level giving extra weight includes only the leading and most selective international journals, series and publishers that account for about 20 per cent of the world's publications. The national councils in each discipline or field of research participate annually in determining and revising the highest level under guidance of the Norwegian Association of Higher Education Institutions. The weighting of publications by type and channel is shown in *table 1*.

Table 1. The weighting of publications

	Channels at normal level	Channels at high level
Articles in ISSN-titles	1	3
Articles in ISBN-titles	0,7	1
Books (ISBN-titles)	5	8

These weights are given to the publications, not to the authors. Co-authored publications are *fractionalized* among the participating institutions in the measurement.

Citation counts (or other measurements of impact that are developing with the world-wide web) are not included in the model because of the short time span (last year's performance counts in next year's budget) and the heterogeneity of disciplines that the model must cover.

Several institutions have adopted the national model on a local level for their internal budgeting of faculties and departments. This is one of several signs of general acceptance of the model, but it has not been uncontroversial. Most of the ongoing debate is about the division of publication channels in a normal and a higher level. Since national publication channels (authors mainly from the same country) cannot be appointed to the higher level, some of the researchers who most frequently use Norwegian channels argue that the model is threatening Norwegian as a scientific and scholarly language. This view is supported by Norway's national publishing industry, but not by the statistics from the new documentation system. Almost one third of Norway's total scholarly and scientific publication output is in the Norwegian language, and this share has not decreased after the model was implemented. From 2004 to 2005, there has been increasing publishing activity on both levels of publication channels.

A more general and just as important result of the implementation is the new focus on scientific and scholarly publishing that now engages all researchers in all types of institutions and areas of research. An interesting interdisciplinary debate on scientific and scholarly standards and publishing traditions has arisen.

4.3.5 Weighting model and ERIH

The Norwegian model is intended not only to monitor the quality of the national system's research output, but also to drive the research strategies of institutions in a particular direction. The resulting prioritization of publication channels and types of publication will therefore involve controversial normative judgements.

These controversies are inevitable, even where such prioritization or categorization is not a matter of evaluation and does not result in funding allocations. This much is evident from the development European Reference Index for the Humanities (ERIH) project which is jointly sponsored by the European Science Foundation (ESF) and HERA. In the absence of an agreed and easily measurable criterion such as a journal's impact factor, assessing whether a

journal is “high-ranking” or enjoys a “strong reputation” will always be a matter of consensus and peer review.

Nevertheless, over a period of 2 years substantial agreement has been reached on the journal lists that ERIH in the majority of disciplinary fields, and it is expected that broad agreement on the categorization and content of the lists for all 15 fields will be achieved this year (2007). The main lessons from success of this project have been that consensus requires:

- The composition of the expert groups and steering committees needs to have widespread credibility;

- The need for stability, continuity and transparency in the peer review process;

- Extensive consultation with the academic communities throughout the process;

- Building in mechanisms for the review of the lists at regular intervals to reflect a dynamic research landscape.

There has been some criticism of the Norwegian model from the point of the view of the humanities. A national or regional focus *per se* in the humanities does not mean a reduction in quality. This and the anxieties concerning the effective down-grading of research published in the Norwegian language may combine to disfavour the humanities disproportionately. However, it is possible that the biases of the system would be overcome by using weighting which are based on ERIH categorizatons, or weightings that were achieved through a similar peer review process.

It is legitimate to question whether the hard-won consensus which has been achieved in the course of compiling the ERIH lists – which after all are intended to be a reference tool - could be sustained if the lists were then to be used for a very different purpose i.e. that of evaluating and benchmarking the outputs of humanities research. It is more likely that a new process – with new criteria and guidelines for expert groups – would need to be initiated.

The Norwegian model presented here has a number of other advantages. The model is deliberately not very sophisticated from a bibliometric point of view: It can be understood by all employees in the higher education sector. Still, the *common documentation system* for all institutions in combination with the dynamic *authority record of publication channels* are innovations from a bibliometric perspective. The documentation system presents "institutional publication lists" in the same structured and controlled manner as we are used to in traditional bibliometric data sources. However, it improves substantially the coverage of these databases.

