Vol. 13: 193–202, 2013
doi: 10.3354/esep00140

ETHICS IN SCIENCE AND ENVIRONMENTAL POLITICS
Ethics Sci Environ Polit

Printed May 2014
Published online November 21, 2013

Contribution to the Theme Section 'Global university rankings uncovered'

FREE ACCESS

Global university reputation and rankings: insights from culturomics

Konstantinos I. Stergiou[1,2,*], Athanassios C. Tsikliras[1]

[1]**Laboratory of Ichthyology, Department of Zoology, School of Biology, Aristotle University of Thessaloniki, UP Box 134, 541 24 Thessaloniki, Greece**

[2]**Institute of Marine Biological Resources and Inland Waters, Hellenic Centre for Marine Research, Aghios Kosmas, 16777 Athens, Greece**

ABSTRACT: In this study, we used culturomics (i.e. analysis of large electronic datasets for the study of human culture) in order to study the use of the names of various universities in the digitized corpus of English books. In particular, we used the Google Ngram viewer (available online: http://books.google.com/ngrams) to produce the frequencies of the names of 13 US, 5 UK and 4 Canadian universities in the English books and examined how these frequencies changed with time (1800 to 2008). We further used these frequencies to establish reputation rankings for these universities. Our results showed that Ngram is an easy-and-cheap-to-apply tool to approximate the reputation and 'intellectual' impact of universities over long time periods. Its reputation-generating capability, at least for top universities, is not worse than the within- and between-system capabilities of commercial tools (i.e. QS, THE and THE World Reputation Rankings). Ngram can, thus, be promising at least for students (and their families), who make choices that are affected by rankings, providing them with additional benefits (e.g. perception of the historical impact of a university) when compared to the short-term, volatile annual commercial rankings.

KEY WORDS: Global university rankings · University reputation · Google Ngram · Culturomics · QS · THE · Reputation rankings

INTRODUCTION

Global university rankings (GURs) are attracting increasing attention on the agenda of stakeholders directly or indirectly related to higher education (e.g. politicians, managers, administrators, policy makers, institutions, academia, students), and the number of agencies performing GURs is increasing with time (e.g. Harvey 2008, Williams 2008, Rauhvargers 2011, 2013, Jarocka 2012, Hazelkorn 2013). Available global ranking systems develop their annual league tables based generally on (e.g. Buela-Casal et al. 2007, Enserink 2007, Federkeil 2009, Huang 2011, Rauhvargers 2011, 2013, Hazelkorn 2013) (1) a variety of quantitative criteria and measures which are given different weights (e.g. number of papers, publications in Science/Nature, number of citations, number of Nobel Prize winners among their staff and alumni, faculty:student ratio); (2) web presence, visibility and access (such as Webometrics); and (3) reputation, such as the World Reputation Rankings (THER), produced since 2010 by Times Higher Education (THE), which is based on an invitation-only survey of academic opinion. The degree of subjectivity of reputation rankings is increasing (see Federkeil 2009, Rauhvargers 2011, 2013, for an extensive discussion on reputation rankings and their shortcomings).

Fame, or reputation, is what is said or reported about a name. Van Vught (2008, p. 169) stated the following. 'The reputation of a higher education institution can be defined as the image (of quality, influence, trustworthiness) it has in the eyes of others. Reputation is the subjective reflection of the various actions an institution undertakes to create an external image. The reputation of an institution and its quality may be related, but they need not be identical. Higher education institutions try to influence their external images in many ways, and not only by maxi-

*Corresponding author: kstergio@bio.auth.gr

mizing their quality.' University reputation, which has different meanings for different groups and scientific fields, is 'a form of social capital within the system of higher education that can be transformed into economic capital, too' (Federkeil 2009, p. 32).

Although fame, on an individual perception basis, might be subjective, it can be objectively measured by quantitatively estimating the frequency of the name appearing in various sources, including books (Michel et al. 2011). The digitization of millions of books available online provides an important source and opportunity to study cultural trends (and human behavior) based on the quantitative analysis of language and word usage in such digitized texts; this new scientific field is known as culturomics (Michel et al. 2011).

Michel et al. (2011) constructed a corpus of digitized books (nowadays making up ~6% of all books ever printed: Lin et al. 2012) and, using the percentage of times a word or phrase appears in the corpus of books (available in 8 languages: English, Spanish, German, French, Russian, Italian, Hebrew and Chinese), they investigated cultural and other trends. Their approach provides insights for different fields and issues (e.g. lexicography, collective memory, fame, censorship, epidemiology) and gives rise to an important analytical tool for social sciences and the humanities. Michel's et al. (2011) computational tool, the Google Ngram viewer (henceforth called Ngram), is available online (http://books.google.com/ngrams). Later, Lin et al. (2012) updated the corpora of the digitized books. Ngram has been recently applied in various fields, e.g. for tracking emotions in novels (Mohammad 2011, Acerbi et al. 2013), for tracking poverty enlightenment (Ravallion 2011), as a grammar checker (Nazar & Renau 2012), for studying the evolution of computing (Soper & Turel 2012) and novels (Egnal 2013), in accounting (Ahlawat & Ahlawat 2012), in poetry (Diller 2013) and for analyzing drug literature (Montagne & Morgan 2013).

Herein, we used Ngram to investigate patterns in the use of university names (i.e. frequency of times appearing in the digitized books) and related such patterns with the rankings derived from 3 different commercial systems QS, THE and THER.

MATERIALS AND METHODS

Ngram estimates the usage of small sets of phrases and produces a graph where its *y*-axis shows how a phrase occurs in a corpus of books during a particular period relative to all remaining phrases composed of the same number of words (Lin et al. 2012). The analysis is available for 1800 to 2008 (Lin et al. 2012). A detailed account of the Ngram technique is provided in Michel et al. (2011) and Lin et al. (2012), whereas a step-by-step guide for its application using examples is available online (http://books.google.com/ngrams/info#advanced).

