Center for
Higher Education
Policy Studies

# Between certainty and comprehensiveness in evaluating the societal impact of humanities research

## Reflections on a decade of Dutch experiences

**CHEPS WORKING PAPER 02/2015**

Paul Benneworth. CHEPS (University of Twente)
p.benneworth@utwente.nl

**Series Editor Contact:**
Paul Benneworth
Centre for Higher Education Policy Studies
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands
T +31 53 – 4893263
F +31 53 – 4340392
E p.benneworth@utwente.nl
W www.utwente.nl/cheps

# Table of Contents

## Abstract

Research evaluation is a tool that can be used for many different purposes, with every different kind trading off comparing and understanding activities and seeking to treat evaluation subjects fairly. Evaluation problems can emerge when an approach that seeks to give one kind of fairness is used for a set of purposes demanding an alternative perspective on fairness. These problems of course afflict all kinds of evaluation, not just in research, but there has been in the last decade an increasing awareness that they are prevalent in research evaluation systems that seek to make judgements within national research systems. Clearly there are the risks that problems may emerge when attempting to use these very limited indicators to measure and reward university research impact in a systemic way. This paper therefore asks how can evaluation of research impact at the systems level –deal with the problem of the very different mechanisms by which different kinds of research produce their impact? We explore this question via a case study of the Netherlands, where policy-maker driven attempts to capture impact within the research evaluation system awoke fears amongst the humanities research community that they would not be treated fairly. On this basis, the paper argues that more reflection is demanded of scholars on what kinds of research impact matters in their field, and how that messiness of impact generation legitimates a multi-disciplinary, judgement- and discretion-based system that ultimately values activities and outcomes which lie beyond the pale of their own scholarly norms.

# 1. Introduction

Research evaluation is a tool that can be used for many different purposes: Molas Gallart (2012) characterises those uses as falling into three classes, namely at the systems level (e.g. allocating resources), the institutional level (controlling activities) and at the operational level (improving practises). Each of these different use classes makes a different trade-off between a need to be able to compare very different kinds of activities, and between getting a fine-grained understanding of those activities (Molas Gallart, 2015). In effect, different kinds of research evaluation that seek to be fair (Huang & Chang, 2011) have to make a choice between two kinds of fairness with regard to the agent (the entity being evaluated). The first of these is *between-agent* fairness, when what is important is to make sure that similar kinds of performance by different actors (even where the activities are very different in their nature) are similarly rewarded (Blockmans, 2007), thereby providing an incentive for different providers to raise their overall performance (*cf.*. Kickert, 1995). The second of those is *within-agent* fairness, where it is important to judge a unit under assessment against in the context of its own circumstances, what that unit has agreed to do, and what counts as good performance given those contextual differences.

Evaluation problems can emerge when an approach that seeks to give one kind of fairness is used for a set of purposes demanding an alternative perspective on fairness. Where excessive attention is paid to individual idiosyncrasies, an evaluation system rapidly loses its ability to make fair distinctions between different actors, making comparisons almost possible, and therefore undermining the reasonableness of resource allocation made on that distinction. Where a system attempts to compare very unlike attributes, the system outcomes reflect arbitrary choices made in the way that particular characteristics are weighted. Under these circumstances, the best outcomes are defined determined by basing evaluations on what all participants are trying to achieve, and rewarding the desired performance, a ground rule of new public management (*cf.* Ferlie *et al.*, 1996). When an evaluation is concerned with a single actor, then basing the evaluation on a standardised template can often lead to unfair treatment, and in particular the disregard for things that are important or valuable in the real context but which have for whatever reason been omitted from the evaluation protocol.

These problems of course afflict all kinds of evaluation, not just in research, but there has been in the last decade an increasing awareness that they are prevalent in research evaluation systems that seek to make judgements within national research systems (Donovan, 2007). Even where conceptually underpinned, research evaluation has as a field has created strong practices and assumptions often on the basis of successful experiments. The whole field of bibliometrics as a tool for the evaluation of scientific excellence has emerged on the basis of available comparable data, provided in databases such as ISI, which have at best a very partial coverage of subjects in the social sciences and humanities(Van Raan, 2005; Pontille & Torny, 2010). But this paper is concerned with another element of research evaluation which has been subject to the same emergent path dependency, not of scientific excellence but rather of societal impact. If bibliometrics as a field has emergent characteristics, then metrics of research impact have emerged not just in an emergent way, but arguably hastily and ill-considered (Bozeman & Sarewitz, 2011). Under significant pressure from policy-makers to justify the value of public research investments, discussion has often focused on a very limited set of research impact measures primarily focused on commercialisation activities (e.g. license income, patents, start-ups) most relevant to a very limited number of disciplines, notably the pharmaceutical sciences (Spaapen & Van Drooge, 2011; Benneworth, 2015).

Clearly there are the risks that problems may emerge when attempting to use these very limited indicators to measure and reward university research impact in a systemic way. Indeed, Sweden recently suspended its plans to introduce a direct financial reward for research impact precisely because of a realisation there were no good metrics to evaluate that research. This paper therefore asks the research question of how can evaluation of research impact at the systems level – aiming to give between-agent fairness – can deal with the problem of the very different mechanisms by which different kinds of research produce their impact? We operationalise this in terms of the question of whether peer-review offers the only mode of fairness, or whether there are quantitative/indicator led approaches that might be able to deliver that fairness at a To explore this question, the paper uses a case study where the problem clearly comes to the fore, the Netherlands, where policy-maker driven attempts to capture impact within the research evaluation system awoke fears amongst the humanities research community that they would not be treated fairly. On this basis, the paper argues that

more reflection is demanded of scholars on what kinds of research impact matters in their field, and how that messiness of impact generation legitimates a multi-disciplinary, judgement- and discretion-based system that ultimately values activities and outcomes which lie beyond the pale of their own scholarly norms.

## 2. The problematic of evaluating arts & humanities research impact

The issue of research evaluation is extremely empirical in its nature (it is driven by a practical set of concerns about managing science bases and policy), and therefore some degree of conceptual clarity is necessary from the outset. Citation analysis emerged as a tool to help scientists better understand and structure their knowledge of the field, and only later evolved into bibliometrics as currently understood (Trolley & O'Neill, 1998). There are many reasons that research might be measured at the systems level, particularly to take resource decisions. There are a number of research funding systems that adopt a formula funding approach (for example in Scandinavia and Flanders, see *inter alia* Debackere & Glänzel, 2004; Erikson, 2011), which typically allocate points to particular kinds of outputs and allocate resources on the basis of those points. In so doing, they make no judgements about quality, rather the basis of the system is to reward creating particular kinds of outputs, be they journal articles, book chapters; outputs that fall outside one of these classes simply have no value as far as the system is concerned. But at the same time, they do not say that these have no value, just that a choice is being made not to reward them. That is quite different from national evaluation systems which attempt to compare on the basis of substantive content, what might be considered as the idea of objective quality or excellence (something which we immediately acknowledge is an intensely contested category). In these systems, such as the UK or French systems, grades are attached to units (which may or may not carry funding consequences) (Martin, 2013; Chatelain-Ponroy *et al.*, 2014).