There is not the same documentation system used in all countries, but countries such as Canada have developed a 'Common CV' system (a repository of curriculum vitae of researchers, including publication details); Slovenia has implemented the SICRIS system[10] which captures bibliographic information on all the output of its research base to internationally-approved standards; There is the Finnish national database (KOTA – referred to in Section 3 above); and the Netherlands has the sophisticated DAREnet[11] repository of academic output from all Dutch research institutions. A number of other European countries, including the UK, are considering constructing similar database systems. If such databases could be electronically linked, and the bibliographic information in the system standardized and supplemented as required, evaluators would then have the best possible database for conducting detailed, accurate, in-depth bibliometric analyses.

For all the criticisms of the Norwegian model, it has at least demonstrated the feasibility of creating comprehensive data of good quality for a nation's total scientific and scholarly publication output.

4.3.6. Database developments

Many of the criticisms outlined in section 4.2 (above) are aimed at the Thomson ISI Web of Science as the most commonly used database for bibliometric evaluation. Thomson's Arts and Humanities Citations Index (AHCI) suffers from all the deficiencies enumerated there. It is acknowledged that there are a number of commercial competitors to Thomson, and that either in themselves or in conjunction with the services provided by AHCI there may be the possibility of

[10] http://sicris.izum.si/default.aspx?lang=eng

[11] http://www.darenet.nl/en/page/language.view/repositories

developing a more adequate bibliometric tool for evaluating humanities research. Google Scholar, for instance, indexes large amounts of scholarly material apart from journals.

There have been no extensive studies of how much these other databases would complement ISI in the humanities. However, Professors Michael Norris and Charles Oppenheim of Loughborough University in the UK were commissioned by the Economic and Social Research Council to assess the coverage by various abstracting and indexing services of the social sciences[12]. Many of their conclusions can be applied – with caution - to the humanities as well.

The study compared a range of UK journals and journal titles with the contents of 4 databases – *Web of Science, Elsevier Scopus, Google Scholar* and *CSA Illumina*.

The study also compared the content of these databases with 581 social science journals from France, Italy, Germany and Spain (derived from the International Bibliography of the Social Sciences (IBSS)).

It concluded that Scopus' coverage in terms of article level and citation frequency was superior to all the others[13]. When combining databases was attempted, the combination of Scopus and CSA Illumina produced the best results. However – and this is important with regard to European benchmarking in the humanities– *non-anglophone journal coverage was poor* overall and attempts to find *an authoritative source of monographs* with which to benchmark the holdings of the databases was unsuccessful. CSA Illumina had the best record in this respect, with over twice as many non-anglophone journal holdings as Web of Science and Scopus.

Google Scholar was noted as a promising resource. It covers monographs and book chapters, and has a significant recall of articles and journals. However, its performance in terms of accuracy and functionality (it is not possible to save or manipulate citation records in any way) were deficient. The sources for its

[12] Bibliometric databases – scoping project, Norris and Oppenheim - http://www.esrc.ac.uk/ESRCInfoCentre/Images/Bibliometric_Databases_Scoping_Project_tcm6-18363.pdf

[13] Although Scopus' holdings do not go back before 1996.

holdings are not transparently recorded. It should be noted that this product is still in its beta phase.

If merging existing databases to combine their strengths is not considered much of an improvement over existing practice, there is still the hope of persuading major commercial publishers to expand their current databases. The commercial incentives in the humanities may be slight, but there is a number of trends in publishing which could militate in favour of this approach: (i) companies looking for an edge in a increasingly competitive climate (for example, Thomson's near monopoly of the market does not look as secure in the face of competition from Elsevier and others); (ii) data input costs have dropped dramatically in recent years with the advent of electronic data interchange (EDI). It would cost Thomson very little to expand coverage.

In the light of these considerations, a coordinated and concerted campaign of lobbying by national agencies, together with the promise of financial or other assistance, would potentially be very effective at this point in time.