We used Ngram for estimating the percentages of the names of the top US, Canadian and UK universities appearing in the corpus of English books during 1800 to 2008. For the US and UK we selected all the universities found in the first 20 QS positions for 2012/13 (Table 1). For the UK we also selected University of Edinburgh, which appeared in position 21. For Canada, we selected the first 4 universities appearing in the QS and THE lists (i.e. University of Toronto, McGill University, University of British Columbia and University of Alberta).

We consequently extracted the QS rankings of all of the US, UK and Canadian universities for all the years that are available (i.e. 2012/13, 2011, 2009, 2008; data are not available online for 2010) and estimated the mean annual rank for each of these universities (Table 1). We did the same using the THE and THER data for the available years (i.e. 2012/13, 2011/12, 2010/11) (Table 1). Based on the mean annual QS, THE and THER scores, we ranked the 13 US, 4 Canadian and 5 UK universities from 1 to 13, 1 to 4 and 1 to 5 (i.e. henceforth called national lists), respectively, for each of the 3 systems. We used the recent Ngram frequencies (1980 to 2000) of the US, Canadian and UK universities to rank them in terms of reputation at the national level. Although we also present the frequencies for 2000 to 2008, we did not use them for the ranking because of technical differences between the data before and after 2000 (Michel et al. 2011). We then compared the Ngram national ranks with the national QS, THE and THER rankings estimated as described above. For this, we estimated the average difference between all combinations of the national QS, THE and THER ranks for all universities examined here. The average difference was 2 and was used as a reference point for comparing the Ngram reputation rankings with those of the 3 systems (i.e. we considered that differences in national rankings between Ngram and each of QS, THE and THER were important when they were >2).

We also produced Ngram graphs for 10 European historical universities and compared their average of the lowest and highest frequency during 1980 to 2000 with the year of their establishment (taken from http://en.wikipedia.org/wiki/List_of_oldest_universities_in_continuous_operation).

Table 1. Annual and mean annual rankings for different top US, Canadian and UK universities according to QS, Times Higher Education (THE) and THE World Reputation Rankings (THER). Alberta is not listed in the top THER lists

Country/ University	Annual world university rankings										Mean annual ranking		
	QS				THE			THER			QS	THE	THER
	2012	2011	2009	2008	2012	2011	2010	2012	2011	2010	2008–2012	2010–2012	2010–2012
US													
Harvard	3	2	1	1	4	2	1	1	1	1	1.8	2.3	1.0
MIT	1	3	9	9	5	7	3	2	2	2	5.5	5.0	2.0
Yale	7	4	3	2	11	11	10	10	10	9	4.0	10.7	9.7
CalTech	10	12	10	5	1	1	2	11	11	10	9.3	1.3	10.7
Chicago	8	8	7	8	10	9	12	14	14	15	7.8	10.3	14.3
Princeton	9	13	8	12	6	5	5	7	7	7	10.5	5.3	7.0
Stanford	15	11	16	17	2	2	4	6	4	5	14.8	2.7	5.0
Columbia	11	10	11	10	14	12	18	14	15	23	10.5	14.7	17.3
Pennsylvania	12	9	12	11	15	16	19	18	19	22	11.0	16.7	19.7
Johns Hopkins	16	16	13	13	16	14	13	19	18	14	14.5	14.3	17.0
Cornell	14	15	15	15	18	24	14	17	16	16	14.8	18.7	16.3
Michigan	17	14	19	18	20	18	15	12	12	13	17.0	17.7	12.3
Duke	20	19	14	13	23	22	24	31	33	36	16.5	23.0	33.3
Canada													
Toronto	19	23	29	41	21	19	17	16	16	17	28.0	19.0	16.3
McGill	18	17	18	20	34	28	35	31	25	29	18.3	32.3	28.3
British Columbia	45	51	40	34	30	22	30	31	25	31	42.5	27.3	29.0
Alberta	108	100	59	74	121	100	127				85.3	116.0	
UK													
Oxford	5	5	5	4	2	4	6	4	6	6	4.8	4.0	5.3
Cambridge	2	1	2	3	7	6	6	3	3	3	2.0	6.3	3.0
University College	4	7	4	7	17	17	22	20	21	19	5.5	18.7	20.0
Imperial College	6	6	5	6	8	8	9	14	13	11	5.8	8.3	12.7
Edinburgh	21	20	20	23	32	36	40	46	49	45	21.0	36.0	46.7

RESULTS

The graphs produced with Ngram show trends in 2 (e.g. name-university: Stanford University) or 3 ngrams (e.g. university-of-name: University of Pennsylvania) during 1800 to 2008. The *y*-axis shows the percentage of the phrase selected when compared to all bigrams (or trigrams) contained in the corpus of the English books.

With respect to the top US universities (Fig. 1), the frequencies of all the university names examined here increased from 1800 to the 2000s with the exception of that for University of Columbia, which peaked in the 1940s and declined thereafter; Stanford University, which peaked in 1970 and slightly declined thereafter; University of Michigan, which reached a peak in late 1970s and then declined; and University of Pennsylvania, which peaked in 1980 and remained stable thereafter. The frequencies for Harvard and University of Pennsylvania were higher than those of the remaining universities during 1800 to 1920. However, Columbia University[1] before 1896 was known as Columbia College, which had frequencies that increased up to 0.0001244 in 1895, being similar to those of University of Pennsylvania for the period up to the early 1870s (graph not shown). During 1920 to 1960 the frequencies for University of Columbia were higher than the remaining ones. After 1960, University of Chicago attained higher frequencies than all the remaining universities, equaling those of Harvard for the years following the 1980s (Fig.1). The frequencies of occurrences of the 13 US universities during 1980 to 2000 were higher than 0.00019, with the exception of that for California Institute of Technology, which was ~0.000045 (Fig. 1).

