Bibliometrics in the research assessment exercise reports of Finnish universities and the relevant international perspectives

Li Wang, Pia Vuolanto & Reetta Muhonen

Research Centre for Knowledge, Science, Technology and Innovation Studies
University of Tampere

Li Wang, Pia Vuolanto & Reetta Muhonen

Bibliometrics in the research assessment exercise reports of Finnish universities and the relevant international perspectives

Research Centre for Knowledge, Science, Technology and Innovation Studies
School of Social Sciences and Humanities
University of Tampere

Table of Contents

Tiivistelmä

Bibliometriikka Suomen yliopistojen tutkimusarvioinneissa sekä Iso-Britannian ja Alankomaiden tutkimusarviointien käytännöissä

Tämä raportti käsittelee bibliometriikan käyttöä Suomen yliopistojen tutkimusarvioinneissa. Selvityksessä käytiin läpi kuusitoista 2000-luvulla toteutettua yliopiston tutkimusarviointia neljässätoista yliopistossa. Suurimmassa osassa arviointeja painottuivat arviointipaneelien antamat laadulliset arvioinnit eikä niissä hyödynnetty bibliometriikkaa lainkaan. Osassa arvioinniesta bibliometrista laskentaa käytettiin informoitaessa arviointipaneeleja, mutta analyysien tuloksia ei esitetty arviointiraportissa. Tässä raportissa tarkastellaan sellaisia tutkimuksenarvioinnin raportteja, joissa bibliometriikka on mukana. Näitä raportteja on tehdyistä 16 tutkimuksenarvioinnista kuusi. Tämä raportti esittelee bibliometrisiä menetelmiä näissä kuudessa tutkimusarvioinnissa. Lisäksi raportissa tehdään katsaus Iso-Britannian ja Alankomaiden tutkimusarvioinnin käytäntöihin.

Selvityksessä etsittiin vastauksia kysymyksiin: Millaisia bibliometrisiä menetelmiä ja indikaattoreita yliopistojen tutkimuksen arvioinneissa on käytetty? Miten bibliometristen aineistojen hankinta ja analyysi on toteutettu yliopistojen tutkimusarvioinneissa?

Bibliometristen menetelmien käyttöä valmisteltiin määrittelemällä, mitä yliopiston yksikköjä tai henkilöstöryhmiä sisällytetään bibliometrisiin analyyseihin. Tutkimushenkilökunnan määrittelyssä käytettiin kahta tapaa: tutkimusarviointiin sisällytettiin määrättynä laskenta-ajankohtana yliopiston palkkalistoilla oleva tutkimushenkilöstö tai koko arviointijaksolla yliopiston työsuhteessa olevat henkilöt. Arviointiyksiköt valittiin joko siten, että käytettiin olemassaolevia hallinnollisesti määrittyneitä yksiköitä (oppiaineita tai tieteenalayksiköitä) tai niin, että tutkijat määrittelivät itse arvioinnin kohteena olevan tutkimusyksikön rajat. Yhdessä yliopistossa arviointi kohdistui tieteenaloihin.

Bibliometrisen aineiston kerääminen toteutettiin ulkopuolisen asiantuntijan tai yliopistokirjaston toimesta. Joissakin tapauksissa yliopistokirjasto valmisteli aineistoa ulkopuoliselle asiantuntijalle. Neljän yliopiston bibliometrinen analyysi tehtiin Leidenin yliopiston Tieteen- ja teknologiantutkimuksen keskuksessa (CWTS) ja kahden yliopiston analyysin teki yksittäinen henkilö. Bibliometrisiin analyyseihin sisällytettiin tieteelliset artikkelit ja konferenssijulkaisut sekä muutamissa tutkimusarvioinneissa myös kirja-arviot ja pääkirjoitukset. Tutkimusarvioinneissa käytettiin seuraavia tietokantoja: ISI Web of Science, Scopus, Tuhat (Helsingin yliopiston tutkimustietokanta) ja Tutkii (Oulun yliopiston tutkimustietokanta).

Viidessä arviointiraportissa noudatettiin Leidenin yliopiston Tieteen- ja teknologiantutkimuksen keskuksen käytäntöjä, jolloin tulosmittareina käytettiin julkaisujen kokonaismäärää ja ositettuja julkaisujen määriä. Vaikuttavuusmittareina käytettiin julkaisun

saamien viittausten keskiarvoa ja tieteenalan, julkaisutyypin ja julkaisuvuoden suhteen normalisoitua viittausten keskiarvoa. Viittaukset suhteutettiin maailmassa eniten viittauksia keräävään kymmenen prosentin joukkoon ja lehden viittausmääriin.

Kolmessa arviointiraportissa sovellettiin muita tapoja bibliometriseen analyysiin. Tällöin tulosmittareina käytettiin julkaisujen kokonaismäärää ja vaikuttavuusmittareina viittausten määrää, normalisoitua viittausten keskiarvoa, h-indeksiä ja summaindeksiä. Näissä arviointiraporteissa hyödynnettiin lisäksi opetus- ja kulttuuriministeriön julkaisutyyppiluokittelua, julkaisufoorumin tasoluokitusta, Norjan julkaisufoorumin tasoluokitusta ja Australian julkaisuluokituksia. Yhdessä arviointiraportissa käytettiin tieteenalakohtaista tarkastelua, jossa julkaisut luokiteltiin yliopiston paikkakunnan mukaan.

Selvityksessä havaittiin, että tutkimusarvioinneissa bibliometrisen laskennan yksityiskohdat oli toteutettu erilaisilla tavoilla, jotka vaikuttivat analyysien tuloksiin. Näiden erilaisten käytäntöjen vuoksi yliopistojen bibliometristen analyysien tuloksia ei voida vertailla keskenään. Suomalaisten yliopistojen tutkimusarviointiraportteja tarkasteltaessa tulisi huomioida, miten arviointiyksiköt on määritelty, mitkä julkaisutyypit on sisällytetty analyysiin, miltä aikaväliltä julkaisujen saamat viittaukset huomioidaan, käytetäänkö bibliometrisissä analyyseissä normalisointia, ositetaanko julkaisut kirjoittajien määrän mukaan ja miten itseviittauksiin suhtaudutaan.

Selvityksessä tehtiin vertailua kahteen kansainväliseen tutkimuksenarvioinnin edelläkävijämaahan. Iso-Britanniassa tutkimusarvioinnit tehdään keskitetysti kansallisella tasolla (Research Excellence Framework) ja arviointeja käytetään tutkimusrahoituksen jakamisen tukena. Tutkimusarvioinnit keskittyvät korkeatasoisimpaan tutkimukseen ja ottavat huomioon sosiaalisen tasa-arvon ja yliopistojen erilaiset työntekijäprofiilit. Iso-Britannian tutkimuksen arvioinneissa yliopistot velvoitetaan toimittamaan yksiköidensä materiaali arviointia varten. Arviointiyksikkönä ovat 36 tieteenalaa ja jokainen yliopisto toimittaa materiaalinsa tieteenalaprofiilinsa mukaisesti. Vertaisarviointipaneelia pidetään keskeisimpänä arvioinnin muotona ja bibliometriikkaa käytetään vain rajoitetusti. Kolme keskeisintä arviointikriteeriä ovat: tutkimustuotokset, vaikuttavuus ja tutkimusympäristö. Vuoden 2014 lopulla päättyvässä arvioinnissa vain 11 arviointipaneelin työskentelyssä käytetään bibliometriikkaa arvioinnin tukena. Näissä paneeleissa viittausaineisto perustuu Scopus-tietokantaan. Indikaattoreina käytetään viittausmääriä ja viittaukset suhteutetaan alan kärkijulkaisujen saamiin viittauksiin. Bibliometrisissä analyyseissä ei käytetä normalisointia eikä osittamista.

Myöskään Alankomaissa bibliometriikkaa ei pääsääntöisesti käytetä tutkimuksen arvioinnissa, vaan tärkeimmässä osassa on laadullinen vertaisarviointipaneelin arviointi. Tutkimusarvioinnit tehdään yliopistojen aloitteesta ja niiden omilla aikatauluilla, mutta kansallisella tasolla on valmisteltu ohjeet arvioinnin keskeisistä kriteereistä (Standard Evaluation Protocol). Arviointiyksiköiden valinnassa kiinnitetään huomiota arviointiyksikön kokoon. Yksiköllä on oltava vähintään kymmenen tutkimustyövuoden panos ja oma tutkimusstrategia. Lisäksi

yksikön perustamisesta on pitänyt kulua vähintään kolme vuotta. Yksikön on myös pystyttävä nimeämään vertailuyksikkö ulkomailta. Yksiköitä arvioidaan kolmella tasolla: tutkimuksen laatu, tutkimuksen yhteiskunnallinen vaikuttavuus ja yksikön elinvoimaisuus. Arviointiyksiköt toimittavat arviointia varten itsearviointiraportin, jossa ilmoitetaan tutkimushenkilökunnan määrä, julkaisuvolyymi, rahoituslähteet sekä tohtoriopiskelijoiden määrä.

