Een systeem voor citatieanalyses in de praktijk

Onderzoeksevaluaties aan de universiteit spelen een steeds belangrijker rol. Standaard onderdeel van deze evaluaties zijn citatieanalyses, die de impact van publicaties van onderzoekers of onderzoeksgroepen kwantitatief in beeld brengen. Bibliotheek Wageningen UR krijgt de laatste jaren steeds vaker het verzoek om deze analyses uit te voeren. In dit artikel wordt uiteengezet hoe dat gaat.

Wouter Gerritsma

Sinds de jaren tachtig is Bibliotheek Wageningen UR al betrokken bij citatieanalyses van onderzoekers. Indertijd werd gebruikgemaakt van de Science Citation Index bij Dialog. Dit gebeurde op kleine schaal. Grotere onderzoeken in het kader van de periodieke onderzoeksevaluaties werden uitbesteed aan het Centrum voor Wetenschaps- en Technologie-Studies (CWTS), de citatiegoeroes van Nederland (van Raan 1996; van Raan 2004). Sinds het begin van de jaren 2000 beschikt de bibliotheek over de webversie van de Science Citation Index, Web of Science (WoS) en zijn citatieanalyses in feite uit te voeren door iedere onderzoeker aan de universiteit. Sindsdien is het aantal verzoeken aan de bibliotheek om zulke analyses uit te voeren alleen maar toegenomen. Een belangrijke ontwikkeling die hieraan heeft bijgedragen is dat de bibliotheek sinds een aantal jaren ook beschikt over de Essential Science Indicators (ESI). Met deze database kan de methode van citatieanalyses zoals ontwikkeld door het CWTS nauwkeuriger worden gevolgd. Exact dezelfde analyses als die van het CWTS zijn met de thans beschikbare databases nog niet mogelijk voor een bibliotheek.

Essential Science Indicators

Tot voor kort waren alleen instituten voor bibliometrisch of scientometrisch onderzoek in staat om meerwaarde te leveren bij citatieanalyses omdat zij de beschikking hebben over alle data van de citatie-indexen van Thomson Scientific. Deze instituten, zoals het CWTS, zijn daardoor in staat om analyses van deze volledige dataset uit te voeren. Bij de webversie van WoS zijn zulke exercities nog een utopie. Daarnaast kunnen instituten als het CWTS verbeteringen doorvoeren in de naamgeving van onderzoekers of onderzoeksgroepen bij de primaire data. De resultaten van deze opschoningacties en de analyses van de totale dataset worden vervolgens vergeleken met de citatiescores van een specifiek instituut, een onderzoeksgroep of individuele onderzoeker.

De meeste bibliotheken beschikken niet over dit soort datasets of over mogelijkheden om dergelijke diepgravende analyses uit te voeren. Sinds een paar jaar verkoopt Thomson Scientific daarom de ESI database. Dit is een analytische database die is gebaseerd op de Science Citation Index en het mogelijk maakt om citatiedata van publicaties die verschillen in leeftijd of onderzoeksveld, vergelijkbaar te maken. ESI is gebaseerd op de tijdschriftenset die ook wordt gebruikt voor de Journal Citation Reports (JCR) en Web of Science. De data in de ESI zijn een analyse over de afgelopen tien jaar plus het huidige jaar in opbouw. ESI geeft ranglijsten voor publicaties en citaties van instituten en universiteiten, onderzoekers, landen, en tijdschriften. Daarnaast geeft het iedere twee maanden een overzicht van de meest geciteerde artikelen en de zogenaamde *hot papers*. De laatste zijn relatief jonge artikelen, minder dan twee jaar oud, die buitenproportioneel vaak geciteerd worden (Small, 2004).

Voor ons doel van citatieanalyses zijn de *baselines* van ESI essentieel. Die geven voor 22 verschillende wetenschapsvelden het verloop van het gemiddelde aantal citaties van een artikel weer, plus het aantal citaties van de 10%, 1% en 0,1% meest geciteerde artikelen. De baselines kunnen per wetenschapsveld nogal van elkaar verschillen. Een illustratie van het verloop van de baselines, en citaties van de top 10% artikelen voor landbouw en voedingswetenschappen en biomoleculaire en biochemisch wetenschappen geeft