We also searched for many other US universities that appear in the first 200 positions (i.e. University

[1]We also searched for Barnard College and Teachers College, both of which are affiliated with Columbia University (graphs not shown here). The frequencies of Barnard College during 1890–2008 were by 1 to 2 orders of magnitude smaller than the frequencies of Columbia University. In contrast, the frequencies of Teachers College increased exponentially from 1900 to a maximum in the early 1930s, with frequencies similar to those of Columbia University during 1927–1931, and since then declined exponentially to frequencies that were 5 to 7 times lower than those of Columbia University during 1980–2000.

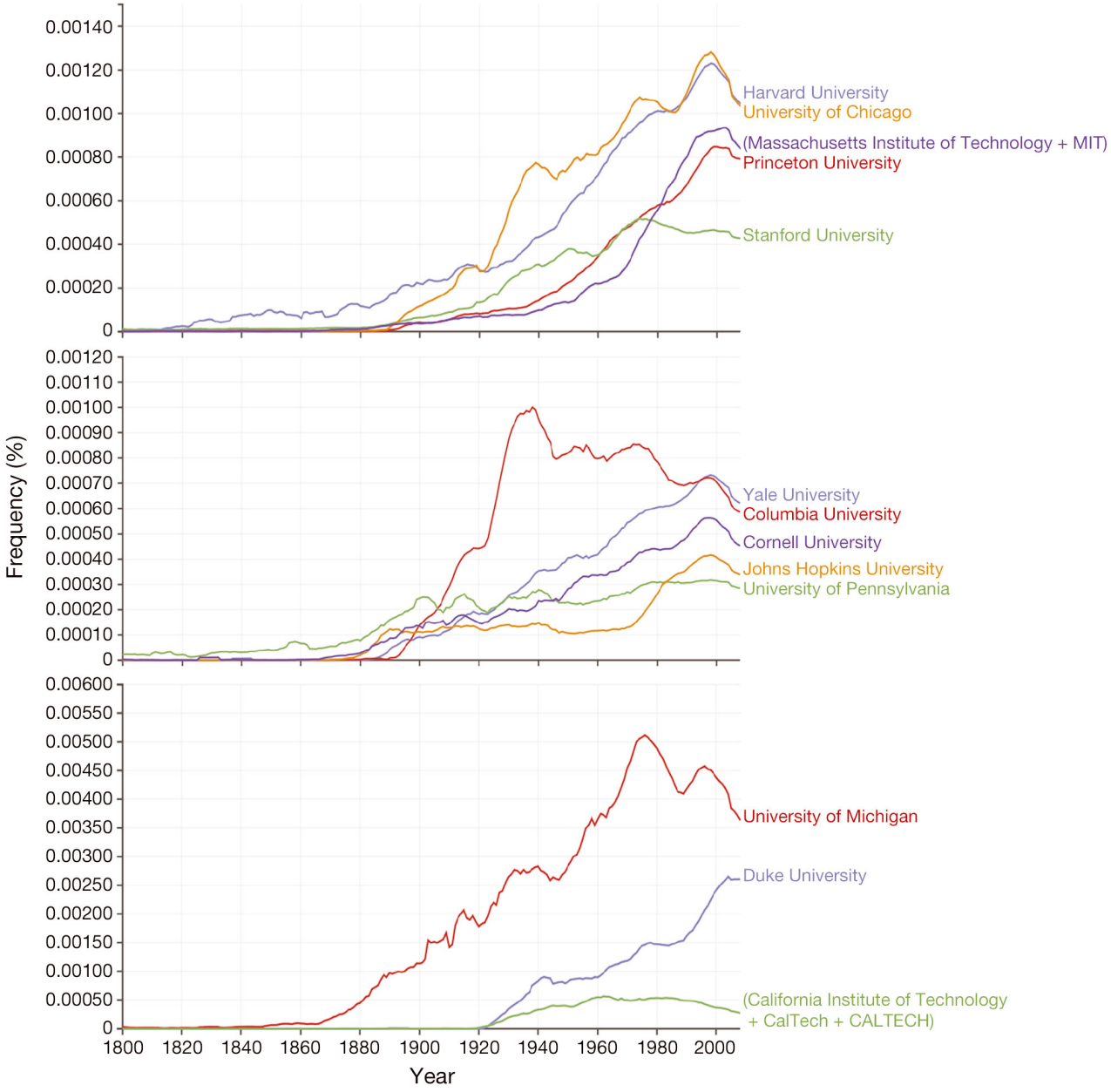

Fig. 1. Usage frequencies (relative) of the names of 13 US universities in the corpus of English books during 1800–2008

of Washington, Rice University, Boston University, Purdue University, Ohio State University, University of Southern California, Northwestern University, Brown University, University of Minnesota, University of Florida; see Fig. S1 in the Supplement at www.int-res.com/articles/suppl/e013p193_supp.pdf), all of which had, during 1980 to 2000, frequencies <0.00016, i.e. smaller than those of the top 13 US universities (but higher than that for California Institute of Technology) (Fig. 1). The only exception was the University of Minnesota with a frequency of 0.00036 in 2000 (~0.00032 for 1980 to 2000), i.e. ranked similarly with the University of Pennsylvania during this period, and University of Washington, which had an average 1980–2000 frequency of ~0.00020, i.e. similar to that of Duke University. When we searched for University of California, its frequency in the corpus of English books was higher than those of the 13 US universities, reaching 0.0015 in 2000 (with an average of ~0.0014 for 1980 to 2000). This is clearly attributed to the fact that this university includes several universities in different cities (i.e. Berkley, San Diego, Santa

Barbara, Los Angeles, San Francisco, Irvine) all of which had, however, frequencies <0.000004, with the exception of University of California, Los Angeles, which, when searched as 'UCLA', its frequency climbed up to 0.00025 in 2000 (with an average 1980–2000 frequency of ~0.00024), thus positioned higher than Duke University and California Institute of Technology but lower than the remaining 11 universities. The frequencies of the remaining University of California sites also increased when we added the frequencies for their acronyms (i.e. UCSB, UCSD, UCI, UCB, UCSF), but all frequencies were <0.00004. This additional analysis showed that the top 13 US universities examined here are generally the dominant ones in terms of frequencies with which their names appear in the corpus of English books.