There is a substantial literature of science policy and higher education sociology that disputes whether research quality or excellence exist in an absolute sense, and therefore research evaluation is doomed to fail. To sidestep that fundamental discussion, we are making clear that here we are using research excellence here as a shorthand to refer to

the extent to which a piece of research produces an idea that influences and stimulates others in their own research activities (*cf.* Sarewitz & Pielke, 2007). Likewise, research impact refers to the extent to which the knowledge created in research projects creates capacities within societies to do more good things (Corea 2007, cited in Benneworth, 2014). The task of research evaluation of those two qualities is therefore seek to get a sense of the relative order of magnitude by which researchers, projects and outputs achieve these two respective outcomes and provide a means for ranking between these orders of magnitude.

I therefore here class this problem as a question of measuring intrinsic quality (rather than of extrinsic outputs), and therefore of raising a problem of absolute fairness between researchers. Their purpose is not simply say Research Unit A has produced more articles than Unit B and deserves more funds, but to say that Unit A is better than unit B. So for these systems to have credibility within the scientific community, they then need to be able to guarantee fairness and objectivity, and indeed in the preceding example for in a clearly defined way Unit A to be better than Unit B. This is not a straightforward issue when attempting to make systems-level evaluations, because there are clear differences, particularly between Science, Technology Engineering and Mathematics (STEM) and social sciences and the humanities (SSH) disciplines in the nature of the outputs of this research and the outputs which are effectively covered by (Huang & Chang, 2011). Our argument is that comparing between disciplines raises a fundamental issue of fairness because of the very different ways in which these disciplines define and deliver impact (e.g. Olmos Penuela *et al.*, 2014).

That is not to say that fairness cannot be achieved. The case of the UK Research Assessment Exercise seeks to make a definitive comparison of all university-based research and to be on some level a fair comparison that allows resources to be allocated on the basis of real quality rather than crude output measures. The Exercise has evolved from a light touch partial subjective review in 1986 to 2014's mammoth bureaucratic exercise with very detailed protocols to process every research unit and reduce that to a single profile. That profile expresses percentages of research publications, environments and impacts that are respectively world leading, world class, internationally excellent, national excellent or not excellent. That simple numerical statement for each unit is the basis for allocating around £1bn core government

research funding annually (Brown, 2013). The fairness in the UK system being provided by a process of peer review with panels given wide-ranging discretion in how to apply their guidance. In theory the review panels were supposed to judge every single item on its merits, and pay no heed to the forum within which appeared, allowing patents and working papers – if sufficiently 'excellent' to acquire the highest rating.

The latest round of the system, in 2014, for the first time included an Impact component and allowed institutions to submit case studies (1 per ten submitted staff) which were then evaluated on the basis of their reach and significance (HEFCE, 2011a). Impact is defined in a very broad way, under 12 very broad subheadings covering almost every facet of socio-economic development (RCUK Undated). Mindful of the potential for there to be disciplinary differences in performance, HEFCE commissioned a set of case studies of impact that found no evidence to back up the assertion that social sciences and humanities would be systematically disadvantaged by this approach (Benneworth *et al.*, 2016). This has proven an extremely expensive process to organise, costing around £250m per research evaluation (PA Consulting, 2008), and in allocating science funding for 7 years, it therefore represents an overhead of around 4% (although unofficial estimates have it much higher, Bowman, 2015). The high financial and opportunity costs of the UK system have long been recognised in a number of accountability reviews of the system, and the exercise has attracted substantial criticism from a variety of angles and in particular for the fact that the system outcomes reflect arbitrary assumptions rather than being an objective comparative framework (Holmwood, 2010; 2011). The UK assessment exercise considered trying to simplify the exercise by adopting bibliometric measures as far as credibility would allow, but a pilot exercise in 2008-09 concluded that such an approach would not be credible (HEFCE, 2014), and attempts to retrospectively model results using bibliometric data have proven unsatisfactory (Mryglod *et al.*, 2014)..

The point of this paper is not to reflect critically on the UK impact system (see for example Martin 2011 for a fuller treatment) but to highlight the relative nature of fairness and how expensive it becomes to deliver an approach that by no means enjoys universal credibility. Policy-makers are keen for research to be held accountable for its societal as well as scientific benefits, and therefore to demonstrate its wider societal

value. The nature of that pressure is made clear by a recent statement from the Australian Research Council who recently noted:

"There is an increasing focus on showcasing or measuring the societal benefits from research, and a need for better coordination in reporting and promoting the impact of these research outcomes. This will become increasingly important in a tight fiscal government environment where returns on investment in research will need to be demonstrated in terms of environmental, economic and social impact. For these reasons and others, key stakeholders including government, industry and the community require more information on the benefits derived from investment in Australian research activities." (ARC, 2015).

This is not merely an issue where particular kinds of outputs can be rewarded, because there is no clarity about how research creates impact, and certainly great hostility amongst academics to the idea that purely economic impact should be rewarded (and that resistance is important given the importance of research evaluation systems to have a degree of credibility). In particular social sciences and the humanities have been particularly aggressive in mounting a defence that there are many different ways that impact can be produced from research (*cf.*. Crossick, 2006; 2009; Bate, 2011; Brewer, 2013; Small, 2014; Bod, 2012) and that any credible research evaluation system should capture the diversity of approaches in ways that allow fair comparison between different kinds of impact. Clearly, the UK system has evolved in response to that pressure to acknowledge a very broad array of the 'pathways to impact'. And it is from that breadth of potential avenues for research impact that the research impact evaluation problem arises – how can a research evaluation system allow a fair comparison the intrinsic value of research impact provided by units seeking to produce that impact through an extremely diverse array of outputs? In the UK, a degree of fairness is provided by allowing peer reviewers huge discretion to evaluate that impact and create their own definitions of impact, in the hope of making credible comparison between disciplines and fields with sometimes very different ways of creating international impact.

The UK provides one mode of ensuring fairness, through peer review, with a drawback that it is extremely resource-intensive in return for producing results that capture

comparable relative research impact. And as the quotation from the ARC above makes clear, in a resource-intensive environment, devoting a significant share of research resources to ex poste valuation raises the question of whether it is possible to achieve the fairness and credibility without the burden of peer review. Clearly, policy-makers have flirted in other countries with attempts to evaluate research impact without this burden, including through the use of indicators. We therefore operationalise our research question to ask the question of can the use of indicators provide another "mode of fairness" in research evaluation. Although there is mixed evidence regarding whether scientific excellence valuations produced through peer reviews can be predicted can be (Myrglod, 2013a; 2013b), we argue there has been much less consideration of this issue around scientific impact. To answer this operational question, we look at a case study of the Netherlands, where there have been attempts to develop metrics into a research impact evaluation system that attempts to produce fair evaluations between units.