Iso-Britannian ja Alankomaiden tutkimuksenarvioinnit havainnollistavat kahta erilaista tapaa käyttää tutkimuksenarviointia. Iso-Britanniassa arviointi kiinnittyy tutkimusrahoituksen allokointiin. Iso-Britannian ja Alankomaiden käytäntöjen avulla voidaan oppia siitä, miten tutkimusarviointeja ohjeistetaan kansallisella tasolla. Bibliometriikan vähäisen roolin vuoksi niitä voidaan kuitenkin käyttää vain rajoitetusti pohdittaessa suomalaisten yliopistojen tutkimusarvioinnin käytäntöjä.

Acknowledgements

Executive Summary

The study considered the methods of bibliometric analyses used in the research assessment exercise reports of Finnish universities, with six reports as the focus of analysis. The report on this research gives an overview of the process of implementation of research assessments and the differences in the methods used in these studies. The two stages of the implementation process are subjected to scrutiny: the preparatory stage (before performance of bibliometric analyses) and the analysis stage.

In the first stage, the reports are compared in terms of how practical matters are organised and managed before the bibliometric analyses are carried out: which staff are being evaluated and how the research unit is defined, what types of publication are considered and how the aggregates of the publications are formed, and who is recruited to perform the bibliometric analyses. This report shows the diversity of the practices employed across the universities. Research-active staff are defined to be those employed on the specified survey date or, alternatively, those who have been employed during the years that the evaluation targets. The choice of definition here determines also which publications are evaluated with bibliometric measures. The research unit may refer to a department or an institute, to a research community that is formed by the researchers themselves via a bottom-up approach, or to a discipline. As for resource recruitment, either the universities solicit the research services of CWTS, Leiden University or an external expert, or they utilise their own library and research staff resources. The types of publication and the indicators involved differ greatly between these two, quite different situations.

In its second stage, the study focused on the bibliometric indicators and compared the definitions and operationalisation of these indicators across the selected reports. Special attention is given to the six reports containing systematic bibliometric analyses. Often the definitions of the indicators are identical or similar from one report to the next, because definitions usually are developed by drawing upon well-established scholarship in the field of bibliometric studies. However, how the calculations for these indicators are actually realised in the research procedure differs greatly: for instance, in whether or not fractionalisation is applied, how normalisation is handled, and what document types are selected for calculation of citation counts. Together, all these differences render it impossible to compare numeric results from different reports. Accordingly, the aim of this report is to illuminate matters for the benefit of Finnish research assessment.

Furthermore, to broaden the perspective, international cases (involving the UK and the Netherlands) are reported on, with the emphasis being on their assessment frameworks and criteria. In the case of the UK, research assessment is organised in a centralised manner for the universities at national level to support the decision-making in funding allocation. This evaluation focuses on the excellence of the research done at the universities, with social equality and staff diversity taken into consideration. Clear contrast to the UK case is provided by the research assessments in the

Netherlands, organised by the universities or units on their own initiative in line with the protocol developed by national organisations. The two cases are presented and discussed for the insights they afford with respect to research assessment in Finland.

1. Introduction

This research concerns the research assessment exercises among Finnish universities, which have become widespread practice in the past few decades. One reason behind this is that the Finnish University Act obligates universities to evaluate their activities regularly. The term 'research assessment exercise' (or 'RAE' for short) generally refers to a research-performance measurement project usually conducted by the university for self-evaluation purposes. By means of qualitative or/and quantitative studies, the research performance of research-active staff at the university is assessed, usually with research output and impact as the two most important dimensions.

Bibliometrics is the commonly used quantitative method applied in RAE on global scale and also in Finland. It involves mathematical and statistical analysis of scientific output employed to evaluate research performance (Puuska, 2014). It can be carried out at various levels, from that of an individual piece of work or researcher to institutions, entire disciplines, or regions and countries (ibid.). In universities' RAE practice, bibliometric studies typically take the research unit as the unit of evaluation. In most Finnish universities, RAE has become a sustained routine of research administration and the exercises are performed every few years. Drawing upon databases such as WoS, Scopus, and the university library's database, the work examines individual research units' number of publications and citation counts – together with the university's as a whole – for what they reveal of the productivity of publishing activity and the resulting impact in scientific communities.

The RAE reports bear witness to the complexity of bibliometric studies. Foremost, the selection of database influences the results obtained as to publication output and citation rate since different databases have different coverage of journals and other publication forums. Furthermore, divergent choices of bibliometric indicators result in multiple ways of presenting the research quality. Moreover, even if seemingly identical indicators are used, differences in the ways of implementing them may lead to confusion. Likewise, in the implementation of an RAE project, various practical details may all affect the evaluation results.

To address this complexity, the research reported upon here was undertaken, with the objectives of studying the bibliometric methods used in the RAEs of Finnish universities and operationalisation-related details such as the definition of the unit and the process of implementation. The project, commissioned by the Ministry of Education and Culture ('the Ministry' hereinafter) came into existence partially in response to the lack of a basis for comparison of the various RAE practices of Finnish universities, but the main intention with the report is to uncover the underlying factors that could lead to differences in the evaluation results. This is approached by examining and comparing the methods of the bibliometric studies in Finnish RAEs.

Simultaneously it is necessary to point out that this research is not intended to identify or judge the flaws or strengths of the universities' RAE practice and bibliometric studies. Neither could this research offer a ranking of Finnish universities in terms of research performance. Overall, one cannot judge the universities' performance on the basis of this study. Nor will we provide practical suggestions for application in policymaking for research management. Instead, what this research focuses on when conceptualising the differences in bibliometric studies is how they arise, for purposes of facilitating understanding of the complexity of bibliometrics. It is not aimed at guessing underlying motivations and reasons or judging whether any given RAE is 'good' or 'bad'.

This research report is organised thus: First, the method according to which the study was conducted is explained. After this, the research findings are presented, with an emphasis on the organisation of the bibliometric studies and the bibliometric indicators. What follows is the two cases of the research assessments in the UK and the Netherlands. In the end, the report is concluded with suggestive final remarks.

2. Methodology

The study, using as its material 16 RAE reports, from 14 universities, was carried out in two stages. In Stage 1, all 16 RAE reports were examined, and those incorporating bibliometrics were taken as material for further analysis. This analysis revealed that six reports contained systematic bibliometric analysis involving all or almost all of the institution's research units. The remaining 10 reports focused predominantly on the qualitative assessment and offered only very limited information pertaining to bibliometrics. Typically these qualitative assessment reports mention the total number of publications and citation counts in the comments from panel review made on some of the units. Alternatively, the total number of publications of the university as a whole and its units are presented in 1–2 tables or charts at most. These reports do not address practical details of how the bibliometric study was done. In Stage 2, the analysis focused on only the six reports whose presentation was sufficient to allow sophisticated analysis of the bibliometric measures and indicators. The main findings described in this report are based mostly on the work in Stage 2. The research material is presented in Table 1.

However, it is worth emphasising that this does not mean only those six universities and not the rest conducted bibliometric study. On the contrary, almost all of the RAE reports indicate that bibliometric measurements were required as part of the material presented to the panels as the basis for qualitative review, though the results of bibliometric study were not published in some of these final reports. Moreover, the analysis of all reports revealed that qualitative assessment was used as the main tool of RAE. In the 10 reports giving more limited information about bibliometric data, qualitative assessment was emphasised, and qualitative assessment by

review panels had a strong role also in the six reports wherein the bibliometric data were presented in full. In those six reports, bibliometric study not only supports qualitative review as part of the material but is also presented in the final reports as an important element contributing to the results of the qualitative review[1].