We ranked the 13 universities in terms of reputation based on their recent frequencies (1980 to 2000) (Table 2). These ranks were compared with the national QS, THE and THER ranks. With the exception of Harvard and MIT, for which all rankings provided the same results, the Ngram reputation rankings differed from the QS ones for 7 universities, with individual differences ranging from 3 to 4, from the THE rankings for 9 universities, with individual differences of 3 to 8, and from the THER rankings for 8 universities, with differences of 3 to 9 (Table 2).

The mean QS and THE university rankings differed for 5 universities, by 3 to 4 positions, whereas the THE and THER rankings differed for 6 universities by 3 to 5 positions, and the QS and THER rankings for 7 universities by 3 to 6 positions (Table 2). Thus, the differences between the Ngram and the QS/THE/THER rankings were generally similar to the differences between ranking systems themselves.

Table 2. National ranks for 22 US, UK and Canadian universities developed from the mean annual ranks of QS, Times Higher Education (THE) and THE World Reputation Rankings (THER) (see Table 1) and from Ngram analysis for 1980–2000

Country/ University	National ranks			
	QS	THE	THER	Ngram
US				
Harvard	1	2	1	1
MIT	3	4	2	2
Yale	2	6	5	2
CalTech	5	1	6	7
Chicago	4	5	8	1
Princeton	6	4	4	2
Stanford	7	3	3	4
Columbia	6	8	11	2
Pennsylvania	6	9	12	5
Johns Hopkins	7	7	10	4
Cornell	7	11	9	3
Michigan	8	10	7	4
Duke	8	12	13	6
Canada				
Toronto	2	1	1	1
McGill	1	2	3	3
British Columbia	3	3	2	2
Alberta	4	4	4	4
UK				
Oxford	2	1	2	1
Cambridge	1	2	1	2
University College	3	4	4	4
Imperial College	3	3	3	5
Edinburgh	4	5	5	3

With respect to the 4 Canadian Universities (Fig. 2), their frequencies in the English corpus increased up to 1980 and then remained stable. University of Toronto and McGill University enjoyed similar frequencies up to 1920. For the years following 1920,

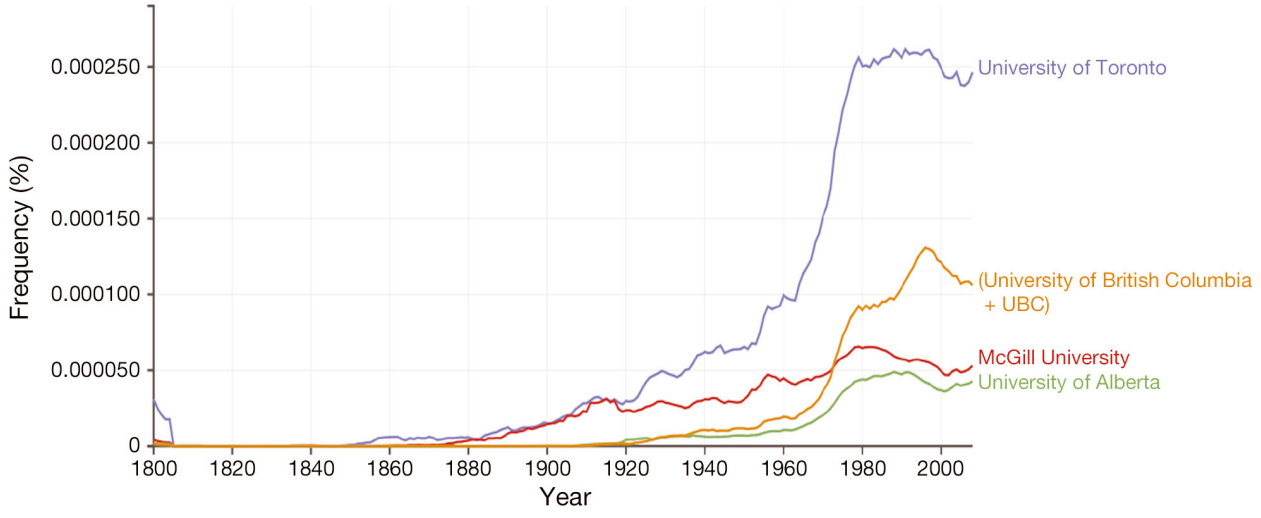

Fig. 2. Usage frequencies (relative) of the names of 4 Canadian universities in the corpus of English books during 1800–2008

University of Toronto dominated, with its frequencies in the years after 1980 (i.e. 0.00026) being 1 order of magnitude higher than those of the remaining 3 universities (Fig. 2). From the latter, McGill had higher frequencies during 1920 to early 1970s, whereas from then onwards the frequencies of University of British Columbia surpassed those of McGill. University of Alberta was characterized by the lowest frequencies throughout the period (Fig. 2). The frequencies of University of British Columbia, McGill and University of Alberta during 1980 to 2000 had ranges of 0.000093–0.00013, 0.000065–0.000049 and 0.000044–0.000049, respectively.

We also searched for other Canadian Universities that appear in various lists (i.e. Université de Montréal, University of Victoria, Dalhousie University, University of Western Ontario, McMaster University, Queen's University, University of Waterloo, University of Calgary; see Fig. S2 in the Supplement) and all had frequencies in 1980 to 2000 of <0.000034, i.e. lower than the ones presented in Fig. 2. The only exception was Queen's University, the frequency of which approached that of McGill in the early 1990s, and surpassed it in late 1990s by a small margin (i.e. 0.000062 and 0.000052, respectively). However, there is more than one Queen's University in the world. The Ngram rankings derived from the frequencies were exactly the same with those of THE and THER, whereas they differed from the QS ones, according to which McGill University is in first place and University of Toronto in second place (Table 1).