## 3. Case study overview & methodology

The issue of a 'mode of fairness' is not a question of objectivity, rather it relates to the question of whether a system which regulates and allocates resources between relatively autonomous actors is held in regard as being credible by those who it regulates (Jasanoff, 2003). To answer this question, we therefore look at the case study of the Netherlands, which since the late 1980s been assessing its research base in order to maximise its research quality. In this research system, every research unit (centre, department or group) has to ensure that within a specified multi-annual period, that it undertakes a review following the Standard Evaluation Protocol (SEP) following a process in which a self-evaluation report is evaluated by international peers following site visits (*cf.* KNAW, 2010). Research units are awarded marks reflecting the quality of their research and then it is left to institutions themselves to decide what to do with those reports.

Research units have considerable freedom to decide how to undertake that, whether individually, as a collection of units or all units in a discipline in the country. In the last decade, there have been two important evolutions (Benneworth, 2014). Firstly the evaluations have become increasingly driven by data relating to inputs and outputs, allowing both a standardisation and comparison between units of assessment but also

the (subjective peer reviewed) judgements to be underpinned by 'objective measures'. Secondly, since 2009, the Standard Evaluation protocol has been iterated to include measures of social impact to demonstrate not only that research is internationally excellent but is contributing to societal development, although societal quality has been an element of evaluation since the system was formally introduced in 1993, albeit as a composite variable relevance (societal-scientific) (Van der Meulen & Rip, 1995; 2000). These two trends set the context for the case study – there has been one tendency for the use of objective data to give foundation to subjective judgements, but at the same time, the introduction of a new 'impact' domain, societal impact, where there is no agreed standards for societal impact.

This case study explores the interplay between these two trends and in particular the negotiations and political tensions that have flowed in seeking to create a system that has credibility, legitimacy and fairness. We here take the example of humanities (what is in the Netherlands called *Geesteswetenschappen*) because it is an area where these tensions are strongest. Whilst in engineering it might be the case that measures such as patents or license income might enjoy a sufficient degree of legitimacy to be incorporated unproblematically into the research evaluation system, in the humanities both generally and in the Netherlands there is no agreement of how impact is produced much less on measuring that impact (Worton, 2006; Crossick, 2006; Belfiore, 2013). Humanities in the Netherlands is strongly institutionalised in terms of the existence of faculties of humanities (whose Deans for a Deans of the Humanities working group), but also through the existence of a Council for Humanities within the Royal Netherlands Academy of Arts & Sciences (KNAW) and a specific funding council within the Netherlands Organisation for Scientific Research (NWO). This institutional framework constituted a discursive space in which a series of actors attempted to agree on a compromise that would allow the agreement of a set of indicators for the social impact of humanities research which enjoyed credibility in both scholarly and policy communities.

The basis for this case study was a qualitative single study which followed the discussion following attempts to agree a set of indicators of measures of social quality in the Netherlands. The case study sought to relate two separate systems, the systems by which humanities researchers created societal impact together with their research

users, partners and publics, in parallel with attempts by a much more restricted tripartite group to create a credible system of humanities research impact indicators, to understand how the tensions between the two tensions played out. The basis for this was a set of 46 semi-structured interviews with researchers, research policy-makers, researchers representatives and societal users. These were carried out in the period September 2011 to March 2012; interviews with researchers and their users started from publically visible examples of how those researchers had claimed to create an impact, and then speaking to users, intermediaries and users of that research (including media, policy bodies and civic society groups). Interviews with the 'indicator system' actors started by identifying and interviewing actors immediately related to the Cohen process, (and its successor the *Regieorgaan Geesteswetenschappen*) (see section 4), and then snowballing out to speak to others interviewed in the process. As part of this, the research team attended a key indicator development workshop in Waassenaar on 25th November 2011. This was paralleled with a wider documentary search of the key documents which emerged as actors attempted to create a definitive list of humanities research societal impact indicators, both through direct recommendation from interviewees but also through following citations and bibliographic searches; this documentation was important to chart the evolution of the proposed indicators after the fieldwork finished. The evidence so gathered is reported upon at length in Benneworth, (2013). A subset of all this material of relevance to the issue of the development of indicators for measuring humanities research impact was sensitively combined to create a narrative account of this attempt to produce credibility between two systems. From this narrative a set of stylised facts was deduced in order to address the operational research question, and thereby provide insights into our overarching research question. It is necessary to acknowledge that what is presented here is not a definitive statement of reality, but rather an attempt to identify critical events and their dynamics conceptualised as the interplay of tensions between two systems, without necessarily claiming that those tensions or systems correspond to real-world phenomena.

## 4. Towards a sustainable humanities in the Netherlands: the Cohen Commission and impact indicators

Although the humanities had formed the mainstay of the Ancient Dutch universities, in the late 20th century they had found themselves under increasing pressure in the context of a national higher education system increasingly concerned with driving technological modernisation. This manifested itself in falling student numbers and research capacities in parallel the increasing importance of project research funding through organisations such as the Dutch Organisation for Pure Research (ZWO, later NWO). The sector was in crisis by the 1980s, and three Commissions of Inquiry, Staal (1991) and Vonhoff (1995) and Gerritsen (KNAW, 2002) sought to find a way to accommodate a humanities sector that was organised on a very different model to the natural, life and social sciences. Although the first of these had resulted in a new funding stream for the humanities, this had not resolved the issue that the sector felt under threat from science system changes which lacked relevance to it. From 2002, Research Funders in particular appreciated that if the problem was that humanities was suffering because it did not fit well in the science system, then a claim of exceptionalism was no longer valid, and therefore it was necessary to change humanities so that it did fit well into that system. In particular, attempts were made to boost a number of norms common elsewhere in the system, including research in larger international consortia, publication in English, the use of large research infrastructures and the adoption of new technologies as seen in the rise of Digital Humanities (see Benneworth, 2013, for more detail).

An important norm here was the idea that humanities research had the potential to be valuable outside the academy as well as representing global excellence. The Gerritsen Commission report had argued that humanities (and in particular rare languages) had public value in terms of creating "a window on the world" that could allow the Netherlands to find stability and security in a world that they were busy discovering was unstable and dangerous (Bosland, 2010). But through the 2000s, the fear persisted that successive modernisations of Dutch science policy, and in particular the rise of the idea of research impact (what is referred to in Dutch policy terms as valorisation) was systematically disadvantaging the humanities. Following a particularly critical parliamentary question the Advisory Council for Science and Technology (2007) produced a report on how humanities and social sciences could create public value and

hence be compliant with the emerging policy fashion for valorisation. Later in that year, another Commission of Inquiry was established to formalise the switch, and to find ways to put humanities in the Netherlands on a sustainable footing.