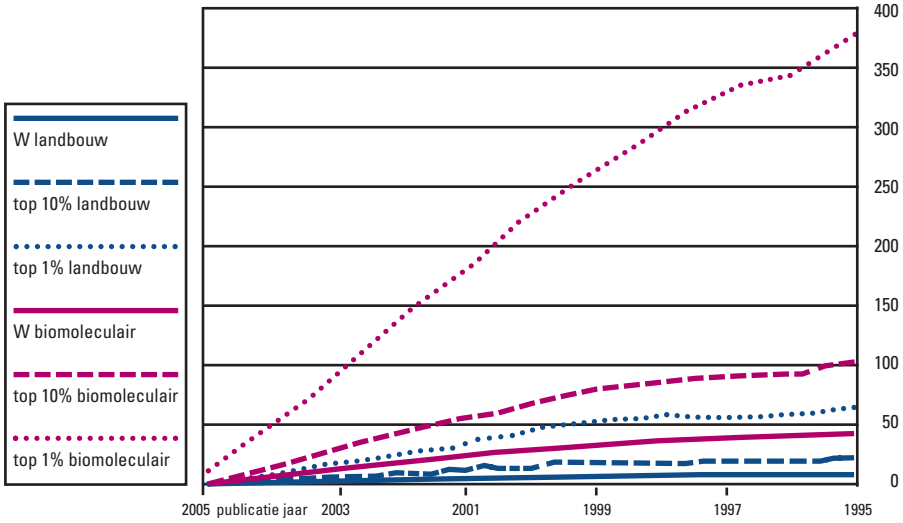
komt tekening

figuur op p. XX.
Aan de hand van deze baselines kunnen citatiedata per wetenschapsveld en per jaar worden gerelateerd aan het wereldgemiddelde. We kunnen uitrekenen hoever een artikel onder of boven het wereldgemiddelde scoort. Daarnaast kan worden aangeven of een artikel behoort tot de top 10% of de top 1% meest geciteerde artikelen in dat veld.

Zelfcitaties

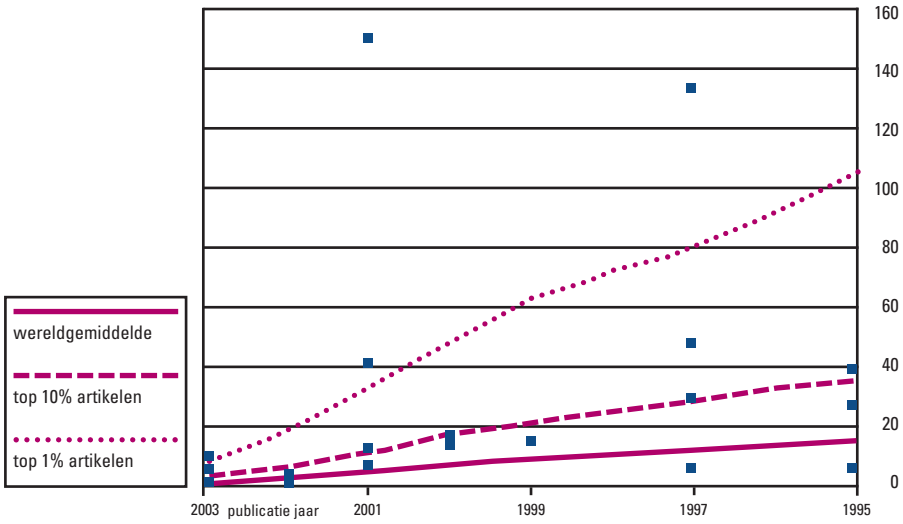
Met het gebruik van WoS en ESI lijken we goed op weg om een volwaardige citatieanalyse af te leveren. Er moet echter nog wel aandacht worden besteed aan zelfcitaties. Op het eerste gezicht lijkt het tamelijk eenvoudig om citatiedata te corrigeren voor zelfcitaties. Waneer P. Jansen zichzelf citeert is dat een eenvoudige ingreep. Wanneer P. Jansen samen met J. Pietersen een artikel schrijft en we onderzoeken de citatie-impact van Jansen, dan is een citatie door Pietersen naar hun gezamenlijke artikel even goed een zelfcitatie als wanneer die verwijzing van zijn co-auteur was gekomen. Echter de andere artikelen van Jansen waar Pietersen naar verwijst, maar die niet mede door hem zijn geschreven, tellen wel weer als een citatie. In het geval van twee auteurs is een en ander nog wel te corrigeren. Bij artikelen met twintig auteurs of meer, die in de moleculaire biologie heel gewoon zijn, wordt dit echter een zeer complexe materie. En helemaal wanneer je nog eens rekening moet houden met eventuele naamsvarianten van de diverse auteurs.
Instituten zoals het CWTS, die van citatieanalyses hun bestaan hebben gemaakt, voeren dit soort correcties minutieus door. Helaas zijn de baselines zoals we die uit ESI halen niet gecorrigeerd voor zelfcitaties. Waneer we vergelijkingen van citatiedata met de baselines van ISI maken, moeten we de citatiedata daarom niet corrigeren voor zelfcitaties.