4.4 Conclusions

Despite a number of promising avenues for bibliometric assessment of the outputs of humanities research, the feasibility of a European-wide benchmark of quality must still be in doubt for the following reasons:

Alternatives to ISI Thomson databases still do not have the range of coverage of publications that would be adequate to benchmarking humanities research outputs across Europe.

Mining citation databases to extend coverage does not address the lack of coverage of non-anglophone journals adequately. It also does not cover an important channel of citation in the humanities – book to book citation.

Using alternative methods of gauging the impact of outputs such as the h-index may represent an advance on traditional citation counts. However it is not clear that – due to current deficiencies in citation databases – that applying this to the humanities would produce distortions and inaccuracies.

Any system of bibliometrics for the humanities will need to address two main issues: (i) adequate coverage of the full range of publications and (ii) agreement on how quality should be assessed. The best hope for such system would appear to be a simplified weighting method as outlined above, with the weights agreed through consultation and peer review. The feasibility of such a system for benchmarking European research outputs depends on the feasibility of the following features of such a system:

Agreement on, and implementation of, European-wide common documentation system of outputs
Agreement on an authority record of publication channels
The existence of a stable, continuous, transparent peer review process
Extensive and regular consultation with the academic communities across Europe
Regular review mechanisms

Clearly, this would be a massive undertaking requiring the mobilization of a large amount of resources within the higher education sector as a whole. Much would depend on the evolution of the ERIH and the community's view of its usefulness as a tool for this purpose.

5. EVALUATION AND PEER REVIEW

5.1 Introduction

Given the criticisms and conclusions outlined in the previous section, it is clear that any evaluation of the outputs and outcomes of humanities research will need to retain peer review as its core element and primary mode of assessment. The issue that this section will address is how the primacy of peer review can be maintained, while also keeping the following desiderata in mind:

(i) The feasibility of conducting large-scale transnational comparisons;
(ii) The method should not be overly burdensome on peer reviewers
(iii) The evaluations should be feasible enough to be conducted on a regular basis to allow the construction of a robust time-series;
(iv) The method should allow the peer review to be informed by and, if necessary, corrected by, relevant quantitative information (in line with **Recommendation 7** above).

There are a number of ways to approach this kind of evaluation. Some of the existing methods have been outlined above in Section 3 on the State of the Art and with reference to the European Reference Index in the Humanities in Section 4 above. In addition to these examples of existing European practice, this report would like to outline 2 case studies of peer-review-led evaluation exercises which would be complement and add value to those methods, while also being consistent with the framework laid out in Section 2.

5.2 Informed peer review and systematic evaluation: case study 1

The US National Research Council conducts a decennial review of PhD programmes in American universities[14]. The PhD assessment is the current version of assessment projects that have been conducted approximately every 10 years since the 60s. While it attempts to assess programs in all fields of arts and sciences over a certain size, the following discussion relates only to that part that is directly related to assessing humanities programs.

[14] The following description owes a heavy debt to a paper presented by Professor Norman Bradburn at the HERA workshop in London on international benchmarking, January 2007. I am very grateful for Professor Bradburn's permission to reproduce that text here.

The project has 4 data parts. The first uses a program questionnaire filled in by university officials responsible for the programs. This questionnaire provides largely descriptive data about the programs, and contains detailed information including the names of the faculty, number of students, levels of support and, most critically, data on drop out rates and time to degree completion. The second part is a faculty questionnaire in which data about faculty activities, research and publications are obtained. The third is an experiment in which we are sending a questionnaire to students in a few fields. For humanities the field is English. The fourth part is data gleaned from public records such as research awards, publications, citations, and honors and awards.

Two quite different methods are used to make assessments. The first is what called *the explicit method*. In the faculty questionnaire respondents are given two sets of criteria that they might use to assess program quality in their field, one for assessing faculty quality and one assessing student quality. For faculty quality, the following indicators are used:

- number of publications per faculty member
- number of citations per faculty member
- receipt of extramural grants for research
- involvement in interdisciplinary work
- racial/ethnic diversity of program faculty
- gender diversity of program faculty
- and reception by peers of faculty members' work as measured by honors and awards

For student quality, the following indicators are used:

- Median Graduate Record Examination (GRE) scores of entering students
- percentage of students receiving full financial support
- percentage of students with portable fellowships
- number of student publications and presentations
- racial/ethnic diversity of the student population
- gender diversity of the student populations
- having a high percentage of international students

For each set of characteristics faculty members are asked to mark up to four that they feel are the most important, then from those four pick the two most important.