Table 1: Overview of the research material

	Report[2]	Studied in Stage 2
1	Aalto University, for 2003–2007 (Krause, Saaristo, Lehtovaara, Sivenius, & Pasanen, 2009; Sandström, 2009)	×
2	University of Helsinki, for 1999–2005 (Haila, Holm, & Niemelä, 2005)	
3	University of Helsinki, for 2005–2010 (Saari & Moilanen, 2012)	×
4	University of Tampere, for 1999–2003 (Sahlander & Soini, 2005)	
5	Tampere University of Technology, for 2005–2010 ('TUT Research Assessment Exercise 2011: Bibliometric report 2005-2010,' 2011)	×
6	University of Jyväskylä, for 2005–2009 (Folea, 2011)	
7	University of Turku, for 2008-2009 (Laine, 2010)	×
8	University of Eastern Finland, for 2010–2012 (Liikanen et al., 2014)	
9	University of Kuopio, for 2000–2006 (Pellinen, Liikanen, & Kalliokoski, 2008)	
10	University of Joensuu, for 2000–2006, in 2008 ('Evaluation of research', n.d.)	
11	University of Oulu, for 2001–2006 (Ruskoaho & Jaako, 2008)	
12	University of Oulu, for 2007–2012 (Eskelinen & Ryyppö, 2014)	×
13	University of Lapland, for 2006–2012 ('International evaluation of strategic research at the University of Lapland', 2013)	
14	Lappeenranta University of Technology, in 2012 ('Research assessment exercise (RAE)', n.d.)	
15	University of Vaasa, for 2005–2009 ('Evaluation of research activities 2005–2009', 2010)	
16	Hanken School of Economics, for 2006–2011 ('Hanken School of Economics Evaluation of Research 2012', 2012)	×

In Stage 1, notes were taken during the read-through to summarise the bibliometric methods referred to in the RAE reports. In Stage 2, qualitative content analysis (Krippendorff, 1980/2012) – a well-established method in social sciences – of six RAE reports was performed. That analysis focused on two elements: the implementation of the bibliometric study and the concepts and operationalisation details associated with the bibliometric indicators. Analysis of the former involved the questions 'who?' (who was involved in organising and carrying out the studies),

[1] With the exception of UoT, where only the bibliometric analyses were presented in the report. The bibliometric questionnaire was complemented by a questionnaire for some units (see Laine, 2010, p. 78–79). Qualitative reviews were being planned at the time the report was issued (*ibid.*, p. 10).

[2] For convenience of presentation, when the six reports for Stage 2 are mentioned in the following text, they will not be referenced in full. The reports are referred to by the abbreviations of the six universities' names. For instance, the report of the University of Helsinki for 2005–2010 (Saari & Moilanen, 2012) is 'UoH' for short.

'what?' (what resources, such as databases and publication lists, were used as input to the projects and what output was delivered), 'when?' (the time window of the bibliometric studies), and 'how?' (which indicators were used and what the focus of study was – output or impact). For the second element, the analysis of the concepts underlying the bibliometric indicators included addressing the questions of how they were defined and, moreover, how they were operationalised. In terms of practice, the analysis was done in a read-through of the reports on paper and simultaneous marking of the relevant ones and noting of the categories, after which these findings were compared in MS Excel.

3. Findings

3.1 Preparatory stage of the bibliometric studies

In very broad terms, the bibliometric studies are carried out in three main stages: definition and formation of the unit, compilation of evaluation material, and the performance of the analysis. This section describes these three stages with the effects on the presented research performance taken into consideration.

3.1.1 Definition of the research unit and staff

For the most part, the first stage started with specifying which personnel are research-active staff, usually researchers including doctoral students, senior researchers, and professors. In some reports but not all, the exact numbers of researchers considered in the assessment were reported on, along with the percentage of the selected researchers. Organisation size, number of research staff, and the proportion of the personnel who are senior scientists all influence the publication output greatly. Therefore, it is of utmost importance to pay attention to which researchers are included in the RAE.

In general, there are two alternative ways of defining research staff. The first considers those researchers employed by the university on a census date. In the other, those included as research staff are the persons who were employed by the university during the period that the evaluation targets. This makes a difference for the collection of publications for evaluation: with the former approach, it is possible for some of the publications included to have been published before the relevant researcher was hired by the university. Likewise, it can be inferred that the publications of any employees who left the university for other employment before the census date are not included, even if they were published during employment with the university. In

the latter approach it also can result in inaccurate representation, since it is rather common that publications come out after a researcher's employment ends.

Furthermore, staff are divided into units for assessment. Units are defined in three distinct ways across the six universities (see Table 2). One way is to ask the researchers to form 'research communities' on their own by applying a bottom-up approach; UoH and UoO are example cases. However, a possible source of confusion is how to treat the output of those who were temporarily employed or affiliated with the university but no longer employed there at the time of research communities being formed. To which units should they be deemed to belong? In general, this is not explained well in the reports. Another way is to organise the units in accordance with the pre-existing organisational structure of the university (in some cases, multiple units can be combined to form a single one for evaluation). The UoT procedure for defining research staff and unit differed from the approach used for the other reports. The UoT analysis was carried out on the basis of discipline. Apparently, the analysis included publications of all staff in the databases used who fell under UoT jurisdiction in certain discipline, but no procedures for defining staff were presented. The publications that included the name of the city of the university were included and divided across 47 disciplines.

Table 2: Definition of units and the coverage of evaluation

Report	Definition of unit	Coverage
Aalto	The unit is a department, institute, or equivalent entity.	All 46 units were assessed. The staff employed on the census date were included.
UoH	Research community is the unit.	136 units were examined, in total. The staff were defined as those who had been employed by the university in 2005–2010.
TUT	The unit is a department or research centre.	23 units, from five faculties, were evaluated. The staff employed on the census date were included.
UoO	Research community is the unit.	There were 49 units in all. The staff were those employed on the census date and those affiliated with the university in 2007–2012 (p. 11).
HSE	The unit is defined in terms of 'Subject Research Priorities', which refers to areas within the departments.	There were 10 areas, from five departments, in total. Nine units were included in the bibliometric study, and one was not. There is no definition of research-active staff.
UoT	Disciplines are the units of analysis. The disciplines at UoT are compared to corresponding disciplines at other Finnish universities. The analysis was not tied to any administrative units, such as faculties or department; instead, it was done by discipline.	There were 47 disciplines analysed. The disciplines were selected to cover the disciplines of the two universities[3] as a whole.

3.1.2 Gathering of the publication material

In the six RAE reports, the list of publications to be assessed is either generated by the university's library or extracted by a hired expert or agency such as CWTS, after which the university verifies it. The quality of the validation is important, because affiliation information is not always required for all publication forums and authors at more than one university may have the same name.

Another factor affecting the results of the bibliometric studies is the publication types included in the material base. Journal articles are included in the bibliometric studies. For conference proceedings, there are no commonly observed conventions. The reports did not

[3] The report analysed the University of Turku and Turku School of Economics, which were merged into one university, UoT.

necessarily explain the ambiguity arising with reference to WoS capabilities, even between reports generated by one institute, CWTS. For example, one report by CWTS states that conference proceedings are not covered by WoS (Saari & Moilanen, 2012, p. 177), while others, also by CWTS, actually include these. One factor is that, as is mentioned in Aalto's report (Sandström, 2009, p. 11), WoS has been expanded since 2008 and conference proceedings have become incorporated into it. Furthermore, the report of UoO (Eskelinen & Ryyppö, 2014, p. 17) informs the reader that CWTS can conduct a separate analysis targeting conference proceedings, based on the WoS Conference Proceedings Citation Index. Therefore, it cannot be taken for granted that all WoS-based studies referred to in the RAE reports include the same types of publications (see Table 3).

Table 3: Corpus of material

Report	Publication material evaluated
Aalto	Material was collected from ISI WoS on the basis of the list of staff names. Articles and proceedings papers are included.
UoH	The library exported the registered publication list for CWTS. Only journal articles are included in the study carried out by CWTS (p. 176). All publication categories defined by the Ministry (as in Appendix 2) except G1–G5 theses were studied by the library.
TUT	The publications were extracted by CWTS from WoS, then verified by the university. Articles, reviews, and proceedings papers were included.
UoO	The library exported the publications in categories A, B, and C in accordance with the Ministry's classification and then asked the researchers to verify the records. The list formed the foundation for the library's study and was also sent to CWTS. Only articles and reviews in journals were included in the study by CWTS. Conference proceedings were analysed for six units separately.
HSE	Working from the list of research staff's names, CWTS extracted the list of publications from WoS, which list was then verified by the library. There is no further specification of which document types the publications encompass.
UoT	The data were collected from Essential Science Indicators and WoS (the Science Citation Index Expanded, Social Sciences Citation Index, and Arts & Humanities Citation Index, including the Conference Proceedings Citation Index) on the basis of the name of the university's city. Publication was included only if the name of the city that is the home of the university appeared in the affiliation information in the publication.

3.1.3 Recruitment for conducting the bibliometric studies

This section of the report focuses on the database and personnel resources utilised for performance of the bibliometric studies as summarised in Table 4. Of the six universities, four solicited the services of CWTS: UoH, TUT, UoO, and HSE. Aalto's bibliometric study was performed by an external expert – Ulf Sandström – from Linköping University, Sweden. In the case of UoT, the statistical analysis was carried out by Päivi Laine, an internal expert with the university.

Furthermore, to complement the studies of CWTS, UoH and UoO conducted bibliometric analysis for some disciplines on the basis of material from an in-house information system hosted in the university library. This was done for two reasons: the coverage of WoS was seen as being too limited for appropriate evaluation of the performance of these disciplines, or the total number of publications of some units did not reach the threshold required by CWTS for statistical reliability reasons. Though these units may be included in the analysis performed by CWTS, they are not necessarily studied by means of the standard procedure applied for examination of other units.

