Agenzia Nazionale di Valutazione del sistema Universitario e della Ricerca

National Agency for the Evaluation of Universities and Research Institutes

POTENZIALITA' E LIMITI DELLA ANALISI BIBLIOMETRICA NELLE AREE UMANISTICHE E SOCIALI.

VERSO UN PROGRAMMA DI LAVORO

Andrea Bonaccorsi
Consiglio Direttivo ANVUR

7 marzo 2012

0. PREMESSA

La valutazione della ricerca si basa sulla considerazione attenta della eterogeneità delle aree scientifiche in riferimento alle procedure di pubblicazione, alla composizione dei prodotti editoriali, alla lingua usata, ai pattern citazionali. Inoltre i metodi valutativi devono tenere conto degli obiettivi concreti della valutazione e del livello di analisi, come ha correttamente affermato il Gruppo di esperti della Commissione Europea su *Assessment of University Based Research* (AUBR). [1]

Vi è accordo sul fatto che le discipline umanistiche e sociali, con alcune eccezioni, e allo stato attuale, condividano i seguenti elementi differenziali rispetto alle scienze dure [2]:

- minore incidenza delle riviste scientifiche nella produzione complessiva
- peso rilevante attribuito alle monografie
- uso più ampio della lingua nazionale
- frequenza di pubblicazione ridotta.

Tra le conseguenze rilevanti di queste differenze si pone il fatto che le tecniche bibliometriche, basate sulla indicizzazione di riviste scientifiche internazionali, prevalentemente in lingua inglese, e sulla estrazione di indicatori citazionali, sono di difficile applicazione e potenzialmente fuorvianti.
Ciò ha sovente condotto alla errata assunzione che le discipline siano nettamente divise in due parti, una sottoponibile a valutazione bibliometrica, l'altra completamente sottratta a qualunque quantificazione e valutabile solo con *peer review*.
Si tratta di una conclusione affrettata ed errata. In tutte le aree scientifiche una corretta valutazione si basa su di un mix valutativo tra peer review e analisi bibliometrica, il cui peso relativo va stabilito non in astratto e una volta per tutte, ma in riferimento alle concrete opportunità offerte dai database disponibili. [3]
Come ha suggerito Henk Moed:

> A bibliometric approach is a quantitative approach. It attempts to calculate statistics of quantitative aspects derived from scholarly publications. Bibliometric indicators result from the statistical analysis of bibliographic information retrieved from the scholarly literature. This determines both their strength and their limitations. The strength of the bibliometric method is that, once established, it can be applied in a uniform or objective manner, eliminating the influence of subjective or personal factors. On the other hand, being a statistical method, it cannot take into account all particularities or special features of the objects to be assessed. As a consequence, bibliometric data should always be applied in combination with qualitative knowledge about the scholars involved and the subdisciplines in which they are active (Moed, 2002).

Ciò significa che, sebbene si debba riconoscere che i database citazionali a copertura universale esistenti (ISI Web of Science e Scopus) non siano adeguati per una corretta valutazione

---

[1] Si veda European Commission (2010).

[2] Le differenze tra scienze dure e scienze umane e sociali dal punto di vista della ricerca sono state discusse in un classico lavoro di Diana Hicks (2004) e riprese successivamente da Moed (2008). Si vedano anche Viale e Cerroni (2003) e Baccini (2010).

[3] L'importanza di estendere, con gli opportuni adattamenti, alle scienze umane e sociali l'approccio bibliometrico, è sottolineata da numerosi Rapporti di accademie scientifiche, Governi e Agenzie governative. Allo stesso tempo si suggerisce di estendere lo spettro di indicatori per coprire nuove aree di valutazione di impatto. Si vedano Solow et al. (2002) negli Stati Uniti, AHRC (2001) e Hugher, Kitson e Probert (2011) per il Regno Unito, Federal Ministry of Education and Research (2007) per la Germania, e Royal Netherlands Academy of Arts and Sciences (2011) per l'Olanda. La Commissione Europea ha promosso un programma di ricerca per il monitoraggio della ricerca nelle aree umanistiche e sociali (Metris). Per il case study riferito all'Italia vedi Metris (2010). Bonaccorsi e Daraio (2004) e Moed e Daraio (2008) suggeriscono l'utilizzo di indicatori bibliometrici come elementi di modelli non parametrici di analisi di efficienza delle istituzioni universitarie.

bibliometrica delle aree umanistiche e sociali (sempre con alcune eccezioni), [4] devono essere intraprese azioni volte alla creazione di nuovi database in grado di estendere gradualmente la applicabilità delle analisi bibliometriche. [5]

Questa posizione deve essere posta a confronto con il dibattito internazionale e con le possibili obiezioni. Si tratta quindi di non eludere un lavoro teorico sulla valutazione, che con questo contributo si intende iniziare.

Una prima linea di critica riguarda l'*oggetto* della valutazione, o il prodotto editoriale. Si ritiene che nelle scienze umane e sociali il prodotto scientifico principale sia la monografia, e che la valutazione di essa non possa essere svolta altro che attraverso la sua lettura integrale. La monografia è il frutto di molti anni di lavoro, spesso delle ricerche di una vita, e ha uno stile editoriale proprio, irriducibile a quello degli articoli su rivista. Quindi ogni attività bibliometrica sarebbe per principio inutile, perché andrebbe a misurare una parte marginale della produzione umanistica (gli articoli su rivista), lasciando del tutto invariata la produzione monografica.
Si tratta di una critica pertinente e seria. Non vi è dubbio alcuno che nelle scienze umane e sociali la monografia sia centrale. Non è a caso che la metodologia scelta dalla VQR in questi settori (con eccezioni) sia la peer review. Quindi si può affermare con chiarezza che la analisi bibliometrica non potrà sostituire la lettura integrale dei testi.
Tuttavia da ciò non segue che ogni attività bibliometrica sia inutile. In primo luogo le monografie sono, al pari di ogni altro prodotto scientifico, citate, e quindi se disponessimo di un archivio di riviste umanistiche e sociali in lingua italiana sarebbe possibile, in linea di principio, misurare le citazioni che le monografie ricevono a partire da articoli su rivista. In secondo luogo, in un ambiente digitale, anche le monografie possono essere utilizzate come fonti di citazioni ad altre monografie, contribuendo a identificare quelle che maggiormente influenzano la ricerca.
Infine anche i sostenitori più forti del ruolo delle monografie devono convenire sul fatto che non tutte si equivalgono a priori. In altre parole, vi sono monografie che vengono pubblicate solo dopo un lungo percorso selettivo, che ne migliora la qualità, mentre altre vengono pubblicate su collane a pagamento, o su collane che non svolgono alcun filtro.
Una buona monografia presso i migliori editori internazionali è tipicamente preceduta da una lista, spesso impressionante, di ringraziamenti. Si tratta dei colleghi che hanno letto e commentato il manoscritto, obbligando l'autore a modifiche e riscritture. Questo processo è richiesto e incoraggiato dagli editori, perché aumenta notevolmente la qualità dei libri. Non sempre accade così nel nostro paese, in cui una quota rilevante di monografie accademiche è letta, prima della pubblicazione, solo da chi l'ha scritta, e (nei casi migliori) dal maestro dell'autore. Fermo restando il principio generale che le monografie vanno sempre lette per intero, è quindi utile aumentare la informazione a priori circa le procedure editoriali. Quindi dal ruolo centrale delle monografie non segue affatto l'inutilità della bibliometria.