This report, Chaired by former Education Minister Job Cohen (published as Cohen (2009)) highlighted a number of key problems for the humanities, of which I highlight two of salience for the discussion about credible impact indicators. Firstly was the lack of concern within the field for the generation of impact (perhaps slightly overlooking the work of NWO in promoting impact as a prerequisite for receiving research funding). Secondly was the importance of an increasing acceptance) the use of indicators as a means of validating the international research excellence (not here framed in terms of impact excellence) of leading Dutch scholars. This report sought to ensure that humanities were compliant with the overarching norms of the science and education system as a whole and no longer pleading exceptional treatment. This led to the committee recommending that the KNAW Council of the Humanities "Take the initiative in developing a system of quality benchmarks for the Dutch humanities which is clear, adequate and as simple as possible" (p. 45). In parallel with this, NWO introduced a 'bonus point' system in its *ex ante* research evaluation where well-argued valorisation was planned. In 2009, the Minister for Education, Culture and Science (OCW) appointed a National Task Force for Sustainable Humanities (Otten, 2015) to implement the Cohen recommendations.

A body was specifically created - *"Regieorgaan Geesteswetenschap*" - with a multi-million Euro budget to make recommendations for the Minister invest a new budget in implementing Cohen's recommendations (OCW, 2009a), clearly to the surprise of some who had felt that the report would – as with its predecessors – disappear without trace. The membership of this Oversight Board was formed from prominent Dutch and Flemish humanities scholars in Europe and the US with a mandated expiry in 2016. In 2009, as an immediate first step, all humanities faculties were entitled to submit plans for strengthening their humanities (OCW, 2009b). Although all investments were notionally provided to universities to allocate as they saw fit, they were at the same time required to report on how the particular allocated sums had been invested in the activities required, thereby avoiding the problem of the Staal resources of failing to be invested in strengthening humanities. The Commission developing longer term

recommendations for the implementation of Cohen, ,in particular around the creation of new Ph.D. positions given the relatively limited support for  funded Ph.D.s in the Netherlands (Regieorgaan, 2010).  A second action area, directly in response to the Cohen recommendation was that the KNAW created a Commission to develop indicators to allow research quality (including that of research impact) to be measured and differentiated.  The *Regieorgaan* in turn linked up with this Commission and committed to using their findings to develop a system to measure and differentiate humanities research quality. The *Regieorgaan* recommended that the KNAW would need to get a good coverage of the report and support both from the national disciplinary research schools as well as the other KNAW subject councils that had undertaken discussions of research impact (notably social sciences and engineering; KNAW, 2005; 2011).

## 5. Quality indicators for Dutch humanities research

The process of developing the quality indicators occurred through two stages, which could be characterised as experimental and regulatory.  The experimental phase involved the KNAW Commission exploring the issues raised by measuring humanities research quality and to present a range of options which could work in practice.  In the regulatory phase, the *Regieorgaan* sought to take an open-ended choice and make it concrete in a way that would meet both academic and policy needs, that it would be academically credible, and serve as the basis for fair comparison, ensuring that evaluation scores in the humanities were accepted as being comparable to performances in other discipline areas.  At the time of writing, the regulatory phase is underway, but it has undergone a number of delays that are at least suggestive that building that academic credibility and fairness has not been as seamless a process as might potentially be hoped.  There were a pair of KNAW projects already underway – Evaluating Research In Context (ERIC) and SIAMPI[1] which were seeking to generate good measures of how research created societal value; these were highly specific.  These projects had developed a framework by which researchers could identify useful impact

[1] Social Impact Assessment Methods for research and funding instruments through the study of Productive Interactions between science and society.

measures, and had been well-received as a way of providing an effective means of what could be thought of providing within-agent fairness. The challenge lay in finding a set of indicators that could serve to provide between-subject fairness.

The Council had deliberately asked the Commission to explore *indicators* for quality rather than *measures*, because it was clear from past experience that past measures would be extremely controversial and even could prevent consensus. The Commission was deliberately guided to avoid making recommendations that would have a strong behavioural effect on researchers:

"if we were to say that quality in publishing is publishing articles of less than ten pages in journals published in north east America, or that it is publishing books, then it is going to change behaviour" (Interviewee, 30.11.2011).

A concrete example of what they did not do was – as happens in Scandinavia and Flanders – was to create equivalence ranking, to say that quality of a book was worth a particle multiple of a journal article; central to the Commission's proposal was that human judgement remained central to the process. In particular, what where being described as indicators had the characteristic of evidence about which peer evaluators could take informed decisions, including by slightly reframing peer review to include signals from the field that the research was of value[2].

The Commission followed a two- step process to finalise its report; it firstly proposed a set of the kinds of indicators that could be used and then piloted those indicators by attempting to get research quality indications from two research units, the KNAW Meertens Institute, and RU Groningen's Research Institute for the Study of Culture (ICOG). This was undertaken by a consultant and was as much a feasibility study as an exercise in prototyping, undertaken in dialogue with the agents in the two research institutions. Part of this dialogue was also with the field, and after the interim report was published (KNAW, 2011b), a symposium was held at the Netherlands Institute for Advanced Studies in Wassennaar (25th November 2011). The mood at the day was

[2] One interviewee noted that some historians produced entries in museum catalogues that were highly respected within their field but outside the field were seen as sub-journalistic.

overwhelmingly positive, that the report was very good problems were identified (see also KNAW, 2012, p. 45). There was much enthusiasm for the recommendations that fields could choose their own indicators, but the (mid-session) discussion also highlighted that any drive towards standardisation would undermine that flexibility[3]. The discussions of the pilots revealed that there were many practical issues and one informed observer noted afterwards that the report marked the beginning of a decade-long process rather than the acceptance of an impact indicator set by the Dutch humanities research community.

The overall hierarchy in the interim and final reports was as with the SEP, to separate out scientific and social impact, then have three kinds of impact in each, publications, evidence of use and evidence of recognition (see Tabel 1 below for the Social Impact indicators).

Table 1 Research impact indicators included in the Commission of Humanities Research Quality Indicators.

[Table 1 goes about here]

The KNAW report was finally published in 2012 indeed the report itself noted that "It has indeed become clear that there are a number of problems that must be solved before it will be possible to further apply the system"[4]. Following the lack of a clear consensus about how to make a definitive statement about indicators that would ensure between-agent fairness, the report proposed in effectively carrying on the work of the group, with a new group supported by the *Regieorgaan* but more focused on harnessing expertise to deal with the challenges involved in agreeing appropriate indicators. Regarding societal

[3] One presenter who had been involved in the consultation reported that as soon as discipline boundaries were hardened then this would hurt multi-disciplinary, trans-disciplinary and inter-disciplinary research at the expense of (more conservative) mono-disciplinary research. Another speaker noted that the introduction of formal journal lists and rankings, even the Euripean Reference Index for the Humanities and Social Sciences, would have a similar uneven impact.