Table 4: Database and personnel resources for the performance of bibliometric studies

Report	Performed by	Database	Sources of citations
Aalto	External expert from Linköping University	WoS	Articles, letters, proceeding papers, and reviews
UoH	CWTS (for all 136 units)	WoS	Articles, letters, and reviews (p. 177)
	Helsinki University Library (66 units)	TUHAT RIS	Study of impact not in terms of citations but via comparison of the publications with the Norwegian, Australian, and ERIH (2007–2008) journal ranking list; Finnish publication forum ranking list; and Norwegian publisher list
TUT	CWTS	WoS	Not mentioned
UoO	CWTS (all 49 units)	WoS and its Conference Proceedings Citation Index	Not mentioned
	University library (17 units)	The university database Tutkii and Scopus	Impact measurement that varies with the discipline, including calculation of citation counts from Scopus and comparison with Finnish Publication Forum ranking list
HSE	CWTS	WoS	Articles, letters, and reviews
UoT	Internal expert of UoT	WoS	Articles and conference proceedings

3.2 Bibliometric indicators

Generally, there are four types of bibliometric indicators involved in the six RAE reports: output (number of publications), impact of publications (citation counts), impact of the journals (their citation rate), and collaboration (frequency of co-authorship and number of authors). The first three are more commonly utilised in the six reports. This section gives an overview of how these indicators are defined and operationalised. The first of its two subsections looks at five universities' reports in which either the research services of CWTS were solicited or the method

used followed CWTS standards closely. On account of the similarities between these studies, they are easier to compare in their selection of indicators. The second (the 'non-CWTS) subsection presents the three RAE reports that, in addition to CWTS data or otherwise, use a bibliometric approach that deviates greatly from CWTS standards. The indicators used in these three reports are explained below.

3.2.1 Indicators adopted from CWTS

Of the five universities that adopted indicators from CWTS, four– UoH, UoO, TUT, and HSE – had their bibliometric analysis conducted by CWTS. Aalto's bibliometric analysis was performed by external expert Ulf Sandström instead of CWTS. The practical details of the statistics implementation for some of the indicators used by Sandström diverged from CWTS's to some extent. The selection of indicators for these five reports is presented in Table 5. While differences readily arise when the bibliometric studies at issue are performed by different experts or organisations, the work being carried out by the same organisation (CWTS) did not always guarantee consistency of results. Variation in statistics operations contributed to this inconsistency.

In the five reports' analyses, output and impact were the two common foci. For output, the measurement involved two indicators: P (the total number of publications) and Pf (the number of fractionalised papers). When a paper has authors from different universities (e.g., five authors, from five universities), it can either be counted as '1' for each university (i.e., a whole paper for each) or be fractionalised, with each university able to count it as 0.2 papers (with fractional counting). Some of the five universities chose whole-number counts, some fractions, and some both. This resulted in differences in the indicator design.

For impact, four indicators are used by all five universities: MCS (mean citation score), MNCS (mean normalised citation score), THCP 10 (top 10% in citation impact), and MNJS (mean normalised journal score). The first three of these represent the citation impact of the publications, and the last one indicates the citation impact of the journals in which the works appear. Though the definitions are generally the same across the corpus, two factors in the operation have potential to lead to divergence in the statistics reported: normalisation and definition of self-citation.

Normalisation is a procedure through which one can obtain a relative score for citation performance to compare with the average level in the field. Details of various statistical procedures – such as the weighting of the values assigned to the papers and setting of the time frame for the reference group – affect the numeric outcomes. For instance, in the University of Oulu report for 2007–2012, CWTS states that the normalisation practice has been improved, with 'source normalisation' employed rather than traditional normalisation (Eskelinen & Ryyppö,

2014, p. 13 (in Appendix 12)). With the new form of normalisation, the world average is 1.1, whilst it is 1 in traditional normalisation (ibid.). The difference between them creates the possibility that numbers from separate reports cannot be compared.

Self-citations were conventionally excluded in the calculation of citation counts, but definitions of self-citation can differ. Usually, the bibliometric studies excluded self-citations. However, it is also possible to define self-citations in a manner encompassing only the first author, not the rest of the authors. The numeric results for all of the citation indicators may change as the definition of self-citation shifts. However, the influence of self-citation is relatively small.

In addition to output and impact, CWTS offers analyses of internal coverage ('Int_cov'), recency of references ('Vitality'), percentage of self-citation ('Psc'), and collaboration in some of the reports. While internal coverage was commonly used in reports applying CWTS standards (since it shows the relevance of WoS to the studies and the effectiveness of the method), the other indicators mentioned above were not commonly used in all the reports. Most of these indicators are not related to impact directly and instead depict other qualities of research; that might have led to their exclusion.

In addition, time window is a crucial factor in what the result of the bibliometric analysis turns out to be. The definition of publication time window is relatively straightforward. For instance, if it is set to 2010–2015, all publications published within that time span are included in the material for bibliometric analysis. Definitions for citation time window, however, vary. There are two, alternative ways of choosing this time window: variable-length citation window and fixed-length citation window. With the former, if the evaluation period is, for instance, 2010–2015, the statistics are based on the publications released during these five years and citation counts from 2010–2016 referring to all the papers. Under the second option, the citation count for a paper published, for example, in 2010 might be calculated on the basis of the counts from 2010 to 2012, while the citation count for a paper published in 2014 based on the counts for 2014–2016. It is evident that the former manner of calculation yields higher citation counts. The standard practice of CWTS is to use a variable-length citation window in the calculation of citation score. Since it always takes some time for citation counts to start accumulating, most of the RAE reports set a variable-length citation window that ends one year later than the evaluation period. However, when analysis of trends is involved – for instance, in comparison of publication output in 2000–2003, 2001–2004, and 2002–2005 – a fixed-length citation window is used.

It is important to note also that the time windows used in the various universities' bibliometric studies differ in their duration. For instance, some universities conduct their studies every four years, others every five years. Naturally, it is not appropriate to compare the numbers when the evaluation interval is not the same.

Table 5: Key indicators adopted from CWTS

Dimension	Indicator	UoH	UoO	TUT	HSE	Aalto
Output	P	×	×		×	×
Output	Pf			×[4]		×
Impact	TCS	×	×	×		
Impact	MCS	×	×	×	×	× ('CPP', 'CPP 2 year')[5]
Impact	PNC	×		×		×
Impact	MNCS	×	×	×	×	× ('NCSf')[6]
Impact	THCP 10[7]	×	×	×	×	× (top 5%)[8]
Impact (of journals)	MNJS	×	×	×	×	× ('NJCS')[9]
Relevance of WoS	Int_cov	×	×	×	×	
Recency of references	Vitality			×		×
Narrowness of speciality	Psc			×		× ('SCIT')[10]
Collaboration	Collaboration			×		× ('IntCOLLm' and 'AUm')[11]
Time window[12]	Publication	2005–2010	2007–2011	2005–2010	2006–2010/11	2003–2007
	Citation	2005–2011	2007–2012	2005–2011	2006–2011	2003–2008

4 There is confusion in the use of P and Pf in the report of TUT. Though in some parts 'P' is used, actually it refers to Pf, according to Appendix 4.

5 What is denoted by 'CPP' (citations per paper) in Aalto University's report is identical to the 'MCS' of CWTS, according to the definition used. In addition to CPP, 'CPP 2 year' (citations per paper with a two-year citation window) is used as an indicator of citation performance.

6 'NCSf' refers to field-normalised citation score – i.e., citations per publication relative to a reference value derived from the global averages for all articles in the relevant sub-fields. Under this definition, it can be inferred that NCSf is similar to the MNCS indicator of CWTS. However, the details of the statistical method might differ from those with CWTS, as is indicated in the report (Sandström, 2009, pp. 13–14). In addition, SCSf (standard field citation score) is used, as an indicator complementary to the NCSf value. SCSf is the number of standard deviations from the average.

7 'PPtop10%' and 'NPHCP10' are also used in the reports, to mean the same thing.

8 In contrast to the top 10%, Aalto University used the top 5%.

9 'NJCS' is similar in definition to the MNJS value used by CWTS. In addition to NJCS, the indicator NCSj (journal-normalised citation score) is used to measure the impact of the journals in which Aalto University staff published; NCSj measures citations per paper relative to the average number of citations per publication in the relevant journals.

10 The definition of self-citation applied in Aalto University's report is different from CWTS's. Under Aalto's definition, a citation is eliminated as self-citation only if the researcher in question is the work's first author; citation of works with other authors is not considered self-citation. At CWTS, self-citation involving any of the authors is excluded.

11 'IntCOLLm' (referring to international collaboration) is used for the mean number of countries per paper from the unit. 'AUm' (for 'author mean') refers to the mean number of authors per paper.

12 Time window is not a bibliometric indicator. This is the time frame during which publications and citations are calculated. It is included in Table 5 for convenience of presentation.