[4] I limiti dei database citazionali esistenti nel caso delle scienze umane e sociali sono ben noti e discussi da una ampia letteratura: Tarantino (2005), Archambault et al. (2006), Nederhof (2006), Hellqvist (2010), Piazzini (2010).

[5] Molti autori sottolineano che la prassi citazionale nelle scienze umane e sociali differisce profondamente da quella in uso nelle scienze dure: Chubin (1980) e McRoberts e McRoberts (1989) offrono una panoramica ampia, anche se datata, dei problemi derivanti dagli indicatori citazionali, mentre Hurt (1987), So (1998), Amin e Babe (2000), Burnhill e Tubby-Hille (2003) e Huang e Lin (2008; 2010) illustrano le specificità delle scienze umane e sociali in riferimento all'uso delle citazioni. Gilbert (1977) ha introdotto la nozione di citazione come persuasione. Moed (2005) resta il classico riferimento generale sulla analisi citazionale (si veda anche Moed, 2000).

Una seconda linea di critica attacca l'importanza delle *citazioni* come indicatore di impatto. Mentre alcune critiche sono generali e sono state largamente affrontate nella letteratura bibliometrica, [6] nelle scienze umane e sociali viene talora avanzata una critica più radicale. L'idea è che la citazione non descriva affatto l'impatto di un lavoro, quanto piuttosto sia indice di conformismo e di adattamento.
Una formulazione particolarmente incisiva di questa posizione è stata offerta da Michel Wieviorka, Presidente della International Sociological Association nel periodo 2006-2010:

> Can research in our disciplines be the object of normative evaluation in this way? Is the good researcher therefore the one who is most frequently quoted or the good journal the one which the evaluators consider the most professional? Can the impact of research be assessed in this way? (…) Sociologists know perfectly well that the most highly rated journals in their discipline tend to be boring and predictable – they tend to embody a prerequisite to earn recognition by one's peers and obtain a post or a promotion (Wieviorka, 2011, 308).

Si tratta di una tesi formulata in un modo che non consente di contro argomentare. Infatti se questa posizione fosse criticata da un autore molto citato, a questi verrebbe obiettato che, in quanto noioso e prevedibile, non ha argomenti accettabili. Un autore poco citato, al contrario, avrebbe tutto l'interesse a non criticare questa posizione. In conclusione, si tratta di un argomento che si auto-conferma.
Al di là del paradosso, occorre insistere sull'importanza delle citazioni come unità elementare della valutazione, almeno su grandi aggregati, per intervalli di tempo adeguati e con normalizzazioni appropriate. In linea di principio, infatti, qualunque prodotto scientifico è fatto per essere utilizzato da altri. Come scrive Moed:

> in a bibliometric approach, it is assumed that important contributions to scholarly progress are sooner or later communicated in scholarly publications (Moed, 2002, 15).

Si può quindi discutere su cosa voglia dire in concreto "sooner or later", cioè su quale finestra temporale vogliamo dare alle scienze umane e sociali. È noto infatti che in questi settori le opere vengono citate più tardi e vengono citate per molti anni, generando finestre citazionali strutturalmente diverse da quelle delle scienze dure. Si possono inoltre introdurre normalizzazioni che tengano conto della estrema specializzazione di molte aree umanistiche e sociali, nonché della possibilità di clusterizzazione delle citazioni per scuole di pensiero. Tutto bene.
Ma occorre insistere che la conoscenza prodotta dalla ricerca scientifica, in qualunque ambito, è *costitutivamente* soggetta a validazione intersoggettiva, quale si esprime *ex ante* nelle procedure selettive per le riviste e nelle regole editoriali per le monografie, *ex post* nel riconoscimento tributato ai colleghi attraverso le citazioni. Altre forme di conoscenza che non usano questi strumenti sono del tutto legittime e sono essenziali alle società democratiche, ma non hanno i caratteri di validità che vengono riservati alla conoscenza scientifica. La comunicazione scientifica, a differenza di quella saggistica, opinionistica, politica o culturale in senso lato, è soggetta a regole di selezione. È generalmente ammesso che nelle scienze umane e sociali, a causa del peso delle monografie e delle riviste in lingua nazionale senza refe raggio, una quota più elevata di prodotti non subiscono di fatto alcuna severa selezione *ex ante* e si sottraggono quindi all'onere di ottenere il

[6] Ricordiamo tra queste:
- la possibilità che la citazione sia negativa, ovvero contenga la confutazione di altri lavori
- il ruolo delle autocitazioni
- le differenze nelle citazioni tra articoli di ricerca e articoli di review
- la possibilità di manipolazione, soprattutto in aree di nicchia, derivanti da *cliques* di autori che si citano reciprocamente
- le restrizioni presenti nei database circa le fonti da cui provengono le citazioni
- le diversità nelle prassi citazionali tra diverse discipline scientifiche.

consenso preventivo dei referee. Pretendere che siano anche sottratti ad un riconoscimento *ex post* attraverso le citazioni è probabilmente un po' troppo.

Una terza obiezione riguarda l'effetto potenzialmente distorsivo della analisi bibliometrica per la produzione scientifica che assume caratteri non ortodossi, critici, deliberatamente irregolari e minoritari, oppure che propone prospettive talmente innovative da poter essere riconosciute (e citate) solo tardivamente. Nelle scienze umane e sociali questo pericolo è particolarmente forte in quanto vige un pluralismo paradigmatico che non può essere considerato un fenomeno transitorio, come di una fase immatura da superare verso un ideale di scienza normale, ma presenta caratteri costitutivi. E poiché la produzione scientifica non è fatta solo di idee, ma anche di gruppi di ricerca, riviste, editori, finanziamenti, vi può essere il rischio di marginalizzare di fatto posizioni minoritarie, che invece sono essenziali alla creatività scientifica.
Questa obiezione è interessante e deve essere tenuta in grande considerazione. Ad esempio l'adozione di indicatori normalizzati per discipline potrebbe penalizzare la ricerca multidisciplinare. Da un altro punto di vista, la pressione per obiettivi quantitativi potrebbe spingere a pubblicare molti lavori di facile accettazione, invece che dedicarsi a lavori più creativi. [7] Sotto un altro punto di vista, gli indicatori citazionali potrebbero scoraggiare nella creazione di nuove riviste. [8]
Rispetto a questa obiezione, tuttavia, si deve ricordare che per tutti i ricercatori, anche per chi sostiene posizioni di minoranza, il riconoscimento da parte dei pari è un obiettivo ambito. Nessun ricercatore serio si rassegna al fatto che le proprie idee non siano accolte dal resto della comunità e i suoi lavori non siano citati. Gli indicatori citazionali agiscono come stimolo per affermare le proprie idee. E la storia della scienza mostra in modo chiaro che le idee buone sono (quasi) sempre riconosciute, anche se con ritardo. Abolire il ruolo del riconoscimento formulato attraverso le citazioni non farebbe, in ultima istanza, il bene della scienza.
D'altra parte, la scelta tra pubblicare lavori più facili e più citabili o avventurarsi verso scoperte più rischiose può essere considerata una delle scelte più caratteristiche dello stile scientifico dei ricercatori. Come hanno mostrato con un elegante modello Dalle e Carayol (2004), chi pubblica in aree già presidiate viene citato di più, ma condivide i riconoscimenti con molti altri, mentre chi rischia su terreni nuovi viene inizialmente citato di meno, ma se compie scoperte importanti ottiene un immediato riconoscimento.
Quindi questa obiezione deve essere considerata molto seriamente quando si valutano singoli ricercatori, in particolare i più giovani. Ma per grandi aggregati e su tempi più lunghi, si deve riconoscere che anche le posizioni minoritarie, se argomentate con rigore, tendono a ricevere prima o poi i riconoscimenti dovuti. Si tratta quindi di una obiezione non decisiva.