For the 5 UK universities (Fig. 3), the frequencies of Oxford and Cambridge, 2 of the oldest European universities, established in 1167 and 1209, respectively, were higher than those of the remaining universities during the whole study period. Their frequencies increased exponentially after 1920 and 1940, respectively. The frequencies of Oxford were consistently higher than those of Cambridge. The frequencies of University of Edinburgh were higher during 1800 to 1910 than in the following years. In 1980 to 2000, the frequencies of University of Edinburgh, Imperial College and University College London were by 2 orders of magnitude lower than those of Oxford and Cambridge (Fig. 3). We also searched for several other UK universities (see Fig. S3 in the Supplement) that appear in top lists (e.g. London School of Economics, University of Southampton, University of Essex, University of Glasgow, Durham University, University of Warwick, University of Lancaster), all of which had frequencies that were by 1 or 2 orders of magnitude lower than those of Cambridge and Oxford. These additional universities had also frequencies that during 1980 to 2000 were lower than those of University of Edinburgh (range: 0.000052–0.000061) and University College London (range: 0.000028–0.000088). The only exception was London School of Economics, which had frequencies ranging from 0.000086 to 0.00011, thus dominating the remaining universities after the mid 1940s but still 1 order of magnitude lower than those of Oxford and Cambridge in recent years (Fig. 3).The Ngram rankings differed by 1 or 2 positions than the other systems (Table 2) because Oxford is ranked first in Ngram and THE and second in QS and THER, whereas the opposite is true of Cambridge.

Across countries, the frequencies of Oxford were higher than those of Harvard and Chicago after 1980 and of Cambridge after 1990. The frequencies of these 2 UK universities after 1995 were 1.5 to 2 times higher than those of University of Chicago and Harvard, whereas the frequencies of the University of Toronto were 1 order of magnitude lower than those of the above 4 universities.

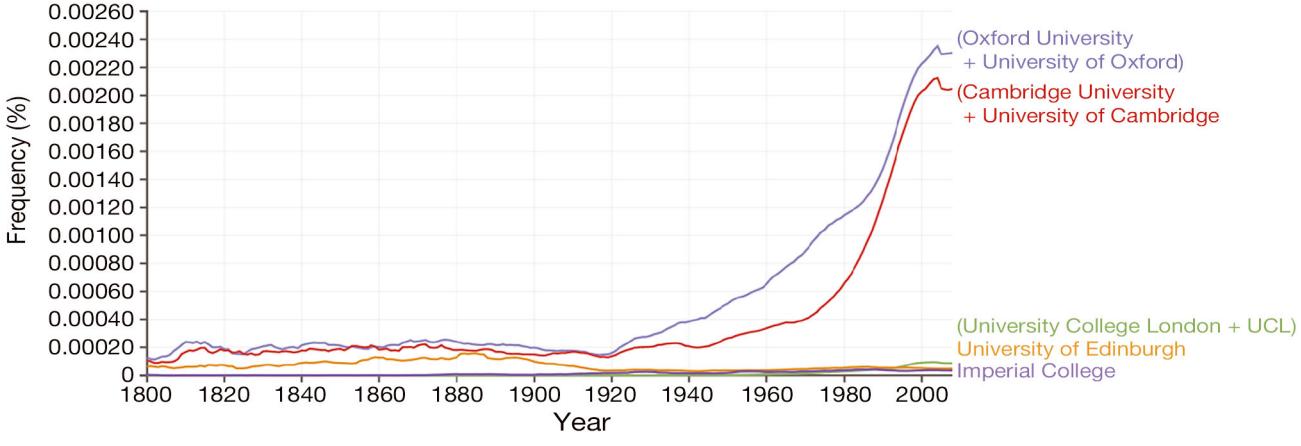

Fig. 3. Usage frequencies (relative) of the names of 5 UK universities in the corpus of English books during 1800–2008

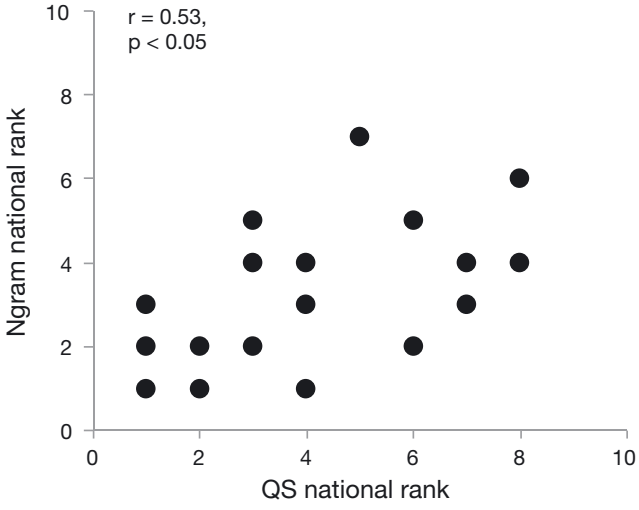

Fig. 4. Relationship between national Ngram and QS 2012/13 ranks for 22 US, UK and Canadian universities

Overall, for all the 22 US, UK and Canadian universities examined here, the national Ngram ranks were significantly correlated with the national QS (Fig. 4) and THER ones (r = 0.53 and 0.46, $p < 0.05$, respectively) but not with those of THE (r = 0.32, $p > 0.05$).