3.2.2 Complementary indicators

In addition to using CWTS's analysis, three universities – UoH, UoO, and UoT – conducted bibliometric studies on their own. The study done by UoH was based on their library databases. UoO's study used both the library database and Scopus. UoT's work used the WoS database. The indicators of the three studies are summarised in Table 6, which depicts greater diversity than does the research applying CWTS's standards.

Overall, in the analysis conducted by the three above-mentioned universities, the research output – i.e., number of publications and their distribution of various sorts – is reported on more extensively than the citation output, which is restricted primarily by the capability of the library databases. Also, research output is a dimension shared by the three universities' studies, whilst the same is not true for journal quality and impact. In the analyses of output, the Ministry's classification of publications (see the appendix to the present work) is commonly drawn upon in the studies of UoH and UoO.

In UoH's report, no citation statistics were shown. In the UoO study, citation was studied on the basis of the Scopus database for the 17 units for which the relevance level with WoS is not ideal, as noted above. The indicators involved are the citation counts, average number of citations per paper, and percentage of un-cited publications. For the UoT report, the bibliometric analysis was conducted on a discipline-by-discipline basis, and the results for each discipline were compared to those for that discipline at other Finnish universities. The analysis also features data from Statistics Finland for the number of publications in certain fields: education, economics, social sciences, political science, administration, media and communication, law, geography, cultural studies, art, languages, and history. The indicators used by UoT include the total number of citations by a unit, mean citation by publication, total number of citations by a unit by year, H-index, and sum index.

Though the universities' analysis of citation is limited, it is complemented by the statistics generated by means of publication rating sources, which offer a view of the quality of the publication channel in which the publications appear. The three universities vary in their choices of rating sources. The rating sources referred to include the Finnish Publication Forum's rankings, Norwegian journal ranking lists, Norwegian/Australian publication rankings, and Australian conference rankings. The numbers of articles / books / conference papers originating from the units assessed are reported as denoting the quality of the journals, publishers, or conferences, for which the significance is similar to that of the mean normalised citation score (the MNJS indicator) for journals as used by CWTS.

Table 6: Complementary indicators

Dimension	UoH	UoO	UoT
Output	- Number of publications of each of the 25 types in the Ministry's classification - Percentage distribution and number of publications by language - Percentage distribution and number of publications by number of authors - Number of publications by subject/faculty/field	- Number of all publications by type - Number of scientific publications by year and type - Number of publications in the Scopus database - Percentage distribution by language	- Total number of publications that contain the university city's name in each of the 47 disciplines (total number by university not reported), not mentioned if fractionalised
Quality of journals or conferences	- Number of articles on the Finnish Publication Forum ranking list and Norway's journal ranking list - Number of publications by the units listed in the Norwegian or Australian publication ranking[13] - Number of conference papers from the units according to Australian Conference Rankings[14]	- Number of publications according to Finnish Publication Forum rating for journals and publication series and for book publishers, along with the distribution by language	No statistics
Impact	No citation statistics	- TCS: Number of citations (from Scopus) - MCS: Average number of citations per publication (from Scopus) - %uncited: Percentage of uncited publications	- Total number of citations by a discipline at a university - Mean citation score by publication (citations per paper) and annual citation by discipline in the university - H-index - Sum index[15]

[13] This analysis covers only biological, agricultural, and veterinary sciences; natural sciences; humanities; and social sciences.

[14] This analysis covers computer science and information systems, telecommunications, and computer science and artificial intelligence.

[15] [SumN(Citations)/CumSum(Papers)], in which the cumulative sum of citations per year [citations(1993)+citations(1994)+citations(1995)+...citations(2008)] is related to the cumulative sum of publications [publications(1993)+publications(1993–1994)+publications(1993–1995)+...publications(1993–2007)]. This approach is aimed at levelling out the effects of differences in size between the universities.

3.3 Discussion

In summary, this section reports on the methods used in the bibliometric analyses conducted for the Finnish universities' RAE reports. The findings are presented in two sections, with Subsection 3.2 focusing on purely bibliometric indicators while Subsection 3.1 discusses the rest of the elements describing the RAE implementation process. It gives an overview of how bibliometric studies can differ from each other, which is based on comparative investigation of the six reports. Continuing with this perspective, the final section brings two other reports by the Ministry into the discussion and comparison (on the basis of a reading of the summaries available in English).

In the two reports, the report from Finnish Citation Index Working Group II (2012) and 'WoS or Scopus?' (2013), the indicators used are similar: number of publications, top 10 index, and field-normalised citation index. According to the descriptions of the method, the similarities and differences between the Ministry's studies and the universities' can be viewed from several angles.

Firstly, the key indicators are similar in definition though the wording differs. The relative citation index indicator is almost the same as CWTS mean normalised citation score by definition, and number of publications, fractionalised number of publications, and top 10 index are no different. However, there are differences in operationalisation between the Ministry's and CWTS's work with respect to citation rate. In the Ministry's, citation is fractionalised and the value of the number of citations is weighted with fractionalisation. The universities' reports do not explain whether or not citations are fractionalised, though in some of the reports the numbers of publications are fractionalised. Furthermore, the aggregate of publications in the Ministry's studies is formed with the criterion of the country code for Finland, whilst the publications considered by the universities are from those defined as research active staff on the basis of either a specified date or employment history. Moreover, in the Ministry's procedure, the publication types included are article, letter, and review, whereas some of the universities also consider conference proceedings. Finally, the time windows used are evidently different between the Ministry's and the universities' work.

4. Research assessment in the UK and the Netherlands

This report considers two cases outside Finland – research assessment in the Netherlands and the UK – with the focus being on not specific universities but the guidelines for research assessment at national level in general. These two countries have been selected because they have a long history of research assessment and their systems are commonly recognised as highly developed, though still not judged impeccable in the public debate in those countries. In this section, international perspectives are brought into the discussion with the two cases studied, both to shed light on how

research assessment could be organised at national level and to inspire our reflection on current practices in Finland.

4.1 The research excellence framework (REF) for 2014 in the UK

4.1.1 An overview of the work flow of the REF

The REF supersedes the earlier 'RAEs', completed in 1996, 2001, and 2008. The new framework – which is developed in an evolutionary process on the basis of RAEs – has been developed in response to the government's intention of reforming the research assessment system in the UK. The REF is the ongoing research assessment project aimed at publishing a final report to the public in December 2014. It is organised by the four funding bodies: the Higher Education Funding Council for Wales, the Higher Education Funding Council for England, the Scottish Funding Council, and the Department for Employment and Learning. The history of the REF/RAE provides orientation to the purpose of informing funding allocation. The university- and unit-level assessment results are important reference for the funding bodies' decision-making. ('Background', 2012)

In practice, the REF involves a series of processes. It starts with the REF team's preparatory work on proposal of a new framework, initial studies and a pilot exercise, panel arrangement and recruitment of members, decision on the panels' work methods, development of a submission guide, online submission system readiness and a user guide. Finally, the research institutes are invited to make submissions. The information flow is open to the public and transparent, with the documents describing the work done in the course of the project available on the official REF Web site ('REF 2014', n.d.).

Secondly, upon the call for submissions, all publicly funded institutes of higher education need to send in the material for evaluation. The unit of analysis is defined as the academic subject, and there are 36 units, in total, for evaluation at any given institution. For each university or other institution, the evaluation material is submitted in 36 separate subsets, assumedly with a different individual submitter in charge of each specific subject. In the online system, all submissions from a single institution are locked and marked as one set. At the end of the submission period, all 36 subsets of the submission material (for the 36 units/subjects) are made ready for the panel review.

Thirdly, in the assessment phase, the 36 units' materials are evaluated by 36 sub-panels, which also form four main panels. In essence, the roles and responsibilities of the main panels consist of ensuring that the sub-panels follow the criteria and work methods decided upon and those of sub-panels assessing the submissions. Each main panel consists of one chair, those chairs of the sub-panels under it, and may also include additional members, with international experience or certain

expertise that is needed. Each sub-panel consists of one chair and 10–30 other members, and it is possible to hire additional assessors when necessary.

The final report is compiled and published by the REF team on the basis of the assessment results generated in the panel review. The contents of the quality profile reports are explained early in the document for the assessment framework and the guidance on submissions. In December 2014, the reports in this iteration of the process will be released to the participating higher education institutes. Later, in spring 2015, the reports, along with material such as minutes of panel meetings, will be published in the Web site for public access.

4.1.2 The evaluation framework for the REF

There are three generic criteria used in the assessments for all submissions: outputs, impact, and environment. Output has to do with the quality of the research outputs in terms of originality, significance, and rigour by comparison with the international standards. Impact is evaluated with respect to societal impact – influence on culture, economy, and society. It is important to note that the impact within academic communities, typically demonstrated by citation rate, is assessed under 'output' and 'environment', instead of 'impact' ('Assessment framework and guidance on submissions', 2011, p. 26–27) . Environment is an element related to the sustainability and vitality of the research environment assessed and to its contribution to the sustainability and vitality of the wider environment of the discipline.