Allo stato della discussione, quindi, non sembra accettabile l'idea che la bibliometria sia dannosa o inutile per le scienze umane e sociali. Si possono compiere importanti passi in avanti, nel rispetto delle differenze epistemologiche, comunicative e sociologiche delle diverse comunità scientifiche.

---

[7] La possibilità che il ranking delle riviste con indicatori bibliometrici possa ridurre l'incentivo dei ricercatori a svolgere ricerca interdisciplinare è discussa da Rafols et al. (2011), mentre Rodriguez-Navarro (2009) discute effetti distorsivi a favore della scienza normale.
Valdecasas, Castroviejo e Marcus (2000) hanno proposto in modo convincente che l'esclusivo uso delle citazioni può danneggiare le intraprese scientifiche di lungo termine, come la ricerca tassonomica per la biodiversità: "Basic taxonomic work is not highly cited, except in 'hot' taxa like the genus Homo. The number of authors citing a paper during the short period of time (ten years) that the SCI uses for its statistics is relatively low. But taxonomy papers continue to be referred to and cited for more than a century after their publication. Almost every good taxonomic paper becomes a classic in the literature"
[8] L'argomento è stato proposto da Lamp (2009), il quale peraltro sostiene che l'ingresso di nuove riviste è reso oggi meno oneroso dalle tecnologie digitali.

Sulla base delle esperienze internazionali, l'ANVUR ritiene che si debbano porre in essere azioni nelle seguenti direzioni:

1. supporto alla candidatura di un consistente gruppo di riviste in lingua italiana, che soddisfano i requisiti editoriali accolti in sede internazionale, per l'istruttoria ai fini della indicizzazione in sede ISI e Scopus
2. rating delle riviste non indicizzate in lingua italiana, con procedure metodologicamente valide e comparabili con analoghe esperienze internazionali
3. pubblicazione di informazioni validate sulle procedure editoriali e di selezione dei manoscritti da parte di editori nazionali
4. creazione di un archivio di metadati e di referenze tratte da monografie in lingua italiana e di riviste italiane disponibili in formato digitale
5. ricerca e sperimentazione di indicatori non citazionali.

Queste azioni hanno l'obiettivo di rendere permanente, oltre i limiti di tempo e normativi della VQR, un sistema di valutazione applicabile alle scienze umane e sociali. [9]
Su ciascuno di questi temi l'ANVUR **ha già attivato dei gruppi di lavoro interni**, i cui risultati verranno messi a disposizione delle comunità scientifiche, dei ricercatori, del mondo dell'editoria e delle biblioteche, entro pochi mesi, al fine di avviare un percorso comune.

Il presente documento illustra le caratteristiche delle linee di azione, le premesse metodologiche, i risultati attesi. [10]

[9] Nel caso italiano, il tema dell'utilizzo di indicatori bibliometrici per le scienze umane e sociali è stato al centro di interventi recenti del CNR (2009) e del CUN (2009).
Per una valutazione dell'impatto della VTR su alcune aree delle scienze sociali, in particolare la sociologia e le scienze politiche, vedi Bartolini (2007), Diani (2008), Chiesi (2008). Marcuzzo e Zacchia (2007) hanno esplorato il potenziale di database bibliografici diversi da ISI e Scopus per l'economia.
Per una valutazione complessiva, vedi Franceschet e Costantini (2009), Reale (2010), Biolcati-Rinaldi (2010), Aru et al. (2010), Costantini e Franceschet (2011).

1. Supporto alla indicizzazione di riviste in lingua italiana

In questa linea di attività l'ANVUR intende **sostenere la candidatura di un consistente gruppo di riviste italiane alla indicizzazione presso ISI e Scopus**.
Appare ragionevole che lo sforzo maggiore sia indirizzato alle riviste di fascia A, definite secondo il rating proposto dai GEV della VQR. Non è esclusa la estensione alle riviste di fascia B, con opportune verifiche.

Preliminare a tale attività è un programma basato su due assi:

- ricognizione delle condizioni formali di accesso ai database
- analisi sistematica della rispondenza ai requisiti da parte delle riviste italiane di fascia A.

Un passo preliminare è la raccolta, anche attraverso interviste dirette e/o audizioni, dell'esperienza delle riviste italiane di area umanistica e sociale che già hanno superato le soglie di accesso in ISI e/o Scopus.
Attraverso un questionario semi-strutturato verranno raccolte indicazioni circa:

- attività necessarie per l'istruttoria
- principali ostacoli alla ammissione
- durata dei tempi di istruttoria
- costi complessivi
- collocazione iniziale in classi di Impact Factor (ISI)
- benefici riscontrati.

Una volta svolta questa istruttoria, ed esaminati i dati della sottoposizione dei prodotti della VQR, sarebbe possibile compilare una lista di possibili candidati.
I direttori delle riviste dovrebbero essere coinvolti in una iniziativa organizzata, dichiarare la propria adesione, concordare modi e tempi di apertura delle istruttorie.
ANVUR offrirà tutto il supporto logistico ed organizzativo, anche rispetto alla trattativa con Thomson Reuters ed Elsevier.

2. Rating delle riviste non indicizzate

Premessa

Atteso che nelle aree umanistiche e sociali si fa largo uso di riviste scientifiche in lingua nazionale, e che esse sono indicizzate in misura marginale nei database internazionali, si pone il problema della possibilità di valutare le riviste stesse.
Mentre per le riviste indicizzate la valutazione della rivista come tale viene effettuata attraverso misure normalizzate delle citazioni medie ricevute dagli articoli pubblicati su di essa (Impact Factor, SJR o altre misure), per le riviste non indicizzate si è provveduto in molti paesi ad un esercizio di rating.
Il rating consiste nella assegnazione di ogni rivista scientifica ad una classe di merito, normalmente con un ordine gerarchico, a seguito di una valutazione esperta.
La premessa metodologica di questo esercizio è che **la qualità della rivista non può essere traslata automaticamente sulla qualità dell'articolo in essa contenuta**. La ragione è che esiste una ineliminabile variabilità della qualità dei singoli articoli all'interno della stessa rivista. Tuttavia la assegnazione delle riviste a classi di qualità **fornisce una informazione a priori circa il valore atteso della qualità dei singoli articoli**. Infatti le riviste di maggiore reputazione tendono ad avere comitati editoriali più prestigiosi, politiche di accettazione più severe, tassi di rigetto più elevati, ricevono una quantità maggiore di sottomissioni, e quindi sottopongono i lavori ad una selezione più spinta, che ne innalza **in media** la qualità. Tale informazione a priori potrebbe naturalmente essere smentita dalla analisi dei singoli lavori, ma la probabilità che ciò avvenga non è uniforme.
Detto in altri termini: mentre la probabilità di trovare un lavoro scadente in una eccellente rivista non è mai pari a zero, tuttavia è minore della probabilità di trovarlo su una rivista di qualità inferiore.
Questo documento esamina le premesse metodologiche di questo esercizio. Più in particolare, il presente documento discute le modalità con cui il rating delle riviste delle aree umanistiche e sociali è stato effettuato al'interno della VQR, **illustrando una metodologia con la quale sarà possibile aggiornare e rendere permanente la valutazione**, rendendola uno strumento valido e accettato anche oltre i limiti temporali e istituzionali della VQR stessa.