Citatieverloop en aantal citaties in 2006
Landbouw en biomoleculair



Het citatieverloop van publicaties in de landbouw- en voedingswetenschappen (landbouw) en de biomoleculaire wetenschappen en biochemie (biomoleculair). De lijnen geven het citatieverloop voor het wereldgemiddelde, de top 10% en de top 1% artikelen weer

Aantal citaties in 2004
van artikelen van een auteur in een wetenschapsveld



Het aantal citaties van artikelen van een auteur in een wetenschapsveld, met de baseline voor het wereldgemiddelde en de aantallen citaties voor de top 10% en 1% artikelen

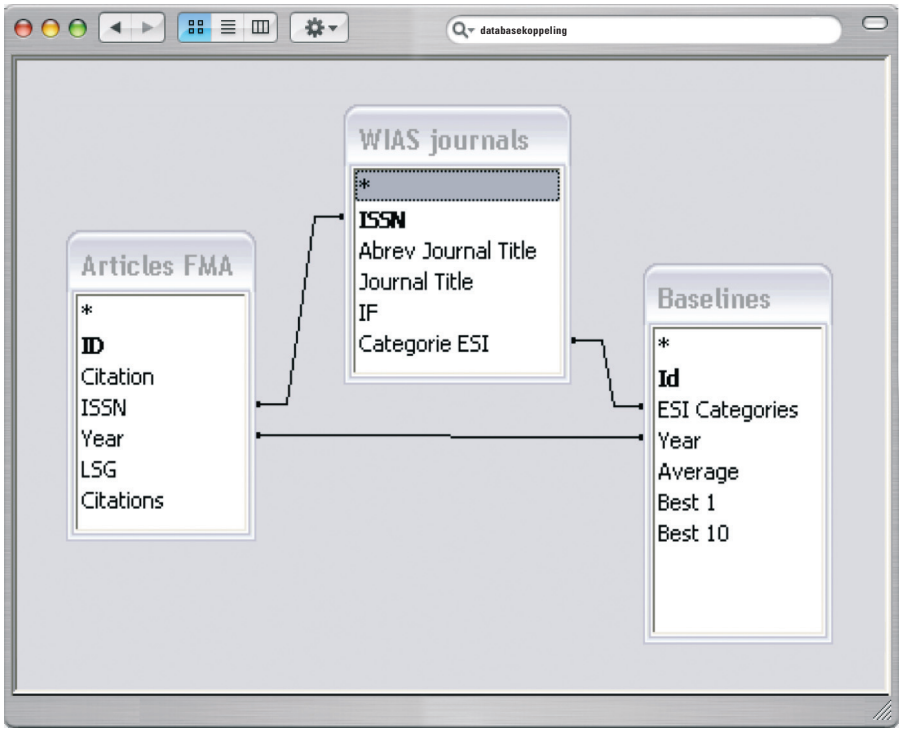
De invloed van zelfcitaties is de laatste jaren een onderwerp van veel onderzoek (Glänzel and Thijs 2004; Thijs and Glänzel 2006). Het lijkt erop dat we kunnen aannemen dat binnen instituten en onderzoeksgroepen in hetzelfde onderzoeksveld er over het algemeen een zelfde soort citatiegedrag is (Glänzel, Debackere et al. 2006). De mogelijkheid voor het toepassen van correcties voor zelfcitaties staat op ons wensenlijstje, maar echt noodzakelijk lijkt het niet wanneer we naar de onderzoeksgroepen kijken.

In de praktijk

De contacten voor opdrachten van citatieanalyses lopen meestal via de secretaris van een onderzoeksschool. De meest concrete vraag die de bibliotheek in dat geval stelt is die naar een complete lijst met namen van de deelnemers van de te onderzoeken groep. Het klinkt gek, maar dat levert meestal de nodige problemen. Zelfs wanneer een onderzoeksregistratiesysteem als Metis volledig op orde is blijken er altijd weer verrassingen voor de dag te komen. Terwijl Metis juist in het leven is geroepen om alle onderzoekers aan de universiteit te registreren en hun output aan wetenschappelijke publicaties in vast te leggen. Maar wanneer er een eerste versie van een namenlijst is uitgedraaid, komen er vervolgens namen bij en vallen er even later net zo makkelijk weer namen van de lijst af.