For the overall quality rate associated with a submission, a weight of 65% is attached to outputs, 20% to impacts, and 15% to environment. The submissions are rated at four levels for these three elements: three-star, two-star, one-star, and unclassified (from the highest to lowest quality). Both the overall quality value for each submission and the sub-profiles for the three component elements will be presented in the assessment results to be published.

In order to unpack the substance of what these three elements/criteria refer to, it is necessary to introduce the material the evaluation is based on. Most importantly, as the term 'research excellence framework' suggests, the assessment targets are set in terms not of all research outputs and research staff of the units but for the excellent research output and researchers. Therefore, the submitting institutions are allowed to submit selectively and present their 'highest quality of research' for evaluation. This is what is recommended by the REF team. However, even though the selection of staff is a matter for the institution's autonomous decision, social equality and diversity must be taken into consideration according to the REF recommendations. For example, if all the research staff selected are white, male, and somewhat older, questions related to discrimination and unfavourable assessment results could arise.

For the evaluation of output, the required material includes not only publication products such as articles and books but also a great variety of other items: devices, images, patents, events,

and so on. The outputs eligible for submission are those that became publicly available during the assessment period – here, 1.1.2008–31.12.2013. A co-authored output will be counted as a single piece without fractionalisation in the same way as work with a single author, but information on the number of authors is required. Citation data may be – but not necessarily – included in the material about output. Citation data are obtained by the REF team from a single source, and the institutions are able to verify these in the submission system.

Impact in the REF context refers to the effects of the research outputs in nearly every other sphere of society but academia – public policy, services, people's health and well-being, environment, etc. Impact is assessed beyond the geographic limits of the UK, to full international extent. The evaluation of impact is based on case studies, the required number of which ranges from two to seven, depending on the profile of the research staff under evaluation. In each submission, the material submitted for the evaluation of impact consists of two parts: a template-based report describing the approach of the unit enabling the impact of the research and several case studies of successful research having a considerable impact on society. The two parts are integrative in that the former elucidates the background and particular circumstances of the case studies submitted.

The environment is evaluated on the basis of the material describing the resources of the unit assessed. In specific terms, the text needs to include an overview, research strategy, number of doctoral degrees awarded, number of research students and staff, research funding, infrastructure and facilities, collaboration, and contribution to the discipline in question.

Besides all of the above-mentioned material, information on the staff who are evaluated is required, such as which researchers are included and to which staff categories they belong. Though the profile of the research staff is not rated, the detailed information not only helps the REF team and the submitting institutions with verification and monitoring but also offers a basis for ensuring that concerns of social equality are addressed reasonably. For example, when evaluating the research outputs of researchers early in their career; part-time researchers; and those whose circumstances include maternity leave, health problems, injury, etc., the sub-panels are able to make judgements about what constitutes a reasonable reduction in output.

4.1.3 The bibliometric analyses and the indicators

The decision-making of the REF on bibliometric indicators – citation scores in this case – started with a pilot exercise, which was based on a subset of the staff population from RAE 2008. In the pilot exercise, the normalised citation scores for the selected staff's publication were analysed from both Scopus and WoS, for comparison to ensure good coverage ('Analysis of data from the pilot exercise to develop bibliometric indicators for the REF', 2011). In addition, sophisticated analyses of percentage distributions of the various groups in terms of particular staff characteristics (e.g., male

vs. female, UK vs. non-UK, or ethnicity) were performed for insight into whether social equality is implemented well.

To make a long story short, based on the above-mentioned pilot exercise the REF team decided to leave it to the sub-panels to judge whether citation data are necessary for their evaluation. Because the importance and necessity of citation data varies greatly from unit to unit within the body of evaluated research and staff, it is feasible to set a formulaic requirement for all units. Later in the process, 11 of the 36 sub-panels decided that they need to be informed via citation data ('Citation data', 2013). Therefore, citation data are provided for only these 11 sub-panels. In other words, no formulaic bibliometric studies are conducted that cover all of the units assessed. Instead, the evaluation material for only around 30% of the units contains bibliometric data.

The citation data were provided by Elsevier on the basis of Scopus, which was selected because of its good coverage, demonstrated in the pilot exercise: 94% of the publications were covered in Scopus ('Citation data', 2013). The REF submission system is interfaced with Elsevier's systems. After the institutes submit their outputs in the system, the outputs can be matched automatically with data in Elsevier's systems. Citation counts are returned and displayed in the REF submission system.

The bibliometric indicators provided to the sub-panels include the citation count, along with contextual citation data that serve as a benchmark for that count. The contextual citation data consist of the mean citation count for the subjects and the citation count values with which a publication can get into the top one, five, 10, and 25 per cent ('Citation data', 2013; 'Contextual citation data', 2013). The data for these indicators are available for both individual outputs and unit of analysis, and they are provided to both the panels and the institutions. Although normalisation was involved in the pilot exercise and evident from the meeting records for the earlier discussions ('Analysis of data from the pilot exercise to develop bibliometric indicators for the REF', 2011; 'Report on the pilot exercise to develop bibliometric indicators for the Research Excellence Framework', 2009), no normalisation or fractionalisation was conducted in the stage of submission ('Citation data', 2013; 'Contextual citation data', 2013).

4.2 The Standard Evaluation Protocol (SEP) in the Netherlands

4.2.1 An overview of the SEP

The SEP is the set of national research evaluation guidelines developed by the Association of Universities in the Netherlands (VSNU), the Netherlands Organisation for Scientific Research (NWO), and the Royal Netherlands Academy of Arts and Sciences (KNAW) to guide the universities' research assessment. The research assessments in the Netherlands usually take place every six years.

The assessment is targeted at examining the research conducted at Dutch universities, university medical centres, NWO, and the KNAW institutes.

In contrast to the UK's REF, which is organised in a centralised manner at national level, the SEP involves only guidelines and the research assessments are organised by the universities/units themselves. In addition, the purpose of the evaluation in the Netherlands differs from that in the UK, in that it has nothing to do with funding allocation. From the available references, it appears that the Dutch submission requirements are relatively flexible and the institutions evaluated also play a role in the final scoring and judgement, which will be further explained below.

The research assessment projects recently undertaken were organised in 2013. The universities and units follow the requirements set in the SEP 2009–2015 materials ('Standard evaluation protocol (SEP) 2009–2015', 2009). However, since the research assessments are organised 'on the ground', the timetables and the evaluation periods vary between universities/units. For most of the universities, the evaluation period is 2007–2012; while for others it is 2006–2011 or 2005–2011. The final reports of the assessed units were separately published in 2013 on the Web site of Quality Assurance Netherlands Universities (QANU) ('Assessment reports', 2013) and the universities' own Web sites. There is no such integrative report available as compares evaluation results across all the assessed units or informs a ranking of these units. At least one could not be found from a search attempted in English.

The upcoming research assessments will be performed in accordance with the SEP 2015–2021 ('Standard evaluation protocol (SEP) 2015–2021', 2014), with changes being made in both assessment dimension and focus. For reasons of being up to date with recent developments, this section of the report refers to the latter protocol, for 2015–2021, though assessment using the guidelines developed in 2014 has not taken place yet. The document introducing the SEP 2015–2021 system, together with a few Web pages of the hosting organisations, is almost the only information source that this study can use.

There are three actors in the implementation of research assessment according to the SEP: the board, the research units assessed, and an assessment committee consisting of external experts. The board acts as the organiser, the responsibilities of which include handling schedule control, composing the guideline, appointing the assessment committee, co-ordinating practical matters between the assessed research units and the assessment committee, publishing the final report, etc. The board is formed by the universities/units on their own, not appointed by national organisations such as NWO and KNAW. The units and the assessment committee, respectively, have the role of presenting the research performance and judging it, in line with the guidance issued by the board. Since the work flow of panel-review-based assessment projects under the SEP and in the UK's REF is more or less similar, the SEP case is overviewed here with comparison with the REF, to avoid redundant narratives.

Relative to the REF in the UK, a number of differences can be noticed in the case of SEP. Firstly, the definition of the unit of assessment is less formulaic and specific. In the research institutes' submission, the research unit might be a research group, a research institute, a research cluster, etc. – as long as it meets several conditions, such as being sufficiently large and having been established for at least three years. Furthermore, the board can decide that some units may constitute an aggregate for evaluation by a single committee. Aggregation of units can also occur at the national level in a discipline-specific assessment. It is not mentioned whether the assessment should cover all outputs of research-active staff or, selectively, only some outputs, as with 'excellent' research in the case of the REF.

Secondly, as is suggested by the SEP, the committees' site visits to the units assessed are important, as is reviewing the evaluation material. In the site visits, the committees can meet the personnel via meetings and interviews and thereby have better understanding of the context of the material presented. In addition, the committee' assessment reports need to be reviewed by the assessed units, for comments. The units can address issues from their point of view and the scoring can be fine-tuned. Thus, the final evaluation results are based on consensus between the committees and the units assessed.