2.1 L'esperienza internazionale

La classificazione delle riviste in classi di merito è stata svolta in diversi paesi e discipline.
In alcuni casi si è trattato di un esercizio originato all'interno di programmi di valutazione, in altri invece di una iniziativa volontaria. Le premesse di metodo sono riconducibili a due problemi:

a) non sono disponibili indicatori citazionali in quanto le riviste sono in lingua nazionale e scarsamente presenti nei database ISI e Scopus
b) gli indicatori citazionali correnti (in particolare IF e SJR) sono ritenuti inaffidabili o almeno inaccurati.

Il problema (a) è centrale nelle scienze umane e sociali, mentre il problema (b) si è posto soprattutto in matematica, dove l'Unione Matematica Mondiale, aderendo alle critiche contro l'IF, ha promosso una iniziativa di classificazione delle riviste basata su gruppi di esperti. [11]

[11] Il gruppo di lavoro dell'IMU (International Mathematical Union) e dell'ICIAM (International Council of Industrial and Applied Matheamatics) ha proposto una classificazione delle riviste in quattro classi, come segue:

Per quanto riguarda le scienze umane e sociali è largamente accettata la nozione che le riviste indicizzate costituiscono una piccola frazione del totale e gli indicatori citazionali disponibili non sono rappresentativi della qualità della ricerca, con alcune importanti eccezioni (in particolare, psicologia, economia e statistica). A partire da questa consapevolezza diversi governi e istituzioni internazionali hanno iniziato esercizi di rating.

In Spagna la classificazione delle riviste è svolta da molti anni, sia con singole iniziative settoriali che con la costituzione di un archivio unificato nazionale. Gimenez-Toledo et al. (2007) discutono estesamente l'esperienza spagnola, con ampia bibliografia. La classificazione delle riviste in lingua spagnola è pubblicata regolarmente[12], si basa su 4 classi (A, B, C, D) più una classe di eccellenza, e viene aggiornata periodicamente, attraverso la consultazione di molte centinaia di esperti.

In Francia una importante iniziativa di classificazione è stata svolta dalla AERES (Agence d'Evaluation de la Recherche et de l'Enseignement Superieur). Nel 2008 l'Agenzia ha pubblicato una ampia lista di riviste di scienze umane e sociali, classificate in tre classi (A, B, C). [13] La pubblicazione ha naturalmente prodotto forti reazioni, anche attraverso la pubblicazione di lettere aperte alla stampa da parte delle riviste classificate nella fascia inferiore. Nel 2011 il rating è stato ritirato e si è proceduto alla sola pubblicazione di una lista di riviste. Tuttavia il documento del 2008, facilmente reperibile in rete, costituisce ancora un punto di riferimento per le università.

L'esperienza dell'Australia è citata spesso come un esempio del fatto che la classificazione delle riviste è dannosa. Ricapitoliamo i fatti. Nel 2008 il Governo ha lanciato l'ERA (Excellence in Research for Australia), programma che includeva un esercizio di classificazione delle riviste scientifiche, basato su consultazione di esperti, che ha superato le 20.000 unità ed è stato pubblicato nel 2010. Nel maggio 2011 il rating delle riviste è stato ritirato dal nuovo governo.

- Tier 1: A top journal in mathematics or a major subfield of it. Almost all papers published are of very high quality, and it regularly publishes papers that are of great significance. Peer-review is applied consistently and rigorously, and editorial work is carried out by leading mathematicians.
- Tier 2: Very strong journal with a carefully run and reliable peer-review process. Papers are generally of high quality, and regularly papers are published which are of significant importance in at least a subfield of mathematics.
- Tier 3: Solid journal that generally publishes reputable work and follows accepted practices of peer review, but are generally less selective than journals of Tier 2, and paper quality is more variable. Such journals may play an important role in speci_c communities, but are usually not considered highly important to mathematics or a sub_eld globally.
- Tier 4: Journals not found to meet the standards of the other tiers.

Reperibile presso http://www.mathunion.org/fileadmin/IMU/Report/WG_JRP_Report_01.pdf. 30 giugno 2011.

La proposta segue la pubblicazione di *Citation Statistics*. A report from International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS), Giugno 2008. Vedi http://www.mathunion.org/fileadmin/IMU/Report/CitationStatistics.pdf.

[12] L'archivio delle riviste è denominato CIRC (Clasificación Integrada de Revistas Cientficas). Vedi http://epuc.cchs.csic.es/circ/categorias.html

[13] http://www.aeres-evaluation.fr/Publications/Methodologie-de-l-evaluation/Listes-de-revues-SHS-sciences-humaines-et-sociales

Alcuni autori [14] avevano sollevato il problema della disparità di trattamento nella assegnazione dei punteggi più alti (classi A e A*), sostenendo che alcune discipline erano state sacrificate. Altri editor di riviste finite in classi più basse avevano protestato vibratamente. Esaminando la letteratura, tuttavia, non si sfugge alla convinzione che i problemi sollevati fossero di scala nettamente inferiore rispetto alla dimensione dell'esercizio e al suo rigore metodologico. Il ritiro dei rating non pare giustificato sulla base delle critiche pubblicate.
Per capire meglio cosa è accaduto, ho interpellato un esperto australiano che ha avuto ruoli importanti nell'Australian Research Council (ARC), che conosce dal di dentro la vicenda e che, per comprensibili ragioni, mi ha chiesto l'anonimato. Ecco la sua ricostruzione dei fatti:

- la motivazione principale del ritiro è che alcuni opinion maker ("politically persuasive individuals") hanno persuaso il Presidente dell'Australian Research Council che i ranking delle riviste venivano utilizzati in modo inappropriato dalle università per indirizzare le attività e le pubblicazioni;
- i ranking delle riviste avevano elevata correlazione con indicatori citazionali normalizzati per disciplina; data l'elevata correlazione alcuni hanno sostenuto che la classificazione delle riviste fosse ridondante;
- alcuni influenti ricercatori (inclusi alcuni editor di riviste scientifiche australiane) hanno protestato affermando di essere stati classificati in modo inappropriato.

A giudizio del mio interlocutore, queste proteste hanno riguardato una piccolissima frazione ("a tiny number") delle oltre 20,000 riviste censite. Si è trattato quindi, a suo giudizio, non di debolezze metodologiche ma di pressioni politiche. È interessante notare che nel caso dell'Australia, essendo le riviste scritte in inglese, sono maggiormente disponibili indicatori citazionali. Nel caso dell'Italia, al contrario, tali indicatori sono sostanzialmente assenti e quindi l'argomento della ridondanza della classificazione delle riviste non si applicherebbe. In definitiva portare il caso dell'Australia come emblematico della impossibilità di effettuare classificazioni di riviste è del tutto inappropriato.