Voor de bibliotheek is de eerste stap om van die namenlijst in Metis de bekende naamsvarianten te achterhalen. Het record staat thans op een naam met acht verschillende synoniemen. Voor al die namen en hun varianten wordt Web of Science uitputtend doorzocht, waarbij dus ook weer rekening wordt gehouden met de idiosyncrasie van WoS.

Bibliografische gegevens van de gevonden artikelen worden als tekstbestand gedownload, waarbij ervoor wordt gezorgd dat ook het ISSN van de tijdschriften wordt overgehaald. Daarnaast wordt in elk geval ook het adres en het document type meegenomen. Deze records worden in eerste instantie opgeslagen als textfile en vervolgens met een aangepast filter ingelezen in EndNote. In EndNote beschikken we over zeven verschillende aanpasbare velden om de diverse codes voor onderzoeksgroepen aan te brengen.



De koppeling van de drie verschillende databases. Het publicatiejaar, het wetenschapsveld (ESI category) en tijdschrift ISSN zijn de sleutels voor de koppeling

‘Slechts van een beperkt aantal tijdschriften, zoals Nature, Science en PNAS wordt de indeling naar wetenschapsveld gemaakt op artikelniveau’

EndNote is voor Bibliotheek Wageningen UR een logische keuze omdat het sterk is in de bibliografische output waarmee wordt gewerkt, en voldoende mogelijkheden geeft om additionele velden toe te kennen met eigen codes.

Drie bestanden

Deze ruwe lijst met publicatie- en citatiegegevens wordt naar alle onderzoekers gestuurd met het verzoek deze te controleren. Dit is een periode waarin we uitgebreid mailen met de onderzoekers, om methodieken nogmaals toe te lichten en twijfels omtrent sommige artikelen en namen uit te sluiten. De EndNote-database wordt gecorrigeerd en daar waar nodig aangevuld. Vanuit EndNote maken we via een speciale style een download met de standaard bibliografische gegevens per artikel, het aantal citaties, de ISSN van het tijdschrift en de indeling van de te onderzoeken groepen. Deze gegevens worden ingelezen in een Access-database.

Een tweede bestand dat een belangrijke rol speelt is de tijdschriftindeling die ESI hanteert. Het gros van de tijdschriften is verdeeld over 22 verschillende wetenschapsvelden. Slechts van een beperkt aantal tijdschriften, zoals Nature, Science en PNAS wordt de indeling naar wetenschapsveld gemaakt op artikelniveau. Het klinkt vreemd maar in ESI is de indeling naar wetenschapsveld in eerste instantie lastig te achterhalen, terwijl deze in In-Cites (in-cites.com/) wel voorhanden is. Wat echter ontbreekt in zowel de helpfiles van ESI als bij de indeling in In-Cites zijn de ISSN-nummers van de tijdschriften. In de loop der jaren hebben we dat echter volledig ingevuld, zodat we beschikken over een tabel met tijdschrifttitel, het ISSN, en het wetenschapsveld waarin het is ingedeeld.

Het derde bestand dat een rol speelt is dat waarin de baselines zijn vastgelegd. Deze worden iedere twee maanden in ESI geüpdatet. De data voor de baselines worden naar een Excel-sheet overgehaald en daar bewerkt om te kunnen koppelen aan de twee andere bestanden. Wanneer deze drie bestanden via het ISSN en de indeling in wetenschapsvelden aan elkaar gekoppeld zijn (zie figuur op XX) kunnen de werkelijke vergelijkingen en analyses gemaakt worden.

Naast de aantallen publicaties en citaties en gemiddeld aantal citaties per artikel geven we een aantal additionele indicatoren. De belangrijkste is de relatieve impact. Dit is de verhouding van het aantal citaties ten opzichte van het wereldgemiddelde. Een relatieve impact van 1,5 is dus 150% van het wereldgemiddelde. We wijken in de berekeningen van de relatieve impact iets af van de methode waarmee de ‘crown indicator’ door het CWTS berekend wordt. Daarnaast tellen we ook de artikelen die behoren tot de top 10% en de top 1% van meest geciteerde artikelen in de wereld.

Resultaten

Om een en ander te verduidelijken geeft de tabel op XX als voorbeeld de geanonimiseerde resultaten van een citatieanalyse van vijf kandidaten voor een leerstoel, ten behoeve van een benoemingsadviescommissie.