Third is self-assessment, the most important part of the body of evaluation material. This is one more means of voicing the units' opinions from the bottom up. The required material includes the conclusions from the previous assessment, the unit's self-assessment, and appendices to the self-assessment. The self-assessment report is limited to 15 pages and is to consist of descriptions of the unit's organisation, strategy, and efforts; the results achieved; SWOT analysis; description of the PhD programmes; and material on research integrity. The report is supported by appended tables providing quantitative information such as a profile of the research staff, the financing structure, output indicators, and a list of the most important publications in scientific or societal terms.

Ultimately, the staff profile requirement is less detailed in the SEP context. In the research staff description, the quantitative information is organised in the dimension only of research experience, with respect to categories of scientific staff (tenured and non-tenured senior scientists), postdocs, PhD students (with funding), support staff, and visiting fellows. Information on gender, age, nationality, ethnicity, special circumstances, and so on is not required as in the REF in the UK.

4.2.2 The evaluation criteria and bibliometric data

There are three criteria for assessment: research quality, relevance to society, and viability. They are explained in this section of the report. Research quality has to do with the units' scientific outputs, such as scientific publications, instruments, datasets, software tools, and infrastructure developed by the unit. In the submission, research quality needs to be presented with statistics on the outputs, their impact in scientific communities (demonstrated with the indicators of citations, how the

instruments or tools are used by peers, etc.), and marks of recognition from peers (demonstrated in the form of prizes, awards, research grants, solicitation of lectures, membership of editorial boards, etc.).

Relevance to society is viewed in terms of the contribution to society economically, culturally, and socially. Likewise, relevance can be demonstrated also from three perspectives in broader societal context: what outputs are developed for certain societal groups, such as articles in professional journals and lectures for the public; how the outputs are used by various societal groups (use of patents or licences, collaborative projects involving various groups in society, etc.); and marks of recognition: prizes, membership of advisory boards in civil society, and so on. Case studies are recommended for presentation of performance in relation to these issues.

Viability is a concept used for the feasibility of the unit's strategy for the future in consideration of the support available from the infrastructure. How the units should prepare the material to show their performance in this dimension is not illustrated.

It is important to note that the above-mentioned indicators supporting the evaluation for these three criteria are all introduced as 'example' indicators, which means that they are not formulaic templates; instead, the units assessed can select these or use other indicators in view of their specific profile and circumstances. Simultaneously, it is recommended that the units adhere to the commonly used definitions and measurement methods or those based on agreement within their institution or research field.

Where bibliometric indicators are involved, there is no specific requirement or set of centralised studies. As the units' evaluation reports from 2013 ('Assessment reports', 2013) show, some but not all of the reports contain quantitative information on the number of publications, with a table. These numbers of publications have no single source: some are based on an external database such as Google Scholar, and others involve self-reporting, without noting the source of statistics. It can be also so that bibliometrics have been used as a background data when informing panels about the units' performance.

For the three criteria discussed above, research performance is marked in terms of four levels: 1) world-leading/excellent, 2) very good, 3) good, and 4) unsatisfactory. Besides the quantitative grading, qualitative assessment too – in terms of both the three criteria respectively and the unit as a whole – should be presented in the committees' report, along with assessment of the PhD programmes and the unit's research integrity and recommendations as to strategies for the future.

4.3 Discussion

Foremost, it needs to be clarified that the presentation of the two international cases is not meant to set out examples for Finland to copy. Different countries are embedded in different situations in

terms of their research staff population and academic culture. Furthermore, different countries are confronted with divergent challenges and have their own development strategies. For instance, that assessment in the UK focuses on 'excellent' research work and the funding allocation is informed by the assessment results does not mean that Finland must follow suit. However, there are insights we can gain from the fundamental principles applied in these cases, helping us reflect on research assessment in Finland.

Firstly, transparency is one factor that contributes to credibility in the UK's REF 2014 case. Instead of only the final report being published at the end of the assessment, all decision-making, step by step, such as that on criteria, procedure, resources, and guideline, is well-explained and made available with documents for public access. The flow of information is open between all parties, including the REF team, the evaluated institutes, the panels, the funding councils, and the public. However, this obviously is costly and time-consuming to implement. For instance, for the REF 2014 process, some preparatory work started even before RAE 2008 was over.

Secondly, in the REF case, staff profiles are required to be reported upon in detail. From the REF team's side, policy is developed to guide the institutions' selection of research staff in such a way as to facilitate social equality and diversity. On the institutes' part, the demographic information of staff and which of the specified categories they belong to are submitted by means of formulaic templates and dealt with by statistical means. All these details are not only helpful for enhancing the well-being of the academic system but also informative when one is comparing staff profiles horizontally across units. For instance, when two units, one with a higher proportion of senior scientists and full-time research staff and the other with a lower proportion of these, are compared, the detailed reports on staff profiles assist to some extent in making a fair judgement.

Thirdly, the SEP guidelines in the Netherlands, with the arrangement of site visits and the opportunities for the units to give feedback on the committee's assessment report, are valuable for promotion of effective communication through provision of more concrete experiences and information beyond the textual material submitted and help to avoid misunderstanding.

5. Final remarks

This report has examined the methods of bibliometric analysis used in the research assessment exercise reports of Finnish universities, with six reports as the focus of analysis. It has given an overview of the process of implementation of research assessments and the differences in the methods used in those studies. Furthermore, for expansion of the horizon, international cases of the UK and the Netherlands were reported on, with emphasis on their assessment frameworks and criteria. The study has concentrated on the bibliometrics aspects of the research evaluation. It would be valuable to study the qualitative evaluations and the state of the art of qualitative expertise as well.

For a general discussion on all the cases studied, the main difference between the UK and the Netherlands, on one hand, and Finland, on the other, can be described as being that the former two countries have national instructions for evaluation while the universities in Finland have their own schedules and procedures, which are defined in line with the universities' needs, with their own logic and schedule. This leads to diversity in practices and results due to differences in operationalisation and conducting of the RAE.

At the same time, the reasoning for the universities' procedures stems from the autonomy of the universities. Also, the universities have conducted the RAEs from the standpoint of their internal needs. Doing so allows them to use the RAEs for their own development and internal profiling. An RAE of an individual university is also a learning process in which the university defines itself anew. In this sense, a made-for-all national protocol could hamper creativity in the university's self-learning process.

The current diversity of practices of Finnish universities in research assessment requires a broad base of knowledge about bibliometrics and bibliometric 'literacy' if one is to be able to make full use of the information provided by the reports. The actors reading RAE materials and their knowledge about bibliometrics vary greatly, so there is a risk that the bibliometric data may be understood in such a manner that its limitations are not considered carefully.

One of the factors explaining the differences between the RAE reports is that the Finnish universities have tried to deal with the issue of discipline-to-discipline differences in publication culture, also a topical issue in bibliometrics itself (Puuska, 2014). To ensure equal treatment of the various disciplines at the operational level of RAE, the universities have used various databases, different indicators, and qualitative assessment. The issue cannot be resolved in a simple way even if national guidelines for research assessment were developed. There the discussion about the usefulness and feasibility of bibliometrics continues. It is important to be aware in Finland of this international discussion and the trends in bibliometrics. The analysis herein of research assessment protocols in the UK and the Netherlands serves as a preliminary starting point for accumulation of this kind of knowledge in Finland.

Appendix 1. Abbreviations of databases, universities and indicators

RAE   Research assessment exercise
CWTS   Centre for Science and Technology Studies, Leiden University, the Netherlands
WoS   Thomson Reuters Web of Science Database
Scopus   Elsevier abstract and citation database
UoH   University of Helsinki
UoO   University of Oulu
UoT   University of Turku (and Turku School of Economics)
Aalto   Aalto University
TUT   Tampere University of Technology
HSE   Hanken School of Economics

Abbreviations of CWTS bibliometric indicators

Dimension	Indicator	Definition
Output	P	Total number of publications – unique papers.
Output	Pf	Number of fractionalised papers.
Impact	TCS	Total citation score. The total number of citations received by P over the entire period, excluding self-citations.
Impact	MCS	Mean citation score. The average number of citations per publication, excluding self-citations.
Impact	PNC	Percentage not cited. The percentage of articles not cited during the period under review (the figure does not consider self-citations).
Impact	MNCS	Mean normalised citation score. The papers citation score in comparison to the international level in the field, relative to specific years and in consideration of the type of journals in which the research units' papers were published.
Impact	THCP 10 (or 'PPtop10%', 'NPHCP10')	The field-normalised proportion of publications belonging to the top 10% in terms of citation impact.
Impact (of the journals)	MNJS	Mean normalised citation journal score. The citation score of the journals in which the researchers publish.
Relevance of WoS	Int_cov	Internal coverage. The average number of the publications' references that are covered by WoS. The higher this number is, the more possible it is that a unit's publications are indexed in WoS.
Recency of references	Vitality	Average age of references per assessment unit.
Narrowness of speciality	Psc	Percentage of self-citation.