2.2 Qualità e quantità

La possibilità di assoggettare le riviste scientifiche ad una classificazione gerarchica non è da tutti accettata. Nel discutere una esperienza svolta nel settore giuridico per le università del Belgio fiammingo, Moed (2002) riporta che il Comitato Interuniversitario delle Facoltà di Legge si era rifiutato di assegnare una classificazione alle riviste giuridiche, adducendo come motivazione il fatto che **le riviste giuridiche mostravano una eccessiva variabilità nella qualità degli articoli pubblicati.**
Si tratta di una posizione comprensibile sul piano della difesa di pratiche consolidate, ma infondata sul piano scientifico. Infatti se si ritiene che la variabilità della qualità interna alle riviste sia elevata, non resta che confrontare la variabilità *tra* le riviste con quella *all'interno* delle riviste, per trarre una conclusione. In effetti, Moed (2002) ha mostrato che, somministrando due questionari ad un ampio numero di studiosi, sia belgi che esteri, era del tutto possibile, con opportuni metodi, estrarre i giudizi esperti e trasformarli in classificazioni di merito. Alla fine dell'esercizio, lo stesso Comitato ha dovuto riconoscere che la classificazione era giustificata e ponderata.

[14] I problemi incontrati nella classificazione delle riviste in Australia è descritto dettagliatamente in Genoni e Haddow (2009) e Haddow e Genoni (2010). Sulle critiche sollevate intorno alla equità tra discipline si veda il botta e risposta tra Vanclay (2011; 2012) e Butler (2011).
Butler (2003a; 2003b) discute l'impatto della valutazione della ricerca in Australia ed in particolare gli effetti sui comportamenti di pubblicazione dei ricercatori

Questo caso mostra un problema più generale, che è utile affrontare subito in modo approfondito. Circola spesso l'idea che nelle aree umanistiche e sociali **la qualità della ricerca sia un elemento incommensurabile, sottratto in linea di principio ad ogni quantificazione e suscettibile solo di giudizi intuitivi e sintetici, non articolabili e replicabili.** Questa idea si fonda sull'assunzione implicita che non sia possibile un consenso intersoggettivo sulle dimensioni della qualità. Questa assunzione è del tutto non dimostrata. Si può al contrario mostrare che gli individui posseggono intuitivamente una nozione multidimensionale di qualità che, se opportunamente indirizzati, possono esprimere nel linguaggio naturale e in forma gerarchica. In altre parole, si dà una validazione intersoggettiva della qualità, anche di oggetti complessi e caratterizzati da una ampia varietà di dimensioni. Il problema centrale è la formulazione di un linguaggio adeguato a catturare le dimensioni, spesso implicite, della qualità. In questo modo gli individui non sono obbligati a "quantificare" il proprio giudizio, ma solo a esprimere qualitativamente un gradimento maggiore o minore. Se tale formulazione viene raggiunta, si può dimostrare formalmente che esistono regole di aggregazione dei giudizi individuali. In un interessante lavoro, due matematici francesi mostrano come sia possibile trasformare giudizi esperti anche estremamente complessi come quelli dei giudici del vino, o dei tuffi olimpionici, o in misura diversa, dei candidati alle presidenziali francesi, in misure quantitative (Balinski e Laraki, 2011; si veda la recensione in Bonaccorsi, 2012). Questa prospettiva sposta il baricentro della discussione, dalla improponibile contrapposizione qualitativo-quantitativo, alla costruzione di linguaggi sufficientemente ricchi per catturare il giudizio, implicito ma molto preciso, che gli individui hanno del mondo.

2.3 Natura della assegnazione di un rating

La assegnazione di riviste scientifiche a categorie di qualità è una procedura che rientra nella più generale classe di procedure volte alla *assegnazione di un oggetto qualsiasi ad una categoria*. Preliminare a tale procedura è la definizione di quale tipo di categoria si tratti.
Per definire una categoria occorre prima di tutto verificare la natura dei dati sulla base dei quali può essere costruita. Esistono quattro tipi di dati, o variabili, a cui corrispondono diverse possibilità (Box 1).[15] **La procedura di rating corrisponde alla creazione di variabili ordinali, ovvero variabili che sono in grado di creare categorie ordinate gerarchicamente**.
È importante rendere chiaro il fondamento concettuale sul quale si basa la definizione di rating. Occorre in altre parole fornire delle definizioni della riviste in ciascuna categoria dalle quali risulti in modo inequivocabile la ragione per cui le categorie possono essere ordinate gerarchicamente (vedi oltre). Si noti che in uno degli esperimenti di classificazione delle riviste umanistiche adottato su scala europea (ERIH), inizialmente fu dichiarato che le categorie A, B e C non andavano intese come categorie di qualità decrescente, ma solo come indicatori di diversità. Ciò allo scopo di prevenire possibili obiezioni. In realtà ben presto tutti compresero che le riviste in classe A (riviste ad ampia diffusione internazionale) erano considerate più importanti delle riviste in classe B (riviste ad ampia diffusione nazionale), e queste delle riviste in classe C (riviste locali). Le obiezioni non furono evitate. Quando un ordinamento tra classi è implicito nella definizione delle categorie, è meglio renderlo esplicito.

Occorre anche ricordare che **con il rating non è possibile giungere ad un ordinamento gerarchico delle singole riviste (ranking)**, cosa che è invece possibile fare disponendo di indicatori citazionali, ad esempio con l'Impact Factor. [16]

[15] Si tratta di una distinzione ormai classica. Si vedano Bryman (2008), Lewis-Beck, Bryman e Liao (2004), Hardy and Bryman (2004).
[16] Le riviste scientifiche sono valutate, dal punto di vista bibliometrico, con una varietà di indicatori, il più importante e consolidato dei quali è l'Impact Factor (IF).

Quindi, nonostante che nel linguaggio comune si parli sovente di "ranking delle riviste", tale denominazione non è corretta, ed è preferibile usare l'espressione "rating". [17]

Box 1
Classificazione della natura delle variabili

Variabili nominali
Si definiscono nominali (*nominal variable*) o categoriche (*categorical variable*) le variabili che identificano categorie tra loro eterogenee in senso qualitativo, per le quali non è possibile un ordinamento di alcun tipo. Ad esempio la colorazione di oggetti (se non viene esaminata sotto il profilo fisico della lunghezza d'onda), oppure alcune patologie mediche per le quali si debba ricorrere ad un giudizio clinico o sintomatico complesso, oppure ancora la classificazione bibliotecaria.
In tutti questi casi si assegnano oggetti a singole categorie, sovente seguendo regole formalizzate, senza che esse assumano un significato gerarchico.

Variabili ordinali
Si definiscono ordinali le variabili che identificano categorie ordinate gerarchicamente, ma per le quali le differenze tra categorie contigue non sono uguali. Con variabili ordinali è possibile costruire scale, ovvero ordinamenti gerarchici.
Ad esempio le stelle assegnate ad un ristorante, o le risposte ad un questionario del tipo "ogni giorno", "2 o 3 giorni alla settimana" e "da 4 a 6 giorni alla settimana".

Variabili di intervalli
Le variabili di intervalli (*interval variable*) sono variabili per le quali le distanze tra le categorie sono identiche lungo tutto il range dei valori ammissibili.