The Ngram graphs for 10 of the oldest universities in the world are shown in Fig. 5. The frequencies of these universities are 2 to 3 orders of magnitude lower than those of the US, UK and Canadian ones, which is expected given the use of the English corpus of books. What is important here is that such historical universities do appear regularly in English books, with percentages fluctuating with time. There is a positive relationship between the age of the university and its frequency in the corpus. Thus, the oldest university, University of Bologna, generally displays the highest frequencies (except during 1950 to 1970 when University of Padua attained higher frequencies), followed by the Universities of Padua, Salamanca, Naples, Coimbra, Toulouse, Siena (its frequency increased exponentially since 1970), Valladolid, Murcia and Macerata (established in 1290), which is not shown in Fig. 5 because of its very small frequency when compared to the remaining ones. Indeed, the year of establishment of these universities was negatively correlated (r = –0.82, $p < 0.05$) (Fig. 6) with their average frequency during 1980 to 2000 in the corpus of English books. It is worthy of mention here that from these 10 universities, only University of Bologna is found in the top 200 QS 2012/13 universities, whereas the Universities of Toulouse, Coimbra, Padua and Montpelier are among the top 500 QS 2012/13 (at positions from 278 to 386).

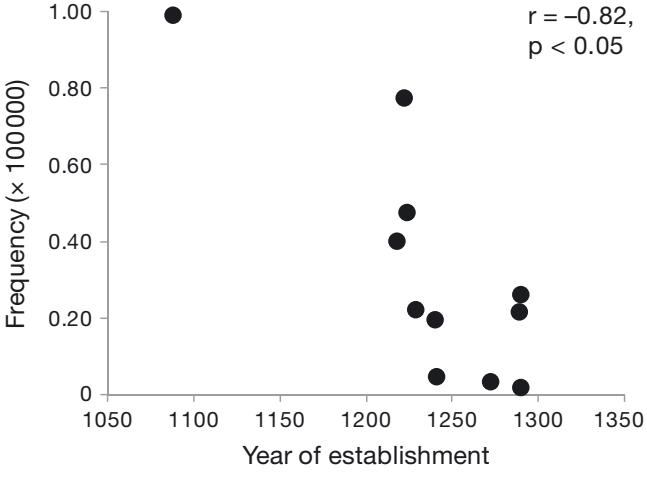

Fig. 6. Relationship between the average 1980–2000 Ngram frequency of 10 of the oldest European universities (shown in Fig. 5), as well as of University of Macerata, in the corpus of English books and their year of establishment (taken from http://en.wikipedia.org/wiki/List_of_oldest_universities_in_continuous_operation)

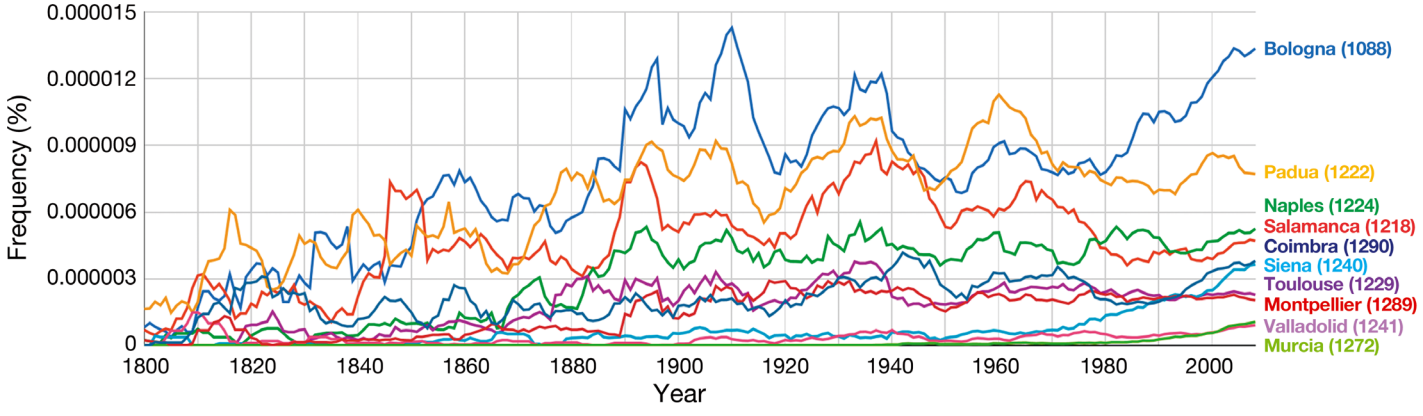

Fig. 5. Usage frequencies (relative) of the names of 10 of the oldest European universities in the corpus of English books during 1800–2008. All universities were searched as 'University of…'. Year of establishment is shown in parentheses (taken from http://en.wikipedia.org/wiki/List_of_oldest_universities_in_continuous_operation)

DISCUSSION

In this study, we used Ngram to produce the frequencies of the names of 22 US, UK and Canadian universities in the digitized corpus of English books, which is comprised by about half a trillion words (Lin et al. 2012), and studied how these frequencies changed with time (1800 to 2008). We further used the frequencies during 1980 to 2000 to establish reputation rankings for these universities. Naturally, books are only one source that can be used to study reputation, with many other sources being also important and useful (e.g. newspapers, magazines, media: Michel et al. 2011; blogs and social networks: Altmann et al. 2011, Dodds et al. 2011, Ratkiewicz et al. 2011).

Our results showed that the differences between the Ngram and the QS/THE/THER rankings for US universities are similar to the differences between the 3 ranking systems themselves, whereas the rankings for UK and Canadian universities were almost identical for the various systems (Table 2). This, together with the fact that Ngram and QS and THER national ranks were significantly correlated, clearly indicates that Ngram generally captures and reflects the reputation to the same extent that commercial rankings do, at least of the very top universities, in each country.