Appendix 2. The Ministry's Publication Type Classification

(Ministry of Education, 2010)

	Code	Publication Type
Refereed scientific article	A1	Journal article, original research
	A2	Review article in scientific journal
	A3	Book section
	A4	Conference proceeding
Non-refereed scientific article	B1	Non-refereed journal article
	B2	Book section
	B3	Conference proceeding
Book	C1	Scientific monograph
	C2	Edited book, conference proceedings or special issue of a journal
Publication for professional communities	D1	Article in trade journal
	D2	Article in a professional manual or guide or professional information system, textbook material
	D3	Professional conference proceeding
	D4	Published development or research report or study
	D5	Textbook, professional manual or guide, dictionary
Publication for public	E1	Popularised article, newspaper article
	E2	Popularised monograph
Public artistic and design activity	F1	Published independent work of art
	F2	Public partial realisation of a work of art
	F3	Public artistic performance or exhibition
	F4	Model or design adopted for production/use
Thesis	G1	Polytechnic thesis, Bachelor's thesis
	G2	Master's thesis, polytechnic Master's thesis
	G3	Licentiate thesis
	G4	Doctoral dissertation (monograph)
	G5	Doctoral dissertation (article)
Patent and invention disclosure	H1	Granted patent
	H2	Invention disclosure
Audiovisual material, ICT software	I1	Audiovisual material
	I2	ICT software

References

Analysis of data from the pilot exercise to develop bibliometric indicators for the REF. (2011). Retrieved 10 October, 2014, from http://www.hefce.ac.uk/media/hefce1/pubs/hefce/2011/1103/11_03.pdf

Assessment framework and guidance on submissions. (2011). Retrieved 9 October, 2014, from http://www.ref.ac.uk/media/ref/content/pub/assessmentframeworkandguidanceonsubmissions/GOS%20including%20addendum.pdf

Assessment reports. (2013). Retrieved 28 October, 2014, from http://www.qanu.nl/en/assessment-reports

Background. (2012). Retrieved 10 October, 2014, from http://www.ref.ac.uk/background/

Citation data. (2013). Retrieved 10 October, 2014, from http://www.ref.ac.uk/subguide/citationdata/

Contextual citation data. (2013). Retrieved 10 October, 2014, from http://www.ref.ac.uk/subguide/citationdata/contextualdata/

Eskelinen, S., & Ryyppö, A. (Eds.). (2014). Research assessment exercise report 2014 – International evaluation of research at the University of Oulu

Evaluation of research. (n.d.). from http://www.uef.fi/en/uef/tutkimuksen-arviointi

Evaluation of research activities 2005–2009. (2010). from http://www.uva.fi/fi/research/publications/finalreport.pdf

Folea, A. (2011). Evaluation of research activities 2005–2009, from https://www.jyu.fi/hallinto/neuvostot/tiedeneuvosto/arviointi/RA2010report

Haila, K., Holm, R., & Niemelä, R. (2005). Research assessment exercise evaluation reports, from http://www.helsinki.fi/arviointi2010-2012/tutkimuksenarviointi_raportti_2005.pdf

Hanken School of Economics evaluation of research 2012. (2012). from www.hanken.fi/media/2109/evaluationofresearch_2013_webben.pdf

International evaluation of strategic research at the University of Lapland. (2013). from http://www.ulapland.fi/loader.aspx?id=b5a141e3-015b-4d6b-90c4-568fe787529c

Krause, O., Saaristo, A., Lehtovaara, K., Sivenius, P., & Pasanen, I. (Eds.). (2009). Striving for excellence – Aalto University research assessment exercise 2009 and bibliometric analysis 2003–2007 Project report.

Krippendorff, K. (1980/2012). Content analysis: An introduction to its methodology. Newbury Park, CA: Sage Publications.

Laine, P. (2010). Tutkimuksen kokonaisarviointi 2008–2009 Loppuraportti, from https://www.doria.fi/bitstream/handle/10024/61586/turunyliopisto_tutkimuksen_arviointi_raportti_20101.pdf?sequence=4

Liikanen, A., Tirronen, J., Keinänen, R., Sagulin, M., Simonaho, S.-P., Taskinen, H., & Mönkkönen, J. (2014). International evaluation of research activities at the University of Eastern Finland 2010-2012, from http://epublications.uef.fi/pub/urn_isbn_978-952-61-1348-7/urn_isbn_978-952-61-1348-7.pdf

Ministry of Education. (2010). Publication type classification, from http://www.aka.fi/Tiedostot/Tiedostot/Liitetiedostot/OKM_julkaisutyyppiluettelo_2010_en.pdf

Pellinen, R., Liikanen, A., & Kalliokoski, P. (2008). Evaluation of the research activities at Kuopio University, Kuopio, Finland, during the years 2000–2006.

Puuska, H.-M. (2014). Bibliometrics and the assessment of research peformance. Tampere: Tampere University Press.

Report from the Finnish Citation Index Working Group II. Finnish research organizations' publications and citations in the Web of Science, 1990–2009. (2012). Retrieved 24 September, 2014, from http://www.minedu.fi/export/sites/default/OPM/Julkaisut/2012/liitteet/okm18.pdf?lang=fi

Report on the pilot exercise to develop bibliometric indicators for the Research Excellence Framework. (2009). Retrieved 10 October, 2014, from http://www.hefce.ac.uk/pubs/year/2009/200939/

Research assessment exercise (RAE). (n.d.). from http://www.lut.fi/web/en/research/research-assessment-excercise-rae

Ruskoaho, H., & Jaako, J. (2008). Research assessment exercise 2007.

Saari, S., & Moilanen, A. (Eds.). (2012). International evaluation of research and doctoral training at the University of Helsinki 2005–2010.

Sahlander, T., & Soini, T. (2005). Research evaluation of the University of Tampere.

Sandström, U. (2009). Aalto University bibliometric report 2003-2007, from http://dept.ics.aalto.fi/calls/rae2009/aalto_university_bibliometric_report_2003-2007.pdf

Standard evaluation protocol 2015–2021. (2014). Retrieved 10 October, 2014, from http://www.knaw.nl/nl/actueel/publicaties/standard-evaluation-protocol-2015-2021

Standard evaluation protocol (SEP) 2009–2015. (2009). Retrieved 10 October, 2014, from http://www.knaw.nl/nl/actueel/publicaties/standard-evaluation-protocol-sep-2009-2015

TUT research assessment exercise 2011: Bibliometric report 2005-2010. (2011). from http://www.tut.fi/idcprod/groups/public_news/@l102/@news/@p/documents/liit/p021003.pdf

TaSTI Working Papers series

1/2008

Himanen, Laura & Pasanen, Hanna-Mari (2008). Julkaisut yliopistojen perusrahoituksen indikaattorina. Australian, Iso-Britannian ja Norjan rahoitusmallien vertailu.

2/2008

Lyytinen, Anu, Marttila, Liisa & Kautonen, Mika (2008). Tutkimus- ja kehitystoiminnan haasteita ja mahdollisuuksia monialaisissa ammattikorkeakouluissa.

3/2009

Pihlajamaa, Elina (2009). Oslon alueen työmarkkinoiden kansainvälistymispalvelujen kehittyminen.

4/2009

Vuolanto, Pia, Vehviläinen, Marja & Ylijoki, Oili-Helena (2009). Female employees in intermediary organisations: tasks, careers and networks.

5/2010

Marttila, Liisa, Lyytinen, Anu, Ylijoki, Oili-Helena & Kaukonen, Erkki (2010). Tutkimusyhteisöjen ja akateemisen työn muutos. Laitosjohtajakyselyn tulokset.

6/2010

Vuolanto, Pia, Puuska, Hanna-Mari & Kaukonen, Erkki (2010). Tutkimusresurssien kehitys Suomen yliopistoissa vuosina 1991-2006.

7/2010

Raunio, Mika, Korhonen, Marjaana & Hoffman, David (2010). Kannattaako kansainvälistyä? Suomen yliopistot kansainvälisinä akateemisina ympäristöinä. Loppuraportti.

8/2013

Auranen, Otto, Leino, Yrjö, Poropudas, Olli & Pölönen, Janne (2013). Julkaisufoorumi-luokitus ja viittausindeksit tieteellisten julkaisujen laadun mittareina. Web of Science -aineistoon perustuva vertailu.

9/2013

Raunio, Mika, Kautonen, Mika & Saarinen, Jukka P. (2013). Models for International Innovation Policy : Transnational Channels and Regional Platforms : Fostering Globalizing Innovation Communities in Finland and Abroad.

10/2014

Wang, Li, Vuolanto, Pia & Muhonen, Reetta (2014). Bibliometrics in the research assessment exercise reports of Finnish universities and the relevant international perspectives.

www.uta.fi/tasti