Sui limiti dell'IF si è sviluppata una imponente letteratura, centrata sulla fissità delle finestra temporale di citazione, sulla assenza di citazioni da riviste non-ISI, sulla asimmetria della distribuzione delle citazioni. Si veda Archambault e Larivière (2009) sulla storia dell'IF. Una delle critiche più penetranti, introdotta da Pinski e Narin (1977) e radicalizzata da Bollen, Rodriguez e van de Soberl (2006) è che le citazioni non sono "pesate", nel senso che hanno lo stesso valore anche se provengono da riviste di prestigio molto diverso. Abramo et al. (2010) mostrano con riferimento all'Italia che l'IF e le citazioni sono fortemente correlate in aggregato per intervalli di tempo estesi, ma poco correlati e in grado di determinare forti instabilità nei ranking per intervalli inferiori.
È unanimemente accettato che la valutazione della rivista non è utilizzabile per la valutazione dei prodotti che in essa vengono pubblicati (Nederhof e Zwaan, 1991; Jarwal, Brion e King, 2009). Nel caso dell'IF Seglen (1997) mostra che la distribuzione delle citazioni agli articoli è estremamente asimmetrica, per cui attribuire ai singoli articoli il valore medio delle citazioni della rivista produce gravi distorsioni. Figà Talamanca (2000) è una formulazione particolarmente incisiva dei limiti degli indicatori riferiti alle riviste. Starbuck (2005) mostra che anche in riviste ad alto IF esistono numerosi articoli non citati.
Più recentemente sono stati proposti la estensione dell'h-index alle riviste (Braun, Glanzel and Schubert, 2006; Norris e Oppenheim, 2010) e l'indicatore SJR (Gonzalez-Pereira, Guerrero-Bote e Moya-Anegon, 2010). Per una visione teorica del tema della valutazione bibliometrica delle riviste si veda Boyossou e Marchant (2011).

[17] Ciò non toglie che anche nelle scienze sociali sia possibile utilizzare ranking citazionali, per le riviste indicizzate. Per il caso dell'accounting, si veda Coyne et al. (2010); per l'economia politica Kalaitzidakis, Mamuneas e Stengos (2003), Ritzberger (2008) e Hofmeister (2011). Altre referenze utili sono Nederhof e Zwaan (1991), Nederhof e Noyons (1992) e Nederhof (2006).

Variabili di rapporti
Le scale a rapporti consentono non solo di ordinare gerarchicamente degli oggetti, ma anche di assegnare un significato al loro rapporto. Si tratta di scale di intervalli con un punto zero fisso.

Questa distinzione è della massima importanza, perché la misurazione del consenso tra valutatori assume forme diverse se il giudizio viene effettuato tra categorie nominali, tra le quali non è possibile istituire rapporti di similarità o vicinanza, oppure tra categorie ordinali o addirittura tra categorie definite da variabili continue. Intuitivamente, se le categorie possono essere ordinate in modo gerarchico (es. Limitato, Accettabile, Buono, Molto buono, Eccellente) **la differenza tra categorie contigue deve essere trattata diversamente rispetto alla differenza tra categorie lontane**. Se due valutatori giudicano lo stesso oggetto Limitato ed Eccellente, rispettivamente, il loro disaccordo è molto più profondo di quello che vi sarebbe se il loro giudizio fosse Limitato e Accettabile.
Come vedremo, questa distinzione si riflette nella costruzione di indici di affidabilità.

2.4 Numero e definizione delle categorie

2.4.1 Numero delle categorie

Un problema metodologico importante è la **scelta del numero delle categorie.**
Essa obbedisce ad un compromesso tra esigenze opposte. Da un lato, più alto è il numero delle categorie, più fine è la classificazione, e minore è la probabilità di giudizi coincidenti dovuti al puro caso (intuitivamente, più numerose sono le celle della matrice, minore è la probabilità che un oggetto cada in ciascuna di esse). Allo stesso tempo, tuttavia, una elevata numerosità delle categorie può essere in conflitto con l'esigenza di contenere i costi e con lo stato delle conoscenze dei valutatori.
Nelle esperienze internazionali sono spesso usate 3 categorie (A, B e C), con una eventuale quarta categoria residuale, oppure con una quarta sotto-categoria definita all'interno della classe più alta (A*).. La scelta è del tutto ragionevole.
I GEV della VQR hanno adottato una classificazione su tre livelli, dove le categorie A e B sono listate nominativamente, mentre la categoria C è intesa come residuale.

2.4.2 Definizione delle categorie

La assegnazione delle riviste alle categorie riflette il giudizio esperto di una serie di valutatori ("rater") i quali utilizzano dei criteri di classificazione.
Nell'esercizio dei GEV non è stata adottata una definizione formale delle categorie, ma si è fatto riferimento ad una definizione sintetica di "reputazione". Essa include una molteplicità di dimensioni.
In sede di aggiornamento dei rating occorrerà standardizzare la definizione delle categorie, in modo da fornire ad ogni valutatore un insieme di informazioni omogeneo e ben chiaro dal punto di vista semantico.
Allo scopo di suggerire alcuni elementi, la tabella 1 riassume le definizioni adottate in alcune esperienze internazionali.

Tabella 1

Definizioni delle categorie di rating delle riviste in alcuni esercizi internazionali

ERIH *	ERA **	**CIRC ****
European Science Foundation	Excellence of Research in Australia	**Clasificacion Integrada de Revistas Cientìficas**
NATional (NAT): European publications with a recognised scholarly significance among researchers in the respective research domains in a particular (mostly linguistically circumscribed) readership group in Europe; occasionally cited outside the publishing country, though their main target group is the domestic academic community. **INTernational (INT):** both European and non-European publications with an internationally recognised scholarly significance among researchers in the respective research domains, and which are regularly cited worldwide. International journals are themselves classified into two sub-categories based on a combination of two criteria: influence and scope: **INT1 Sub-Category:** international publications with high visibility and influence among researchers in the various research domains in different countries, regularly cited	**Overall criterion: Quality of the papers** **A*** Typically an A* journal would be one of the best in its field or subfield in which to publish and would typically cover the entire field/subfield. Virtually all papers they publish will be of a very high quality. These are journals where most of the work is important (it will really shape the field) and where researchers boast about getting accepted. Acceptance rates would typically be low and the editorial board would be dominated by field leaders, including many from top institutions. **A** The majority of papers in a Tier A journal will be of very high quality. Publishing in an A journal would enhance the author's standing, showing they have real engagement with the global research community and that they have something to say about problems of some significance. Typical signs of an A journal are lowish acceptance rates and an editorial board which includes a reasonable fraction of well known	**Grupo A (gA)**: integrado por las revistas científicas de mayor nivel. Pertenecerían al mismo las revistas internacionales de mayor prestigio que han superado procesos de evaluación muy exigentes para el ingreso en diferentes bases de datos. - Indexadas en Science citation index, Social sciences citation index o Arts & humanities citation index según los master lists de 2011. - Indexadas en las listas European reference index for the humanities (European Science Foundation) con una calificación de INT. **Grupo B (gB)**: compuesto por revistas científicas españolas de calidad pero que no alcanzan un alto nivel de internacionalización aunque son revistas que reciben cierto grado de citación y que respetan los estándares de publicación. Asimismo forman parte de este grupo aquellas revistas científicas internacionales con un menor pero aceptable grado de prestigio y difusión. - Indexadas en el primer cuartil según promedio de citas de cualquiera de las categorías del Índice de impacto de las revistas españolas de ciencias sociales o del Índice de impacto de las revistas españolas de ciencias jurídicas (Grupo EC3). Se toma como referencia los impactos acumulativos de los años 2005-2009. - Indexadas en DICE (Difusión de las revistas españolas de ciencias sociales y humanas) (Iedcyt) y que cumplen con el requisito de contar con evaluación por expertos y además estar presentes en el