The within- and between-systems differences in rankings can generally be high albeit less so for the very top universities (e.g. Dichev 2001, Marginson 2007, Usher & Savino 2007, Federkeil 2009, Huang 2011, Chen & Liao 2012). The same was also true of the QS, THE and THER rankings for the years used here. For instance, from Table 1 it is evident that, with the exception of Harvard, MIT, Johns Hopkins, University of Michigan and Oxford for which the differences in mean annual ranks between QS and THE are <1, the differences for all remaining universities were from 2.6 to 31 positions. Thus, one has to wonder about the usefulness of the exact annual rank of a university (e.g. McGill University: position 18 or 32; University of Alberta: position 85 or 116) (Table 1), which reflects noise rather than news (Dichev 2001), as opposed to some index referring to a relatively long period.

Our results showed that Ngram is an easy-and-cheap-to-apply tool to approximate the reputation and 'intellectual' impact of universities over long time periods. Its reputation-generating capability, at least for top national universities, is not worse than the within- and between-systems capabilities of the commercial tools, which are generally regarded as providing 'reliable' information. However, if the reputation ranking of universities can be obtained by just typing their names in Ngram and checking their frequencies, then there is probably no need to resort to the very expensive procedures of the commercial reputation ranking systems, which take into account a large number of variables and their reputation scores of universities are practically meaningless for universities below the top 50 (Rauhvargers 2013). In addition, contrary to various indicators used in commercial ranking systems that can be 'manipulated' by institutes for climbing up the rank (e.g. see Table 1 in Hazelkorn 2009), Ngram cannot. Ngram can, thus, be promising at least for students (and their families), who make choices that are affected by rankings to an increasing extent (e.g. Sauder & Lancaster 2006, Bowman & Bastedo 2009, Hazelkorn 2009) and pay particular attention to reputation (Federkeil 2009). Naturally, student decisions on selecting a university are a multidimensional process that depends also on other factors (e.g. other reputation and prestige indicators such as tuition fees and instructional expenditure for liberal arts: Bowman & Bastedo 2009; student's economic status: Clarke 2007). Students might have additional 'educational' benefits by using the Ngram tool. For instance, they will also have a perception of the historical impact of a university, something that is not true for the short-term, volatile rankings (the earliest GUR system is available since 2003), which might mislead students when making their choice. Indeed, the 10 oldest universities examined here might not appear in top 100 lists, but historical universities have undoubtedly driven the evolution of modern universities and higher education in general. This contribution and historical perspective can be felt when someone is visiting their campuses and especially their libraries (e.g. University of Coimbra, University of Salamanca, Trinity College in Dublin).

In general, one might expect that references to old universities have decreased during the last few decades, because more and newer institutions are now competing for reputation. However, with few exceptions (e.g. Columbia University, Stanford University, University of Michigan, University of Salamanca, University of Padua: Figs. 1–3, 5) for which the frequencies consistently declined for an extended period, the frequencies of the universities examined here have generally increased with time during the last 100 yr. This is most probably explained by the fact that the increase in the number of universities competing for reputation parallels a global large increase in the references to universities.

Although people are becoming more famous nowadays than before, they are also forgotten more rapidly (Michel et al. 2011). In contrast, as mentioned above, universities are generally characterized by rather continually increasing fame, which must be attributed to the fact that universities are there forever and their fame is accumulated from generation to generation. This agrees with the positive relationship between Ngram frequency and age of universities. As universities are the productive units of scientific knowledge, this fame accumulation certainly reflects the accumulation of knowledge and thus the continually growing importance of science to the well being and future of our societies.

Our work suffers from certain biases in the estimations of frequencies. For instance, when searching university names using their acronyms, Ngram might be counting the frequency of acronyms that also refer to other entities. For example, when searching for University of California, Berkeley, as 'UCB', the corpus will obviously provide the sum of the frequencies of all the occurrences of this one ngram acronym (e.g. University of Colorado at Boulder, United Christian Broadcasters, if they are occurring), irrespectively of its actual reference. Thus, there is a risk of having a bias in the frequency count. One might need to use very sophisticated disambiguation algorithms to determine the correct reference of an acronym in a given context, and, with a limited context window of one ngram, this can be rather hard. This problem of ambiguity also applies to the case of universities that are also publishing houses. In this case, part (ranging from relatively small, e.g. University of Michigan, to large, e.g. Cambridge and Oxford) of the frequency count of the names of these universities will be because of the citations of the books by this publisher. Although the frequencies related to university publishing houses are most probably part of a university's reputation, one would need to measure the impact of works published by authors affiliated to other universities and printed by other publishing houses to make up for that extra bonus that is given to the universities with publishing houses. In that sense, this is also a source of bias that needs more complex statistical procedures, algorithms and analyses applied on the downloaded whole dataset in order to be controlled (see, e.g. Acerbi et al. 2013).

The analysis presented here might also have important cultural and historical implications, which, however, are outside the scope of this work. For instance, the frequencies of the 10 oldest European universities displayed characteristic periodicities of ~20 yr that might reflect important historical and cultural events (see Gao et al. 2012, for analysis of long-range correlations in ngram frequencies). The same is also true of the alternating patterns in terms of frequency dominance between universities (e.g. Universities of Coimbra and Toulouse: during 1800 to 1870 and 1940 to today, University of Coimbra has higher frequencies than University of Toulouse, whereas the opposite is true of 1870 to 1940). Another interesting issue is the relationship between the increasing frequencies of the University of Bologna since 1985 (Fig. 5) and the Magna Charta Universitatum Europaeum that was proposed by the University of Bologna in 1986 and the Bologna Declaration of 1999 towards the reform of Higher Education in Europe. Finally, the prominent declining pattern in the frequency for Columbia University after 1940 (Fig. 1) may be related to particular historical facts that might have affected its reputation (e.g. atom research and the Manhattan Project in the 1940s; intense student activism in the 1960s resulting in the President's resignation; links between the university and the Vietnam War; Columbia College did not admit women until 1983, see http://en.wikipedia.org/wiki/Columbia_University, section Columbia University, 1896–present [accessed 19 August 2013]).

Acknowledgements. The authors thank C. Apostolidis and 2 anonymous reviewers for their valuable comments and suggestions.

