GÖTEBORGS UNIVERSITET
GÖTEBORGS UNIVERSITETSBIBLIOTEK

Notiser – april 2012

Reliabilitetsaspekter på bibliografisk information vid utvärdering av forskningsprestationer

Bo Jarneving, Digitala tjänster

I denna rapport granskas relationen mellan bristfällig bibliografisk information och utvärdering av forskningsprestationer. Med utgångspunkt i ett obundet slumpmässigt urval påvisas och diskuteras diskrepans mellan adressinformation i GUP och i WoS.

Inledning

I takt med att användningen av publikations- och citeringsdata för utvärdering av vetenskapliga prestationer ökar, har reliabilitetsaspekten på sådana data och information blivit alltmer väsentlig. Göteborgs universitet (GU) har sedan 2004 regelmässigt sammanställt publikationsdata i en egen publikationsdatabas (GUP). Denna databas används för utvärdering av GU:s forskning utifrån olika behov och senast 2011 genomfördes en omfattande utvärdering av GU:s forskning, där också publikationsdata från GUP användes (RED10).

Sedan 2010 används publikations- och citeringsdata vid fördelningen av statliga medel till universitet och högskolor. I de fallen citeringsbaserade indikatorer används i utvärderingssammanhang krävs kompletterande data från någon internationell, multidisciplinär citeringsdatabas. Vanligen används citeringsdata från Thomson Reuters *Web of Science* (WoS). Det är av avgörande betydelse att den bibliografiska informationen är korrekt, d.v.s., att publikationer och citeringar kan tillskrivas rätt universitet. Ett känt problem i detta sammanhang är de många stavningsvarianterna på universitetsadresser.

Publikations-och citeringsdata används också vid fakultetsvis fördelning av medel inom GU varje budgetår. För en del fakulteter baseras fördelningen av medel enbart på publikationsdata medan andra utvärderas med hjälp av både publikations- och citeringsdata. Oberoende av vilket gäller att informationen i bibliografiska beskrivningar av publikationer måste vara korrekt så att en forskningsprestation i form av publicering associeras till rätt instans. När citeringsdata används vid forskningsutvärdering kompliceras situationen ytterligare eftersom den bibliografiska informationen i WoS måste matchas mot bibliografisk data i GUP. Kvaliteten på den bibliografiska datan i GUP är avgörande för säker matchning. Universitetsbiblioteket arbetar därför kontinuerligt med att säkerställa kvaliteten genom en s.k. granskningsfunktion där felaktig information kan rättas till. I linje med detta är det av vikt att utvärdera reliabiliteten i den information som ligger till grund för framtagandet av indikatorer för utvärdering.

Syfte

Avsikten med denna rapport är att få kunskap om eventuella brister i matchningen mellan GUP-data och WoS-data samt hur affilieringen till universitet/organisation i GUP överensstämmer med affilieringen i WoS. Utifrån kända affilieringsproblem i samband med registrering av publikationer i GUP skapades för ändamålet fyra ömsesidigt uteslutande kategorier (se Data och metod).

Avgränsningar

Undersökningen omfattade observationsperioden 2005-2011. Populationen omfattade endast publikationer sådana att de också var indexerade i WoS.

Data och metod

Populationen utgjordes av 14 588 distinkta publikationer där varje sådan uppfyllde följande villkor:

- Registrerad i GUP
- Minst en författare av publikationen är affilierad med GU i GUP
- Indexerad i WoS

Ur populationen drogs ett obundet slumpmässig urval där *n* = 200 i april, 2012. Efter jämförelse av den bibliografiska informationen i GUP med den i WoS tilldelades sedan varje publikation en av fyra kategorier:

1. Inga fel.
    a. En korrekt affiliering har gjorts till GU i enlighet med punkt 2.
2. Felaktig GU-adress.
    a. Två adressvariationer räknades som korrekta:
        i. University of Gothenburg (Univ Gothenburg).
        ii. Gothenburg University (Gothenburg Univ).
3. Annan organisation.
    a. GU-författare (enligt GUP) affilieras med organisation annan än GU eller Sahlgrenska universitetssjukhuset i WoS, men ej med GU eller Sahlgrenska universitetssjukhuset.
4. Sahlgrenska.
    a. GU-författare (enligt GUP) affilieras med Sahlgrenska universitetssjukhuset, men inte med GU, i WoS.

Med avseende på kategori 2 så får vi hålla i minnet att WoS standardiserar organisationsadresser. Till exempel avbildas *Göteborg University* på *Univ Gothenburg* vilken i nästa steg förkortas till *Univ Gothenburg*. Korrekta GU-adresser omfattar i praktiken således fler än två variationer på källdokumentnivå.

För att kunna undersöka reliabiliteten med avseende på matchning mellan bibliografisk data i GUP och bibliografisk data i WoS granskades varje enskild bibliografisk beskrivning (publikation) i stickprovet manuellt, i första hand med avseende på titel. I de fall titelinformationen ansågs otillräckligt användes också annan bibliografisk information såsom källa, sidnummer och publikationsdatum (nr, volym, publikationsår).

De framräknade relativa frekvenserna användes sedan som punktestimat kring vilka konfidensintervall beräknades. Ett 95-procentigt konfidensintervall användes i samtliga fall. Eftersom en del relativa frekvenser var små användes ”justerat Waldintervall”[1] vid beräkningen av konfidensintervallen.