[4] Following the *Regieorgaan* recommendation, one of the innovations was that the report foresaw that the National Research Schools (formed from all active research groups in a particular field) would have a responsibility for helping to further define which indicators and indeed approaches were suitable for their respective discipline[4].

impact indicators, the Commission gave a statement that could potentially be interpreted as seeking to defer a difficult question, noting that ERiC and the National Valorisation Commission were developing impact indicators and therefore it made sense to have an extended discussion of these partners, also bringing in the KNAW councils for Social Sciences and Engineering own impact indicator experts. The Regieorgaan supported this plan and a successor committee was appointed on 29 June 2013, chaired by the Dean of Humanities of the University of Amsterdam, to develop a national implementation plan and report back on how to achieve that by the end of 2015 (*Regieorgaan*, 2013).

At the time of writing, that process is underway, but it is clear that despite a huge amount of preparatory work have been undertaken in the previous decade in understanding impact, and even with the relatively weak criterion of the Algra Commission that indicators were evidence and not measures, it has proven very difficult to develop effective impact indicators in the humanities. Indeed, the discussions revealed that even within those disciplines that had apparently been able to agree impact indicators (both scientific and societal) there was increasing resistance to the negative effects produced, not just in the humanities but also more generally. The Science in Transition group emerged from scientists concerned with the strongly negative consequences on the health of Dutch research as a whole as a result of detailed top-down steering (Dijstelbloem *et al.*, 2013). Critical questions in Parliament over perverse stimuli led the most recent Science White Paper (OCW, 2014) to place emphasis on moving away from scientific productivity measures to attempt instead to look at quality in evaluation activities and allow scientists the space to develop quality research. In summary, the indicator approach in the humanities in the Netherlands appears unable to have been able to provide an alternative mode of fairness that would serve as the basis of effective between-agent comparisons, even drawing heavily on peer review, and there are persistent issues with the credibility and scientific support of such systems.

## 6. Fairness and comparison in Dutch research impact indicators.

In this paper, I have sought to answer the question of whether indicators can provide an alternative mode of fairness in research impact evaluations making comparisons and

therefore demanding between-agent fairness. In the example, I presented a case study of as yet unsuccessful attempts to develop an indicator set that could measure or assist in evaluating humanities societal research impact. Yet, the case study need not be regarded as a failure – over the course of the Algra Commission, experts articulated a set of definitions of humanities' societal impact. This was based around the idea that excellent research would be undertaken and then taken up by various scientific and societal user groups: the quality of impact was the extent of user uptake in these two classes. And there was broad support amongst a range of constituencies for this approach, and the plausibility for using the Table 1 framework as the basis for evaluating that impact. At the same time, there was continual resistance from within the humanities community out of a deep-seated fear that whatever was being proposed would be *unfair* to humanities, and in particular present research reports in which their societal impact was an additional reason to regard them as less societally valuable that other subject areas.

But what is interesting is that where there is signs of consensus about the credibility about the approach in terms of providing cross-agent fairness is where other disciplines have faced the same problems. So for both social sciences and engineering sciences KNAW committees had explored possibilities and had concluded that flexibility in impact indicators was a precursor for fairness between the (cognate) disciplines under consideration (where one interviewee suggested it is easier to appreciate that their societal impacts, although different to ones' own, are still useful,, much generally more difficult between disciplines with fundamentally different views of how knowledge is created, *cf.* Olmos Peñuela *et al.*, 2013). A number of interviewees noted that once the issue of between-agent fairness was dealt with, then it is possible for a research system based on considerable flexibility and discretion to nevertheless function with a degree of credibility. The Algra report created a framework for defining impact, both scientific and societal, then subdivided that into publications, joint activities, and signals of user value, that were broadly applicable across disciplines and derived from a reasonably objective definition of research impact (as established through ERiC and Siampi (Spaapen & Van Drooge, 2011).

What arguably has made a difference here is that although there is a compulsion of those involved to participate (for the reward of additional humanities structural funding from

2015), no direct link has been created between university financing and performance against those measures. Indeed, the idea for indicators came out of an apparently sincere desire to raise the credibility of humanities with respect to other disciplines by addressing the tendency of humanities to claim that creating impact did not matter to them. This has allowed the debate presented here to focus on issues of fairness rather than worrying of the potential later unfairness arising from the way the system has been allowed to unfold. In the absence of baseline data it is hard to say whether Dutch humanities is indeed responding as Plasterk had hoped in setting up the committee, to become more open-looking and concerned with creating societal benefits. But clearly – and not purely as a result of a the dialogue in Algra – if we believe the negative picture portrayed in earlier commissions and articulated by our interviewees, there is a greater awareness of the issues of research generating impact. This is both in terms of it being an important policy goal, but also the fact that it need be built on high-quality fundamental research and it is in tune with the broader norms in the field, whatever they might be.

## 7. Concluding discussion – fairness, indicators and evaluation

The main research question in this paper is how can evaluation of research impact at the systems level – aiming to give between-agent fairness – can deal with the problem of the very different mechanisms by which different kinds of research produce their impact? The starting point for this paper was the apparent contradiction between the need for research evaluations to guarantee two kinds of fairness, between ensuring that research agents are fairly assessed against what they can reasonably be expected against them, and that there is fair treatment between the agents being evaluated. There was a fairly strong consensus between the policy and scholarly communities about what a good – fair set of indicators would involve. Policy-makers' concerns lay in allowing humanities to represent themselves fairly to other disciplines, and at the same time to encourage engaged and open behaviour as a wider system norm in order to raise the dynamism of the system. Conversely, generating returns on past scientific investments was not a particularly high priority for policy-makers, which reduced the pressures to highlight particular kinds of economic activities. For academics and their managers were primarily concerned that there was sufficient congruency of the administrative

proposals being made, and behaviours which they recognised as being good-practice in their field.

We can see here the distinct elements of for a research impact evaluation system in which credibility is created. Firstly, there was an overall framework for between-agent fairness, provided by adopting the same distinctions of scientific and societal excellence in terms of publications, activities and uptake evidence as used in other fields. Secondly, there was a framework for within-agent fairness provided through the case studies that showed that with a degree of sensitivity, discretion and judgement the approach could be sensibly operationalised. Thirdly, the credibility of the process was reinforced by an overlap of the aims and intentions in the system, creating something to demonstrate a characteristic of research, quality, rather than to count outputs or allocate resources. Fourthly, there was a use of peer review to provide accountability but also allow for discretion. There was a nesting of the two forms of fairness – the within-agent fairness allowed the process to be credible to researchers within the field; between-agent fairness was provided by a framework in which dissimilar agents were subject to similar but not identical processes (mediated through the SEP).