LITERATURE CITED

Acerbi A, Lampos V, Garnett P, Bentley RA (2013) The expression of emotions in 20th Century books. PLoS ONE 8(3):e59030

Ahlawat S, Ahlawat S (2012) An innovative decade of enduring accounting ideas as seen through the lens of culturomics: 1900–1910. American Institute of Higher Education, 7th Int Conf, Williamsburg, VA, 7–9 March 2013, p 8–19

Altmann EG, Pierrehumbert JB, Motter AE (2011) Niche as a determinant of word fate in online groups. PLoS ONE 6(5):e19009

Bowman NA, Bastedo MN (2009) Getting on the front page: organizational reputation, status signals, and the impact of US News and World Report on student decisions. Res Higher Educ 50:415–436

Buela-Casal G, Gutiérrez-Martínez O, Bermúdez-Sánchez MP, Vadillo-Muñoz O (2007) Comparative study of international academic rankings of universities. Scientometrics 71:349–365

Chen K, Liao P (2012) A comparative study on world university rankings: a bibliometric survey. Scientometrics 92: 89–103

Clarke M (2007) The Impact of higher education rankings on student access, choice, and opportunity. High Educ Eur 32:59–70

Dichev I (2001) News or noise? Estimating the noise in the US News university rankings. Res Higher Educ 42: 237–266

Diller HJ (2013) Culturomics and genre: wrath and anger in the 17th Century. In: McConchie RW, Juvonen T, Kaunisto M, Nevala M, Tyrkkö J (eds) Selected Proc 2012 Symp New Approaches English Historical Lexis (HELLEX 3), Cascadilla Proceedings Project, Somerville, MA, p 54–65

Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM (2011) Temporal patterns of happiness and information in a global social network: hedonometrics and twitter. PLoS ONE 6(12):e26752

Egnal M (2013) Evolution of the novel in the United States: the statistical evidence. Soc Sci Hist 37:231–254

Enserink M (2007) Who ranks the university rankers? Science 317:1026–1028

Federkeil G (2009) Reputation indicators in rankings of higher education institutions. In: Kehm BM, Stensaker B (eds) University rankings, diversity, and the new landscape of higher education. Sense Publishers, Boston, MA, p 19–33

Gao J, Hu J, Mao X, Perc M (2012) Culturomics meets random fractal theory: insights into long-range correlations of social and natural phenomena over the past two centuries. J R Soc Interface 9:1956–1964

Harvey L (2008) Rankings of higher education institutions: a critical review. Qual High Educ 14:187–207

Hazelkorn E (2009) Rankings and the battle for world-class excellence: institutional strategies and policy choices. Higher Educ Manag Policy 21:1–22

Hazelkorn E (2013) How rankings are reshaping higher education. In: Climent V, Michavila F, Ripolles M (eds) Los rankings univeritarios: mitos y realidades. Ed Tecnos, p 1–8

Huang MH (2011) A comparison of three major academic rankings for world universities: from a research evaluation perspective. J Lib Inf Stud 9:1–25

Jarocka M (2012) University ranking systems: from league table to homogeneous groups of universities. World Academy of Science. Eng Technol 66:800–805

Lin Y, Michel JB, Aiden EL, Orwant J, Brockman W, Petrov S (2012) Syntactic annotations for the Google Books Ngram corpus. Proc 50th Annu Meet Assoc Comput Linguistics, Vol 2: Demo Pap (ACL '12)

Marginson S (2007) Global university rankings: implications in general and for Australia. J High Educ Policy Manag 29:131–142

Michel J-B, Shen YK, Aiden AP, Veres A and others (2011) Quantitative analysis of culture using millions of digitized books. Science 331:176–182

Mohammad S (2011) From once upon a time to happily ever after: tracking emotions in novels and fairy tales. Proc 5th ACL-HLT Workshop Language Technol Cult Heritage, Soc Sci, Human (LaTeCH), Portland, OR, 24 June 2011, p 105–114. Association for Computational Linguistics, Stroudsburg, PA

Montagne M, Morgan M (2013) Drugs on the internet, Part IV: Google's Ngram Viewer analytic tool applied to drug literature. Subst Use Misuse 48:415–419

Nazar R, Renau I (2012) Google books N-gram corpus used as a grammar checker. Proc EACL 2012 Workshop Comput Linguistics Writing, Avignon, 23 April 2012, p 27–34. Association for Computational Linguistics, Stroudsburg, PA

Ratkiewicz J, Conover MD, Meiss M, Goncalves B, Flammini A, Menczer F (2011) Detecting and tracking political abuse in social media. Proc 5th Int AAAI Conf Weblogs Soc Media, ICWSM'11, p 297–304. Association for the Advancement of Artificial Intelligence Press, Palo Alto, CA

Rauhvargers A (2011) Global university rankings and their impact. Eur Univ Assoc, Brussels, available at www.eua.be

Rauhvargers A (2013) Global university rankings and their impact: Report II. Eur Univ Assoc, Brussels, available at www.eua.be

Ravallion M (2011) The two poverty enlightenments: historical insights from digitized books spanning three centuries. Poverty Public Policy 3:1–46

Sauder M, Lancaster R (2006) Do rankings matter? The effects of US News & World Report rankings on the admissions process of Law Schools. Law Soc Rev 40: 105–134

Soper DS, Turel O (2012) An n-gram analysis of communications 2000–2010. Commun ACM 55:81–87

Usher A, Savino M (2007) A global survey of university ranking and league tables. High Educ Eur 32:5–15

Van Vught F (2008) Mission diversity and reputation in higher education. High Educ Policy 21:151–174

Williams R (2008) Methodology, meaning, and usefulness of trankings. Aust Univ Rev 50:51–58

Editorial responsibility: Penny Kuhn, Oldendorf/Luhe, Germany

Submitted: June 21, 2013; Accepted: September 11, 2013
Proofs received from author(s): October 23, 2013