It is expected that the relative weight given to the various characteristics will vary by field. For example, it is expected that history faculty may weight extramural grants or number of publications differently from faculty in linguistics. Data from the faculty survey is used to construct weights for each of the characteristics separately for each field and separately for faculty and student characteristics. Then assessments are constructed based on these weights. Note that each of the characteristics can be measured by objective data, except perhaps number of citations in some of the fields. Nonetheless, there is important peer-review input in constructing the weights and the method gives due regard to the disciplinary variation in **Recommendation 2**.

The other method used is called *the implicit measure.* This is considerably more complicated. In past assessments, the principal criteria used were reputational measures obtained by using faculty raters who graded programs in specific universities on a 6-point scale from outstanding to poor. A number of dimensions were rated, but because of the high degree of correlation among the ratings of the dimensions, the single rating of quality of research done by the faculty of the program became effectively the only rating used. University programs were then ranked according to this criterion producing a ranked set of universities that provoked a considerable amount of controversy, even amongst those who were rated highly using this method.

Reputational measures have been heavily criticized for a number of reasons, and mirror many criticisms of the use of peer review in evaluation. Among the most serious relate to bias and/or ignorance. It is said that reputational measures (i) favor large programs (ii) are not very reliable once one gets away from the best known programs and (iii) they may be based on limited - perhaps outdated - information, prejudice, and on some occasions ignorance. Despite these problems, it has been observed that reputation is an important *social fact* that does capture in summary form many aspects of the quality of programs that even a weighted average of the objective measures does not. The key issue is how reliable measures of reputation can be constructed from peer inputs.

In the most recent dicennial review of PhD programmes, it was decided that direct reputational measures would not be used as a basis for quality assessment of the programs. However, the evaluation faced a dilemma in that there are no other measures that come close to being a gold standard, and many people are as critical of the explicit method as are critical of using a reputational measure. Based on some statistical analysis that has been done on the reputational measures in the 1995 assessment, it is believed that it is possible to construct a statistical model separately for each field that will enable reputational measures to be inferred as if they had been directly obtained for every program that is assessed.

The proposed procedure involves the following steps: First, separately for each field, a small sample of programs that will serve as anchors for the statistical model are selected. Second, a sample of faculty active in that field - from universities other than the sampled programs - are selected. These act as raters for the sampled programs. These raters are given descriptive information about each of the sampled programs, such as the names of the faculty in the program and some of the information related to characteristics that were deemed important as judged by the faculty in response to the questions on explicit criteria of quality. The raters are asked to rate how familiar they are with the program being rated, and then to rate the program quality on a 6 point scale. Third, after the ratings are in, a large number of variables are regressed: these variables are thought to be related to quality on the ratings of the sampled programs that serve as the anchors. The resulting equation is used to impute quality ratings for the remainder of the programs. This is called the anchoring study. This method is called an implicit method because it attempts to use statistical methods to decompose the global reputation ratings into component parts that are implicitly used by the raters and can be objectively measured.

In all of these exercises there are two crucial steps that make the resulting ratings believable or not. The first is the selection of the measures to be used in the ratings, and the second is how the measures are combined into an overall index of quality. The first is much easier than the second. The first step can be approached by asking faculty members to indicate what measures they think are important in assessing quality in their field. The evaluation also has a committee of experts composed of former university presidents and provosts, graduate deans, and distinguished scholars, who advise on the relevant measures. In this way, the credibility of the measures used is established.