all over the world. **INT2 Sub-Category:** international publications with significant visibility and influence in the various research domains in different countries. **W Category Journals:** journals which published their first issue three years or less before the closing date for feedbacks for a given panel". Closing dates list is available here.	researchers from top institutions. **B** Tier B covers journals with a solid, though not outstanding, reputation. Generally, in a Tier B journal, one would expect only a few papers of very high quality. They are often important outlets for the work of PhD students and early career researchers. Typical examples would be regional journals with high acceptance rates, and editorial boards that have few leading researchers from top international institutions. **C** Tier C includes quality, peer reviewed, journals that do not meet the criteria of the higher tiers.	Catálogo Latindex. - Indexadas en la base de datos Scopus según su List of titles de abril de 2011 y catalogadas en las categorías Social sciences (code 3300) y Arts and humanities (code 1200). - Indexadas en las listas European reference index for the humanities (ESF) con un calificación de NAC1 o NAC2 **Grupo C (gC)**: se incluirían en este grupo las revistas científicas españolas de segundo orden que, o bien son poco citadas, o bien no cumplen con los estándares de publicación científica. También se incluyen las revistas internacionales de menor relevancia. - Indexadas en el segundo, tercer o cuarto cuartil según promedio de citas de cualquiera de las categorías del Índice de impacto de las revistas españolas de ciencias sociales o del Índice de impacto de las revistas españolas de ciencias jurídicas (Grupo EC3). Se toma como referencia los impactos acumulativos de los años 2004-2008. - Indexadas en DICE (Difusión de las revistas españolas de ciencias sociales y humanas) (Iedcyt) pero sin cumplir con el requisito de contar con evaluación por expertos. - Indexadas en el Catálogo Latindex **Grupo D (gD)**: este último grupo estaría conformado por todas aquellas publicaciones no incluidas en ninguna de las categorías anteriores y, por tanto, con un dudoso status científico. - Cualquier revista que no está indexada en alguno de los productos reseñados anteriormente. **Grupo de excelencia (gEx)**: integrado por las revistas con mayor grado de impacto científico, entendiendo como

		tales las posicionadas en el primer cuartil de los rankings internacionales de citación. - Para ciencias sociales: revistas indexadas en el primer cuartil según el Impact factor de cualquiera de las categorías del Journal citation reports (Thomson Reuters). - Para ciencias humanas: revistas indexadas en el Scimago journal rank (SJR, Elsevier) en las áreas arts & humanities y que están al mismo tiempo o bien en el A&HCI o bien en ERIH clasificadas como A.

Legenda

* Vedi http://www.esf.org/research-areas/humanities/erih-european-reference-index-for-the-humanities/erih-foreword.html

** Vedi http://www.arc.gov.au/era/era_2012/journal_list_dev.htm

*** Vedi http://epuc.cchs.csic.es/circ/categorias.html

Come si vede, nelle esperienze internazionali che hanno potuto beneficiare di periodi molto prolungati di preparazione della valutazione (circa due anni, nel caso dell'Australia, periodi di durata simile per Francia, Spagna e ESF) si è addivenuti a definizioni formali piuttosto elaborate.
Nell'esercizio dei GEV non si è giunti ad una definizione formale delle categorie, per le ragioni che saranno discusse *infra*. Già da ora sono peraltro relativamente chiari i criteri formali che potranno essere utilizzati per l'aggiornamento dei rating.
Come si è espresso il GEV 11 nei propri Criteri di valutazione:

> La prossima revisione sarà particolarmente importante. Si potranno allora utilizzare criteri oggettivi che non era possibile applicare retrospettivamente in modo meccanico (sta qui la ragione dell'ampio ricorso fatto in questi elenchi alla reputazione così come giudicata dalla comunità scientifica). (…)
>
> Tra questi criteri vi saranno:
> 1.I risultati della VQR 2004-2010;
> 2.La presenza in banche dati internazionali, come ISI e Scopus;
> 3.La presenza nei grandi repertori internazionali online, come J-STOR o Project Muse;
> 4.L'indicizzazione nei più rilevanti strumenti bibliografici internazionali;
> 5.La presenza nelle maggiori piattaforme digitali italiane;
> 6.L'utilizzo di una *peer review* ben organizzata, a doppio cieco e all'occorrenza verificabile;
> 7.La pubblicazione nelle principali lingue di cultura, oltre l'italiano;
> 8.La presenza nei cataloghi delle principali biblioteche italiane e internazionali;
> 9.La varietà e l'ampiezza del bacino da cui sono stati ricevuti gli articoli pubblicati, e la diffusione territoriale almeno su scala nazionale, e preferibilmente internazionale;

10. La pubblicizzazione della percentuale degli articoli “invitati” (cioè ricercati dalla rivista) su quelli pubblicati;
11. La pubblicizzazione della percentuale di articoli rifiutati sul totale di quelli ricevuti;
12. La regolarità e puntualità di pubblicazione;
13. La presenza di un buon sito internet.

Questo elenco non è naturalmente esaustivo, e va da sé che non verrà richiesta la conformità a ciascun criterio, ma che piuttosto sarà comparata la maggiore o minore adesione all’insieme di essi. E’ tuttavia importante sin d’ora offrire alle migliori riviste italiane chiare indicazioni sui possibili sviluppi futuri, così da stimolarne la qualità, il livello di internazionalizzazione e la visibilità.

Si pone quindi il problema di definire in modo univoco le categorie e di basare l’aggiornamento del rating dei GEV su tali definizioni.
Nell’immediato sarà utile iniziare un esercizio comparato, sia rispetto alle definizioni usate nel contesto internazionale, sia rispetto alla eventualità che riviste italiane siano state valutate anche in esercizi internazionali.

2.5 Scelta dei valutatori

Nelle esperienze internazionali si riscontrano varie soluzioni:
a) panel di esperti
b) società scientifiche
c) *referee* anonimi
d) consultazioni on-line.

2.5.1 Panel di esperti

Con la costituzione di panel di esperti si procede alla nomina, con procedure diverse da caso a caso, di gruppi di ampiezza variabile di studiosi di grande esperienza, potenzialmente in grado di esprimere giudizi su ampie classi di riviste.
I panel possono essere monosettoriali o plurisettoriali. La durata del lavoro di rating delle riviste non è definita a priori.
La valutazione dei panel può essere distorta se la loro composizione non riflette accuratamente la distribuzione degli interessi scientifici dell’intera comunità. In particolare è possibile che singoli membri del panel siano influenzati, anche involontariamente, da preferenze individuali e idiosincratiche relative a specifiche direzioni o aree di ricerca.
Per mitigare questa distorsione è possibile **sottoporre ai membri dei panel delle Linee Guida molto stringenti e dettagliate**. In aggiunta, è opportuno sottoporre i rating prodotti dai panel al giudizio di singoli esperti esterni.
Infine è rilevante la procedura di formazione del consenso all’interno del panel. Si possono formare regole di maggioranza, di veto o di unanimità. È importante acquisire informazioni dettagliate sulle procedure interne. Tuttavia la procedura più corretta consiste nella formulazione di giudizi indipendenti da parte di ciascuno dei membri del panel, con registrazione scritta e separata dei giudizi individuali su appositi formulari. In questo modo è possibile ricostruire l’effettivo grado di consenso sulle proposte di rating. È noto infatti dalla psicologia sociale che la dinamica di un piccolo gruppo può distorcere il giudizio individuale, allontanandolo da quello che si sarebbe formulato sotto condizioni di maggiore indipendenza (*groupthink:* Janis (1982)).
Più precisamente, **la metodologia che l’ANVUR raccomanda è la compilazione da parte dei membri del panel di schede individuali, redatte in modo indipendente dagli altri membri del panel**. La eventuale formulazione di un giudizio di sintesi, attraverso regole di formazione del consenso, è una informazione aggiuntiva ma potrebbe non essere utilizzata nella procedura.