This suggests that the call by policy-makers for indicators of humanities research impact that are comparable and compatible with other disciplinary areas may have two kinds of effect. It may stimulate this nesting process, as these communities and their representative organisations attempt to articulate their own internally fair versions of what engagement is – and indeed we see particularly in the UK but also in other countries including Canada and Australia (CHASS, 2015; British Academy, 2010; 2014; IDEAS, 2014). Alternatively, it may serve to drive a wedge between these two firms of trust, by creating suspicion between that the indicators are not fair, either they are not relevant to the agents themselves or they are not describing that against which other agents are held to account. Which of the two eventualities arises is a function of a range of choices, whether policy-makers accept they will not get a definitive answer comparing research impact, other agents accept that any measurement system has problems and legitimating judgement and discretion, and agents within a field positively identifying how impact is produced from researchers who are accepted as being emblematical of good practice. This paper therefore echoes Molas-Gallart's (2015) call for policy-maker to be clear about what precisely they want to achieve with an

evaluation. More reflection is demanded of scholars on what kinds of research impact matters in their field, and how that messiness of impact generation legitimates a multi-disciplinary, judgement- and discretion-based system that ultimately values activities and outcomes which lie beyond the pale of their own scholarly norms.

## Acknowledgements

This paper draws on research undertaken within the project HERAVALUE (Measuring the public value of arts and humanities research) is financially supported by the HERA Joint Research Programme which is co-funded by AHRC, AKA, DASTI, ETF, FNR, FWF, HAZU, IRCHSS, MHEST, NWO, RANNIS, RCN, VR and The European Community FP7 2007-2013, under the Socio-economic Sciences and Humanities programme. The author would like to thank all the Dutch participants in interviews, focus groups and the project seminars for their input to the project. Any errors or omissions remain the responsibility of the author.

## Bibliography

ARC (2015) "Research impact: principles and framework" http://www.arc.gov.au/general/impact.htm 28th January 2015 (Accessed 3rd February 2015).

AWT (2007) A radiant future – policies for 'valorisation' of the humanities and social sciences. The Hague: The Advisory Council for Science & Technology. English Language Summary available here: http://www.awti.nl/upload/documents/publicaties/engels/a70_uk.pdf (Accessed 20th January 2015).

Bate J (2011) The public value of the humanities. London: Bloomsbury Academic.

Belfiore, E. (2013) "The" rhetoric of gloom" vs. the discourse of impact in the humanities: stuck in a deadlock?" In E. Belfiore, & A. Upchurch, (eds). Humanities in the Twenty-First Century: Beyond Utility and Markets. Palgrave Macmillan, 2013, pp. 17-43.

Benneworth, P. (2014) "Tracing how arts and humanities research translates, circulates and consolidates in society.. How have scholars been reacting to diverse impact and public value agendas?" Arts and Humanities in Higher Education 1474022214533888, first published on May 14, 2014 as doi:10.1177/1474022214533888.

Benneworth, P. S., Gulbrandsen, M., & Hazelkorn, E. (2016) The impact and future of arts and humanities research, London: Palgrave (forthcoming)

Blockmans (2007) "The underestimated humanities and social sciences" in Quality Assessment in Higher Education pp. 89-94 Available online at: http://www.portlandpress.com/pp/books/online/QAHEE/001/0089/0010089.pdf(Accessed 4th February 2015).

Bod, R. (2013) New History of the Humanities: The Search for Principles and Patterns from Antiquity to the Present, Oxford: Oxford University Press (tr. L. Richards).

Bosland,J. (2010) De waanzin rond Wilders: psychologie van de polarisatie van Nederland, Amsterdam: Balans.

Bowman, R. (2015) "REF guesstimate sums" in Times Higher Education, 12th February 2015, Available online via: http://www.timeshighereducation.co.uk/news/academic-estimates-real-cost-of-ref-exceeds-1bn/2018493.article (Accessed 15th Ferbary, 2015).

Bozeman, B., & Sarewitz, D. (2011). Public value mapping and science policy evaluation. Minerva, 49(1), 1-23.

British Academy (2010) Past Present and Future. The public Value of the Humanities & Social Sciences. The British Academy. [Accessed 11 September 2013], < http://www.britac.ac.uk/news/bulletin/BAPPF.pdf >.

British Academy (2014) Prospering wisely: how the humanities and social sciences enrich our lives, London: British Academy Available on-line at: http://www.britac.ac.uk/prosperingwisely/pub/pdf/prospering-wisely.pdf (accessed 30th October 2014).

Brewer, J. D. (2013) The public value of the social sciences", London: Bloomsbury

Brown, R. (2013). Everything for Sale?: The Marketisation of UK Higher Education. London: Routledge.

CHASS (2005) Measures of quality and impact in publically funded research in the humanities, arts and social sciences, CHASS Occasional Paper 2, Council for Humanities, Arts & Social Sciences: Canberra, Australia. Available on-line at http://www.chass.org.au/papers/pdf/PAP20051101JP.pdf (Accessed 4th February 2015).

CHASS (2013) Submission to: "Assessing the wider benefits arising from university-based research" Prepared on behalf of the Council for the Humanities, Arts and Social Sciences (CHASS): Canberra, Australia. Available on-line at http://www.chass.org.au/submissions/pdf/SUB20130816PM.pdf (Accessed 4th February 2015).

Chatelain-Ponroy, S., Mignot‐Gérard, S., Musselin, C. & Sponem, S. (2014) "The use of indicators in French universities, in W. Blockmans L.Engwall & D. Weaire (eds) Bibliometrics: Use and