The more difficult issue is how to combine these measures into an index that will allow quality comparisons among programs. Two methods for assigning weights in combining the measures will be used. The first will be the explicit weights derived empirically from the results of the faculty surveys. The second will be the weights derived from the statistical models built on the anchoring study that uses peer reputation as the standard. The statistical method has the advantage that it quantifies some of the sources of uncertainty. It is expected that the two methods will produce largely similar ratings, but as the exercise has yet to be completed this has not yet been demonstrated. The ratings will be published in several large quality groupings and not in an explicit rank order as was done in the past. Such ranking gives a false sense of precision and deflects attention from the content of the assessment to a university's standings in a league table. The results will be published at the end of 2007.

The method is a useful case study in that it combines a high degree of peer input with a large degree of quantitative information in a way that allows a large number of programmes to be evaluated and compared. It can be seen that peer input is solicited at three distinct levels: (i) in the selection of indicators (ii) in the weights applied to these measures and (iii) in assigning the overall quality measures to the anchoring studies in the implicit method.

5.3 Informed peer review and systematic evaluation: case study 2

In response to common criticisms of the Thomson ISI database's deficiencies, the Australian National University's Research Evaluation and Policy Project (REPP) had embarked on an Australian Research Council (ARC) Linkage project to investigate the development of novel indicators that could potentially be used to evaluate and assess research performance. An important part of this project was the construction of an extensive database of Australian university publications. The need for this initial database construction needs to be borne in mind when assessing the possibility of a European-wide implementation of the method outlined below.

One of the projects concerned the production of a ranking of conferences into prestige tiers in the field of information and computing science (ICT). This field was chosen as conferences and conference proceedings are an important channel of dissemination and publication of research findings, but they are poorly represented in commercial bibliographic databases. The comparisons with the

humanities are clear, and the following method - which is fast gaining a reputation for best practice – could be used to determine rankings of esteem measures or publishers. The role of peer input is crucial.

The project team identified all conferences from ICT departments were identified, and supplemented by similar outputs from multidisciplinary departments incorporating ICT disciplines. The initial task of cleaning conference names took considerable effort, as they are far less standardised than journal titles. The project team used additional bibliographic details supplied by universities (such as conference date and location) to assist in the process where the identification of the conference lacked certainty.

Additional sources of information that might prove useful for ranking conferences were sought, and wherever possible contextual details were added. The type of information elicited covered a University of California Davis ranking of conferences, and their classification to sub-discipline; CiteSeer rankings and citation data; ISI Web of Science citation data (extracted in the course of the ARC Linkage project); conference acceptance rates; and the number of publications reported by Australian universities.

A workshop of ICT researchers from a number of universities and across the range of disciplines was convened to continue the ranking process. Their tasks were to:

- decide on the number of tiers;
- draft the descriptors for each tier;
- delineate the ICT .areas. or disciplines;
- edit the draft rankings; and
- add any conferences missing from the lists.

Once the revised conference lists had been finalised, they were distributed throughout the academic community for further input and comment. This was undertaken by peak bodies in computing, information sciences and electrical engineering. The aim of this last step was to comprehensively populate the top three of four proposed prestige tiers. All remaining conferences were classified to the fourth “unranked” tier. Respondents were asked to justify any proposed movement between tiers. The descriptions for the four tiers are shown in the box overleaf.

As can be seen from the descriptors used, particularly for Tier 3, typical Australian bluntness was used to ensure that the differentiation between tiers was explicit and easy to interpret. An early version of Tier 1 included the statement that it could also be typified as one where “people from overseas congratulate you for getting in, and you shout drinks to your colleagues”.

Two further steps were planned, but have yet to be undertaken. The conference lists were to be sent to international groups with an interest in the process for external validation. As a final step, the project team would develop performance measures based on the rankings and test them using data from universities, after which a final
workshop will be convened to assess the indicators. Only then will it be possible to determine the efficacy of measures based on conference rankings.

Descriptors for Conference Tiers

Overall criterion: Quality of the papers presented at the conference

Tier 1
Typically a Tier 1 conference would be one of the very best in its field or subfield in which to publish and would typically cover the entire field/subfield. These are conferences where most of the work is important (it will really shape the field), where researchers boast about getting accepted, and where attendees would get value from attending even if they didn't have a paper themselves. Acceptance rates would typically be low and the program committee would be dominated by field leaders, including many from top institutions (such as Stanford, MIT, CMU, UC Berkeley, Cornell, UWashington, UTexas, UIllinois, Oxford, Cambridge, Edinburgh, Imperial College, Microsoft Research, IBM Research, and so on). Tier 1 conferences would be well represented in the CV of a junior academic (assistant professor) aiming for tenure at a top 10 US university.