Van de kandidaten voor de leerstoel springen kandidaat B en D er in positieve zin uit. Kandidaat B heeft nog niet hetzelfde aantal artikelen (mee)geschreven als A, C en D maar de relatieve impact is het hoogst. Dit beeld wordt verder aangevuld met zeer groot aantal artikelen dat binnen de 10% meest geciteerde artikelen

Tabel 1. Voorbeeld van de resultaten van een citatie analyse voor een benoemings adviescommissie

Auteur	# Artikelen	# Citations	Relatieve	RI	RI	# papers	# papers
	1994-2003		Impact (RI)	1994-1998	1999-2003	top 10%	top 1%
A	80	1565	1,64	1,76	1,52	4	2
B	65	498	1,93	1,84	1,95	17	1
C	93	972	1,15	1,39	0,9	8	0
D	88	1886	1,86	1,69	1,94	16	3
E	57	346	0,75	0,58	0,83	3	0

Resultaten van een citatieanalyse voor een benoemingsadviescommissie

valt. Kandidaat D heeft op een na de meeste artikelen (mee)geschreven en laat ook een goede progressie zien in relatieve citatie impact over de twee onderzochte periodes. Het aantal artikelen dat tot de top 1% en de top 10% van meest geciteerde papers behoort bevestigt de kwaliteit ten opzichte van de andere kandidaten. Een tabel als deze geeft voor een benoemingsadviescommissie voldoende stof voor discussie.

In tabel 2 worden als voorbeeld de gegevens getoond van een onderzoeksinstituut met drie verschillende onderzoekprogramma's. Deze hebben elk hun onderlinge overeenkomsten en verschillen. Alle drie zijn sterk op het gebied van landbouw en voeding, terwijl programma 1

Alternatieven

Het monopolie op citatiedata van Thomson Scientific (voorheen ISI) komt steeds meer onder druk te staan. Google Scholar wordt door de wetenschappers vaak aangedragen als een alternatief waarin meer citaties te vinden zijn, maar aan de betrouwbaarheid van Google Scholar kleven grote bezwaren. Naast Google Scholar zijn in de wereld van de betaalde databases thans goede alternatieven aanwezig. Zoals Scopus, PsychInfo en Scifinder of Chemical Abstracts (CA). Scopus is ongetwijfeld in potentie de grootste concurrent, omdat het hier een breed georiënteerde bibliografie betreft die een groter aantal tijdschriften dekt dan Web of Science. PsychInfo en SciFinder/CA zijn ieder specifieke databases voor een vakgebied die op hun eigen terrein zeer goede citatiedata bieden. De implementatie van deze functionaliteit laat echter een en ander te wensen over. Naast deze betaalde bibliografieën zijn er op het web talloze alternatieven die volop in ontwikkeling zijn. Hieronder volgt een overzicht van belangrijkste alternatieven.

Citebase • www.citebase.org/
Gebaseerd op e-prints software. Geeft zowel downloads als citaties van artikelen die in repositories met het OAI-PMH protocol geharvest worden

Citeseer • citeseer.ist.psu.edu/
Citatie database op het gebied van computers en informatietechnologie, ontwikkeld aan Penn State University in samenwerking met NEC

SmealSearch • smealsearch2.psu.edu/index.html/
Gebaseerd op de software van Citeseer, maar dan voor het domein van de business literatuur.

Scitation • scitation.aip.org/
Is onderdeel van de American Institute of Physics en dekt naast de eigen tijdschriften een aantal tijdschriften van kleinere society uitgevers in het veld van de natuurkunde.

Meer alternatieven worden gegeven in (Roth 2005). Het verkrijgen van citatiedata is echter slechts één punt. Een analyse van alle citatiedata om tot goede baselines te komen is in geen van de in deze box genoemde indexen tot nu toe uitgevoerd.

Tabel 2. De relatieve impact van drie onderzoeksgroepen

	Alle groepen	Groep 1	Groep 2	Groep 3
Landbouw & voiding	3,82	3,86	3,87	3,60
Biologie & biochemie	0,91	1,55	0,44	1,09
Chemie	1,76		1,76	
Geneeskunde	1,73	1,81	1,11	
Microbiologie	1,70	0.57		1,73
Gemiddelde impact	2,06	2,08	2,26	1,84

De relatieve impact van drie onderzoeksgroepen van een instituut, uitgesplitst naar de verschillende wetenschapsvelden waarin ze actief zijn

'In EndNote beschikken we over zeven verschillende aanpasbare velden om de diverse codes voor onderzoeksgroepen aan te brengen

een belangrijk accent heeft in de geneeskunde, programma 2 in de chemie en programma 3 in de microbiologie. De relatieve impact van een onderzoeksinstituut als geheel is bijzonder goed te noemen, met een gemiddelde citatie-impact van ongeveer twee keer het wereldgemiddelde.