Abuse in the Review of Research Performance Wenner-Gren International Series, volume 87, London, Portland Press, pp. 129-141.
Cohen, J. (2010) "Sustainable Humanities" Report from the Committee on the National Plan for the Future of the Humanities, Amsterdam, NL: University of Amsterdam Press. Available online at http://www.regiegeesteswetenschappen.nl/images/uploaded/92/editorial/id=344.pdf (Accessed 3rd February 2015).
Crossick, G. (2006) "Knowledge transfer without widgets: the challenge of the creative economy", Lecture to the Royal Society of Arts, Leeds, 31st May 2006.
Crossick, G. (2009) "So who now believes in the transfer of widgets?" paper presented to Knowledge Futures Conference, Goldsmiths College, London, 16th-17th October 2009.
Debackere, K. and Glänzel, W. (2004) Using a bibliometric approach to support research policy making: the case of the Flemish BOF-key. Scientometrics 59, 253–276.
Dijstelbloem, H. Huisman, F., Miedema, F., Mijnhardt, W. (2013) "Waarom wetenschap niet werkt zoals het moet en wat daar aan te doen is" Available online at: http://www.scienceintransition.nl/wp-content/uploads/2013/10/Sience-in-Transition-Position-paper-versie-2.pdf (Accessed 4th February 2015).
Donovan, C. (2007) The qualitative future for research evaluation, Science and Public Policy, 34(8), October 2007, pages 585–597, DOI: 10.3152/030234207X256538.
Eriksson, L. (2013). "The Performance-based Funding Model: Creating New Research Databases in Sweden and Norway". July 2013, Ariadne Issue 71 http://www.ariadne.ac.uk/issue71/eriksson
Gascoigne T and Metcalfe J (2005) Commercialisation of research activities in the humanities, arts and social sciences in Australia. CHASS Occasional Papers Nº1, Council for Humanities, Arts and Social Sciences, Australia, May.
Grimson, J. (2014) "Measuring impact: not everything that can be counted counts, and not everything that counts can be counted" Not everything that matters can be measured" in W. Blockmans L.Engwall & D. Weaire (eds) Bibliometrics: Use and Abuse in the Review of Research Performance Wenner-Gren International Series, volume 87, London, Portland Press, pp. 29-41.
HEFCE (2011) Decisions on evaluating research impact, HEFCE Guidance Note 2011.1, Bristol: HEFCE. Available online http://www.ref.ac.uk/media/ref/content/pub/decisionsonassessingresearchimpact/01_11.pdf (Accessed 30th January 2015).
HEFCE (2014) "Bibliometrics pilot exercise" Last updated 11th December 2014 (http://www.ref.ac.uk/about/background/bibliometrics/) Accessed 30th January 2015)
Holmwood, J. (2010) „Sociology‟s misfortune: disciplines, interdisciplinarity and the impact of audit culture‟ British Journal of Sociology 61: 639-58.
Holmwood J. (2011) "The impact of "impact" on UK social science Methodological innovation online 6 (1) pp. 13-17. Available online at http://www.methodologicalinnovations.org.uk/wp-content/uploads/2013/11/5.-Viewpoint-Holmwood-13-17-proofed.pdf
Huang, M. H., & Chang, Y. W. (2008). Characteristics of research output in social sciences and humanities: From a research evaluation perspective. Journal of the American Society for Information Science and Technology, 59(11), 1819-1828.
IDEAS (2015) The impact of humanities and social sciences research, Working Paper October 2014. Federation for the Social Sciences and humanities: Ottowa Canada. Available online at : http://www.ideas-idees.ca/sites/default/files/2014-10-03-impact-project-draft-report-english-version-final2.pdf (Accessed 4th February 2015).
Jasanoff, S. (2003). Technologies of humility: citizen participation in governing science. Minerva, 41(3), 223-244.
Kickert, W. (1995), Steering at a Distance: A New Paradigm of Public Governance in Dutch Higher Education. Governance, 8: 135–157

Kickert, W. (1995), Steering at a Distance: A New Paradigm of Public Governance in Dutch Higher Education. Governance, 8: 135–157

KNAW (2002) Vensters op de wereld: de studie van de zogenoemde klein letteren, Rapport van de Adviescommissie Kleine Letteren (Commissie Gerritsen) , Amsterdam, The Royal Netherlands Academy of Arts & Sciences. Available online at https://www.knaw.nl/nl/actueel/publicaties/vensters-op-de-wereld/@@download/pdf_file/20011106.pdf

KNAW (2005) Judging research on its merits An advisory report by the Council for the Humanities and the Social Sciences Council, AMSterdam, The Royal Netherlands Academy of Arts & Sciences, Available online at: https://www.knaw.nl/nl/actueel/publicaties/judging-research-on-its-merits/@@download/pdf_file/20051029.pdf Accessed 4th February 2015.

KNAW (2010) Standard Evaluation Protocol 2009-2015: protocol for research assessment in the Netherlands: revised June 2010, Amsterdam: The Royal Netherlands Academy of Arts & Sciences in association with VSNU and NWO. http://www.knaw.nl/Content/Internet_KNAW/publicaties/pdf/20091052.pdf

KNAW (2011a) kwaliteitsbeoordeling in de ontwerpende en construerende disciplines, https://www.knaw.nl/shared/resources/actueel/publicaties/pdf/20101065.pdf Amsterdam, The Royal Netherlands Academy of Arts & Sciences. Available online as (Accessed 4th February 2015).

KNAW (2011b) Quality indicators for research in the humanities Amsterdam, The Royal Netherlands Academy of Arts & Sciences. Available online as https://www.knaw.nl/shared/resources/actueel/publicaties/pdf/20111024.pdf (Accessed 4th February 2015).

KNAW (2012) Kwaliteit en relevantie in de geesteswetenschappen Naar een adequaat systeem voor de beoordeling van wetenschappelijk onderzoek, Amsterdam: The Royal Netherlands Academy of Arts & Sciences. https://www.knaw.nl/shared/resources/actueel/publicaties/pdf/20121018.pdf (Accessed 30th January 2015)

Martin, B. R. (2013) "The Research Excellence Framework and the 'impact agenda': are we creating a Frankenstein monster" Research Evaluation, 20(3), pp. 247–254

Molas-Gallart J (2012) Research governance and the role of evaluation: A comparative study. American Journal of Evaluation 33(4): 577–592.

Mryglod, O., Kenna, R., Holovatch, Y., & Berche, B. (2013a) Absolute and specific measures of research group excellence. Scientometrics, 95(1), 115-127.

Mryglod, O., Kenna, R., Holovatch, Y., & Berche, B. (2013b) Comparison of a citation-based indicator and peer review for absolute and specific measures of research-group excellence. Scientometrics, 97(3), 767-777.

Mryglod, O. Kenna, R. Holovatch, Y.· Berche, B. (2014) "Predicting results of the Research Excellence Framework using departmental h-Index" Scientometrics (forthcoming) via http://arxiv.org/pdf/1411.1996.pdf (Accessed 30th January 2015)

OCW (2009a) "instellingsbesluit Commissie Regieorgaan Geesteswetenschappen" Regeling nr. OWB/FO/109823, 2nd July, 2009. Available online at: http://wetten.overheid.nl/BWBR0026155/geldigheidsdatum_04-02-2015 (accessed 4th February 2015).

OCW (2009b) "Eerste extra middelen geesteswetenschappen verdeeld" Press release of the Dutch Ministry of Education, Culture & Sicence, dated 12th November 2009, available on-line at http://www.rijksoverheid.nl/nieuws/2009/11/13/eerste-extra-middelen-geesteswetenschappen-verdeeld.html (Accessed 3rd February 2015)

OCW (2014) Ministry of Education, Culture & Science: the Hague, NL. Available online at http://www.rijksoverheid.nl/bestanden/documenten-en-publicaties/rapporten/2014/11/25/wetenschapsvisie-2025-keuzes-voor-de-toekomst/wetenschapsvisie-2025-keuzes-voor-de-toekomst.pdf (Accessed 4th February 2015)

Olmos-Peñuela, J., Benneworth, P. & Castro-Martinez, E. (2013) "Are STEM from Mars and SSH from Venus? Challenging stereotypical perceptions of differential social usefulness of academic disciplines", Science and Public Policy first published online October 3, 2013 doi:10.1093/scipol/sct071.