Tier 2
Publishing in a Tier 2 conference would add to the author's respect, showing they have real engagement with the global research community and that they have something to say about problems of some significance. Attending a Tier 2 conference would be worth travelling to if a paper got accepted. Typical signs of a Tier 2 conference are lowish acceptance rates and a program committee and speaker list which includes a reasonable fraction of well known researchers from top institutions (as well as a substantial number from weaker institutions), and a real effort by the program committee to look at the significance of the work.

Tier 3
Tier 3 covers conferences where one has some confidence that research was done, so publishing there is evidence of research-active status (that is, there is some research contribution claimed, and a program committee that takes its job seriously enough to remove anything ridiculous or ignorant of the state of art), but it's not particularly significant. This is where PhD students might be expected to send early work. It also includes places whose main function is the social cohesion of a community. Typical examples would be regional conferences or international conferences with high acceptance rates, and those with program committees that have very few leading researchers from top international institutions.

Tier 4
All the rest.

Results

Details of over 11,000 conference publications from ICT and related departments were identified and standardised (by name and abbreviation). While close to 1,500 discrete conferences were identified, just 72 conferences accounted for half the ICT publications reported by universities, and 150 took in two-thirds of the total. The initial classification process focused on these conferences and any others identified in published rankings. Additional conferences were added during the initial workshop, at which stage 479 conferences were ranked into the top three tiers within 13 discipline areas. A further 348 were discussed and allocated to the fourth "unranked" tier.

When approaching the task of classifying the conferences, the lack of additional information was disappointing. Of the 827 conferences that were classified, we only had ISI citation data for 359. In most cases these were based on a handful of publications and the data was not considered robust. CiteSeer rankings contained data for only 194 of these conferences, with many newer conferences missing from the most recent listing (2003). Acceptance rates could only be found for 96 of the conferences. The most useful sources of additional information were the published rankings which provided information on 441 conferences.

Lack of additional qualitative information for many conferences meant that the project relied more heavily on peer opinions than expected. However the classification process was considerably helped by the tongue-in-cheek, yet highly effective, set of descriptors for the tiers. The biggest difficulty encountered related to the delineation of disciplines in this rapidly evolving field, and the inappropriateness of the standard Australian research classification scheme.

Initial feedback from the final community-wide consultations suggest few changes will be made to the penultimate ranking list, though there still exists considerable disquiet with the delineation of disciplines. The success of the project in identifying performance measures based on conference rankings will ultimately rest on the results of testing their application using "live" data.

There are clear echoes of the process used for the European Reference Index for the Humanities in this method. Although the conference rankings project managed to achieve a degree of consensus through a painstaking method of consultation, this consensus is more easily achieved when the academic community is confronted with the option of more unpalatable metrics and indicators (such as peer-reviewed research income) being imposed, as is the

case in Australia. As the HERA survey on impact and quality assessment practices discovered, there are very few European countries where a similar policy environment is currently to be found. This may of course change in the near future, as governments increasingly look to more transparent methods of allocating research resources and evaluating their outputs.

5.4 Conclusions

Both the case studies above give examples of how informed and systematic evaluation of research fields can take place without the need for direct peer review of each output. It will be important to investigate such methods if credible and robust methods for the international benchmarking of the outputs and outcomes of humanities research are to be feasible. Nonetheless, there is a place for sustained and continuous peer input which is consistent with the recommendations outlined in Section 2 above.

However each of these methods require the marshalling of large amounts of resources in order to function properly – large amounts of accurate and comparable data-gathering in the case of case study 1 and the prior creation of a database of the outputs of the higher education sector in case study 2. The feasibility of such methods in the European case are dependent on the willingness of public agencies to commit resources and time to this vital bases of any evaluation that is to have the confidence of policy-makers and the academic community alike.