[1] Agresti A, Coull BA. Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions. The American Statistician. 1998;52(2):119-26.

Resultat och diskussion

Med början i matchningsproblematiken så fann vi inga felaktig matchning mellan GUP och WoS i stickprovet. Givet en konfidensgrad på 95 % får vi ett konfidensintervall $0,00 < \pi < 0,02$. Det vill säga, i 19 av 20 fall finner vi populationsparametern inom detta intervall och kan som mest förvänta oss att två procent av samtliga matchade publikationer är felaktiga. Det finns således inget underlag för att förbättra matchningsmetoden.

Går vi sedan tabell 1 ser vi att 73 % av samtliga publikationer i stickprovet var korrekt affilierade. Med 95 % sannolikhet vågar vi påstå att populationens andel ligger inom intervallet $0,66 < \pi < 0,78$. *Som mest* kan vi förvänta oss att 78 % av populationen är korrekt affilierad och omvänt att 34 % i någon mening är felaktigt affilierad. Detta är förstås ett ganska nedslående resultat, men säkert inte unikt. Om vi omsätter detta till reda tal skulle nära nog 5000 publikationer behöva redigeras med avseende på affilieringen, givet vår populationsdefinition.

Med avseende på kategorin ”Sahlgrenska” så ser vi (tabell 1) att punktestimatet indikerar att hela 18 % av populationen var affilierad med GU (Sahlgrenska akademin) i GUP men inte i WoS där publikationerna i stället var affilierade med Sahlgrenska sjukhuset. Detta procenttal reflekterar förstås storleken på Sahlgrenska akademin i termer av antal publikationer, men otvetydigt kan beräkning av fördelningsindikatorer i olika sammanhang ge substantiella skillnader beroende på hur man betraktar sjukhusaffilieringarna. Med 95 % sannolikhet kan vi som mest förvänta oss att 24 % av populationen tillhör denna kategori och som minst 13 %. Konfidensintervallet är således ganska brett.

I 8,5 % av publikationerna i stickprovet angavs en universitets- eller organisationsadress i källdokumentet annan än GU eller Sahlgrenska universitetssjukhuset (tabell 1). Om denna skattning stämmer med verkligheten så faller 1357 publikationer samt eventuella citeringar till dessa bort vid en extern evaluering av GU. Konfidensintervallets bredd ($0,05 < \pi < 0,13$) ger dock vid hand att denna siffra kan variera avsevärt. En närmare granskning av de publikationer som tillhörde denna kategori gav vid hand att 14 av 17 publikationer troligen var felaktigt affilierade med GU i GUP, medan 3 publikationer var korrekt affilierade med GU i GUP men saknade affiliering till GU i källdokumentet.

Med andra ord kan denna diskrepans också innebära att publikationer och citeringar felaktivt tillskrivs fakulteter under budgetåret.

Slutligen skall vi granska variationer av GU:s universitetsnamn. Vi ser att vi kan förvänta oss som högst 3 % och som minst 0 % av populationen tillhör denna kategori, givet den valda konfidensgraden (tabell 1). Punktestimatet gör gällande att 1,5 % eller 219 publikationer har felstavad GU-adress. Som nämnts under Data och metod, så döljer sig flera namnformer under WoS:s standardiserade adresstermer. Givet den implicita ambitionen att samtliga GU-författare skall använda en och samma namnform, den som anges på GU:s startsida (University of Gothenburg), är andelen felstavade adresser förstås underskattad. Dessutom är det enkelt att finna inkonsistenta avbildningar av namnformer i WoS:

Källdokument		WoS
University of Göteborg	→	Univ Goteborg
University of Göteborg	→	Univ Gothenburg
Göteborg University	→	Gothenburg Univ
Göteborg University	→	Univ Gothenburg

Allt som allt indikerar detta såväl tekniska- som administrativa problem vad gäller universitetsnamn. Uppenbarligen är informationsåtervinning på basis av universitets- och organisationsnamn behäftat med tekniska problem, men den enklaste lösningen torde vara att införa en regel om användningen av standardiserade adressstermer (i.e. University of Gothenburg).

Tabell 1. Punktskattningar med 95-procents konfidensintervall för fyra kategorier.

Kategori	Nedre	Punktskattning	Övre
Inga fel	0,659	0,729	0,782
Sahlgrenska	0,128	0,181	0,240
Annan organisation	0,053	0,085	0,133
Felaktig GU-adress	0,000	0,005	0,031

Sammanfattning

I samband med utvärdering av forskningsprestationer i termer av publicerade vetenskapliga rapporter finns flera tänkbara felkällor. I denna rapport har sådana granskats som främst berör publicering och registrering (indexering) av publikationer. I de fall registreringar i GUP medför felaktiga affillieringar kan detta medföra två oönskade konsekvenser:

1. Uteblivna publikationer och citeringar vid extern utvärdering av GU.
2. Missvisande underlag baserade på publikationer och citeringar vid lokal utvärdering av GU.

Procentsatsen baserad på punktskattningen var oroväckande hög (8,5 %).

Namnvariationer med avseende på GU verkar vid första påseende inte spela någon avgörande roll, eftersom majoriteten fångas upp och unifieras i WoS. Inkonsistens med avseende på denna funktion är dock oroande.

Matchningen av GUP-data med WoS-data verkar fungera tillfredsställande.