Olmos-Peñuela, J., Castro-Martínez, E., & D'Este, P. (2014). Knowledge transfer activities in social sciences and humanities: Explaining the interactions of research groups with non-academic agents. Research Policy, 43(4), 696-706.

PA Consulting (2000) Better accountability for higher education, London: PA Consulting.

Otten, W. (2015) https://divinity.uchicago.edu/willemien-otten-0 Accessed 3rdFebruary 2015).

PA Consulting (2004) Better accountability revisited: review of accountability costs 2004, London: PA Consulting. Available online at http://dera.ioe.ac.uk/4985/1/rd06_04.pdf (Accessed 30th January 2015).

PA Consulting (2008) RAE 2008 Accountability Review, London: PA Consulting. Available on-line at: http://www.hefce.ac.uk/media/hefce/content/pubs/2009/rd0809/rd08_09.pdf (Accessed 30th January 2015).

Regieorgaan, Geesteswetenschappen (2010) Advies inzake Implementatie Duurzame Geesteswetenschappen, Available online at: http://www.regiegeesteswetenschappen.nl/images/uploaded/92/editorial/id=353.pdf (Accessed 4th February 2015).

Regieorgaan Geesteswetenschappen (2011) Tussenrapportage 2009-10inzake Monitoring Duurzame Geesteswetenschappen, Regieorgaan Geesteswetenschappen, Utrecht. http://www.regiegeesteswetenschappen.nl/p/26.html?article_id=10&m=20 (Accessed 29th January 2015).

Regieorgaan Geesteswetenschappen (2012) Tussenrapportage 2011 inzake Monitoring Duurzame Geesteswetenschappen, Regieorgaan Geesteswetenschappen, Utrecht. http://www.regiegeesteswetenschappen.nl/p/26.html?article_id=12&m=20 (Accessed 29th January 2015).

Regieorgaan Geesteswetenschappen (2013) Tussenrapportage 2012 inzake Monitoring Duurzame Geesteswetenschappen, Regieorgaan Geesteswetenschappen, Utrecht. http://www.regiegeesteswetenschappen.nl/p/26.html?article_id=17&m=20 (Accessed 29th January 2015).

Regieorgaan Geesteswetenschappen (2014) Tussenrapportage 2013 inzake Monitoring Duurzame Geesteswetenschappen, Regieorgaan Geesteswetenschappen, Utrecht. http://www.regiegeesteswetenschappen.nl/p/26.html?article_id=20&m=20 (Accessed 29th January 2015).

Pontille, D., & Torny, D. (2010). The controversial policies of journal ratings: Evaluating social sciences and humanities. Research Evaluation, 19(5), 347-360.

RCUK (u.d) Pathways to impact

Sarewitz, D. & Pielke, R. A. (2007) The neglected heart of science policy: reconciling supply of and demand for science. Environmental Science and Policy, 10 (1): 5‑16.

Small, H. (2014) The value of the humanities, Oxford: Oxford University Press.

Spaapen, J., & van Drooge, L. (2011). Introducing 'productive interactions' in social impact assessment. Research Evaluation, 20(3), 211-218.

Staal, F., & S.A. Bonebakker, E. Gene Smith and H.J. Verkuyl, (1991) Baby Krishna: rapport van de Adviescommissie Kleine Letteren, The Hague.

Trolley, J., & O'Neill, J. (1999). The evolution of citation indexing—From computer printout to the web of science. In History & Heritage of Science Information Systems, 1998 Conference Proceedings (pp. 124-126). Available online at: http://webdoc.sub.gwdg.de/ebook/s/2001/chf/www.chemheritage.org/historicalservices/asis_documents/asisbook.pdf#page=136 (Accessed 3rd February 2015).

Van der Meulen, B., J. R. & Rip, A. (1995), Indicatoren en indicaties voor de beoordeling van maatschappelijke kwaliteit van onderzoek, Eindrapport, Commissie Overleg Sectorraden (Centrum voor Studies van Wetenschap, Technologie en Samenleving, Universiteit Twente, Enschede).

Van der Meulen, B., J. R. & Rip, A. (2000) Evaluation of societal quality of public sector research in the Netherlands Research Evaluation, 8 (1) pp. 11–25

Van Raan, A. F. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. Scientometrics, 62(1), 133-143.

Vonhoff, H. (1995) Men weegt kaneel bij 't lood, Final report of the Commission on the Future of Humanities, Utrecht.

Worton, M. (2006). Of Models and Metrics: The UK Debate on Assessing Humanities Research. Prague Peer Review 2006. Peer Review: Its Present and Future State, Prague, Czech Republic. Available online (Accessed 3rd February 2015). http://discovery.ucl.ac.uk/14322/

## Appendix 1 Tables and figures

Table 1 Research impact indicators included in the Commission of Humanities Research Quality Indicators.

| | | |
|---|---|---|
| 4. Civil-society publications | Articles in specialist publications (not being primarily scientific/scholarly journals) | • List<br>• Selection of key publications |
| | Monographs for non-scientists/scholars and interested individuals | • List<br>• Selection of key publications |
| | Chapters in books for non-scientists/ scholars and interested individuals | • List<br>• Selection of key publications |
| | Other civil-society output, for example collections for non-scientists/scholars and interested individuals, editorships of specialist publications, handbooks, dictionaries, editions of texts, databases, software, exhibitions, catalogues, translations, advisory reports on policy | Quantitative and/or qualitative information to be requested as determined according to the context |
| 5. Civil-society use of research output | Projects carried out in collaboration with civil-society actors | Simple statement with dates (years) |
| | Contract research | Simple statement with dates (years) |
| | Demonstrable civil-society effects of research | Simple statement with dates (years) |
| | Other types of civil-society use, for example reviews, citations in policy reports, use of publications, media attention, books sold/loaned | Quantitative and/or qualitative information to be requested as determined according to the context |
| 6. Evidence of civil society recognition | Civil-society prizes | Simple statement with dates (years) |
| | Other evidence of civil-society recognition, for example civil-society appointments, invitations to give lectures, invitations for media appearances, advisory positions/membership of advisory committees | Quantitative and/or qualitative information to be requested as determined according to the context |

Source: KNAW, 2012, p. 58.

The Center for Higher Education Policy Studies (CHEPS) is a research institute (WHW, Article 9.20) located in the Faculty of Behavioural and Management Scienes within the University of Twente, a public university established by the Dutch government in 1961. CHEPS is a specialized higher education policy centre that combines basic and applied research with education, training and consultancy activities.

http://www.utwente.nl/bms/cheps/