Als laatste voorbeeld van de resultaten wordt in figuur 3 een aantal artikelen van een onderzoeker getoond die zijn gepubliceerd in tijdschriften die behoren tot het wetenschapsveld ecologie. Het aantal citaties per artikel medio 2004 is aangegeven als stip. De lijnen zijn de baselines voor dit wetenschapsveld. Tot de top 1% van artikelen qua aantal citaties behoren dus vier artikelen. Drie artikelen vallen onder het wereldgemiddelde.

Wat het verder oplevert

De citatieanalyses zijn voor de bibliotheek vaak grote klussen die op basis van bestede uren worden vergoed. Maar afgezien van deze financiële vergoeding zijn er een paar sterke pluspunten die het interessant maken om deze klussen aan te trekken. Met deze exercities gaan alle publicaties een aantal keren door je handen, je krijgt zo een zeer goede indruk van de tijdschriften waarin en hoe vaak gepubliceerd wordt. Je ziet ook beter waar onderzoekers mee bezig zijn en waarover ze publiceren. Daarnaast treden de bibliotheekmedewerkers in dialoog met alle onderzoekers en zij zien dat de bibliotheek meer is dan alleen maar een verzameling boeken en tijdschriften. Tijdens die dialoog blijkt dat er heel vaak advies gegeven kan worden over alternatieve tijdschriften, of publiceren in bijvoorbeeld open access tijdschriften. Kortom de citatieanalyse is altijd een goede binnenkomer bij een zeer belangrijke groep gebruikers. Kwamen onderzoekers vroeger als vanzelfsprekend naar de bibliotheek, met de sterke ontwikkeling van de digitale bibliotheek is een citatieanalyse een goede reden om de onderzoeker zelf op te zoeken.

Een punt van aandacht voor onderzoekers en het management van groepen naar aanleiding van de citatieanalyses is de naamgeving. Vaak wordt onomwonden duidelijk gemaakt dat individuele onderzoekers, onderzoeksgroepen of de universiteit onder meerdere naamsvarianten door het leven gaan. Het betreft in dat geval niet alleen vrouwelijke AIO's die beginnen te publiceren onder hun meisjesnaam, en vervolgens verder publiceren met de naam van hun partner. Het blijft wat dit betreft vreemd dat een communicatieafdeling van de universiteit wel oog heeft voor de logo's en het briefpapier dat er gebruikt wordt, maar geen duidelijke richtlijnen heeft voor naamgeving en adressen zoals die gebruikt worden in wetenschappelijke artikelen. Die laatste vormen toch een van de meeste belangrijke outputs van een universiteit. <

Wouter Gerritsma is informatiespecialist plantenwetenschappen bij Bibliotheek Wageningen UR en blogt over dit soort onderwerpen op www.wowter.nl/blog.

Literatuur

] Glänzel, W., K. Debackere, B. Thijs & A. Schubert (2006). A concise review on the role of author self-citations in information science, bibliometrics and science policy. *Scientometrics* **67**(2): 263-277.

] Glänzel, W. & B. Thijs (2004). The influence of author self-citations on bibliometric macro indicators.

] *Scientometrics* **59**(3): 281-310.

] Van Raan, A.F.J. (1996). Advanced bibliometric methods as quantitative core of peer review based evalaution and foresight excerises. *Scientometrics* **36**(3): 397-420.

] van Raan, A.F.J. (2004). Measuring Science. Capita Selecta of Current Main Issues. *Handbook of Quantitative Science and Technology Research*. H.F. Moed, W. Glänzel & U. Schmoch. Dordrecht, Kluwer Academic Publishers: 19-50.

] Roth, D.L. (2005). The emergene of competitors to the Science Citation Index and the Web of Science. *Current Science* **89**(9): 1531-1535.

] Small, H. (2004). Why authors think their papers are highly cited. *Scientometrics* **60**(3): 305-316.

] Thijs, B. & W. Glänzel (2006). The influence of author self-citations on bibliometric meso-indicators. The case of european universities. *Scientometrics* **66**(1): 71-80.