Il vantaggio della proposta è che diventerà possibile misurare il grado di consenso tra giudizi individuali, cosa che sarebbe impossibile ricevendo solo il giudizio aggregato.

2.5.2 Società scientifiche

Una società scientifica è per definizione un gruppo di esperti di una data materia. Inoltre il rating delle riviste formulato da una società scientifica ha il vantaggio di acquisire autorevolezza e di facilitare il consenso nella comunità di riferimento.
Esistono tuttavia rischi connessi alla possibilità che gli organi di vertice delle società, verosimilmente più coinvolti nell'esercizio di valutazione, siano portatori di **visioni idiosincratiche della ricerca**. Inoltre potrebbe non esservi trasparenza circa la possibilità di **conflitti di interesse** tra membri degli organi direttivi delle società scientifiche e direttori di riviste o membri di comitati editoriali.
In generale quindi la valutazione da parte delle società scientifiche deve essere associata ad altri metodi. [18]

2.5.3 Referee anonimi

Una soluzione consigliabile consiste nella attivazione del parere delle società scientifiche e successivamente nella richiesta di opinioni indipendenti ad esperti internazionali, in forma anonima. Tali esperti dovrebbero non solo leggere riviste in italiano (quindi essere italiani che insegnano all'estero oppure colleghi stranieri in grado di consultare riviste italiane), ma anche utilizzarle ordinariamente nel lavoro scientifico.
Con una o due iterazioni tra società scientifiche ed esperti anonimi si dovrebbe convergere verso una classificazione accettabile.
L'uso di referee anonimi potrebbe anche essere attivato indipendentemente dalla procedura di validazione ex post di giudizi delle società scientifiche, ma come inizializzazione del processo.

2.5.4 Consultazioni on-line

In alcune esperienze internazionali (es. Spagna) si è attivata una **procedura di consultazione allargata di intere comunità scientifiche, attraverso l'utilizzo di apposite piattaforme software**. Preliminare a tale esercizio è il lancio di una campagna di sensibilizzazione allo scopo di ottenere un elevato tasso di risposta.
Allo scopo di incentivare la formulazione dei giudizi, potrebbe essere consigliabile strutturare la piattaforma in modo da consentire l'anonimato della valutazione, assegnando allo stesso tempo ad ogni docente o ricercatore un solo diritto di voto.
Naturalmente la consultazione on-line **potrebbe essere soggetta a rischi di manipolazione**, laddove una rivista possa contare su un ampio numero di sostenitori mobilitabili secondo proporzioni che potrebbero distorcere il giudizio finale.
Tale rischio potrebbe essere mitigato introducendo alcune regole, ad esempio:

- la consultazione non è valida se non intervengono almeno una certa proporzione dei docenti di una certa area (per evitare che una piccola minoranza possa manipolare il risultato);

[18] Un esempio interessante proviene dalla classificazione delle riviste giuridiche. La valutazione delle riviste scientifiche attraverso rating nel settore giuridico è discussa estesamente in Campbell, Goodacre and Little (2006), Svantesson (2009) e van Gestel e Vranken (2011). Su altri aspetti della valutazione in ambito giuridico si veda Sorensen (1994), Moed (2002).

- nel caso in cui vi sia una polarizzazione dei giudizi agli estremi della valutazione per una rivista, il giudizio viene sospeso (se alcuni giudicano eccellente una rivista e altri la giudicano modesta, è possibile che i primi siano portatori di conflitto di interesse);
- viene sottoscritta una dichiarazione di conflitto di interesse secondo la quale chi risponde si astiene dal valutare riviste nelle quali siede nel comitato editoriale, o inoltre svolge o ha svolto attività di direzione.

Si tratta di studiare con attenzione le potenzialità offerte da varie piattaforme software, il regime di anonimato implementabile, la strutturazione del testo per la assegnazione dei rating.

2.6 La procedura GEV all'interno della VQR

All'inizio della VQR i GEV hanno utilizzato la metodologia dei panel, con una procedura mista iterativa tra società scientifiche e referee anonimi. Ciò assicura autorevolezza ai rating prodotti.
Ecco come il documento di un GEV (Area 11) descrive la procedura seguita:

> Seguendo le indicazioni ricevute dall'ANVUR, la procedura che ha portato a questo risultato è stata articolata in quattro stadi:
>
> 1.Si sono chiesti alle Società e alle Consulte degli elenchi divisi in due fasce (A e B) delle riviste italiane, e nel caso internazionali, rilevanti per ciascun SSD, nonché delle riviste intersettoriali e interdisciplinari per esso più importanti. Per ciascuna fascia sono stati indicati dei tetti quantitativi. Solo una Società su più di 20, la Società Italiana di Filosofia Teoretica, ha ritenuto di non poter dare un elenco graduato nel modo richiesto, cosa di cui ci si rammarica e che non esclude un'auspicata collaborazione futura;
> 2.Questi elenchi sono stati sottoposti a dei revisori (*referee*) italiani e stranieri (in genere tre per elenco), scelti tra specialisti delle discipline che avevano trasmesso gli elenchi e selezionati in modo da evitare la sovra-rappresentazione di orientamenti particolari;
> 3.Il risultato dei referaggi è stato poi sottoposto alle Società e alle Consulte, che hanno fatto le loro controdeduzioni;
> 4.Gli elenchi così rivisti sono stati infine presentati ai sottogruppi in cui è stato suddiviso il GEV ANVUR di area 11.
>
> Tale procedura, approvata dall'ANVUR, ha dato risultati soddisfacenti. Naturalmente essa non può essere riuscita ad evitare del tutto gli errori, ma è sperabile che i quattro filtri utilizzati li abbiano ridotti al minimo.

Dal punto di vista metodologico, **la procedura ora identificata è corretta**, in quanto:

- attraverso il coinvolgimento delle società scientifiche ha esercitato, sia pure in tempi brevi, una mobilitazione della comunità scientifica, ovvero degli esperti depositari della conoscenza valutativa rilevante
- attraverso il ricorso a referee esterni anonimi, ha consentito la validazione delle proposte, mitigando i rischi di conflitto di interessi e di distorsioni
- con la imposizione di un tetto quantitativo riferito ai SSD ha di fatto mitigato il rischio di *grade inflation*, cioè di rincorsa all'inserimento delle riviste in fascia A allo scopo di competere con altre aree scientifiche (vedi oltre)
- con la distinzione tra riviste disciplinari e interdisciplinari ha preservato la varietà delle forme di comunicazione scientifica
- infine, con il ricorso alla approvazione finale in sede di GEV ha valorizzato la metodologia dei panel.

Si è trattato, in sostanza, tenuto conto dei limiti di tempo, di implementare una buona regola delle scienze sociali, che **impone di "triangolare" le osservazioni su un dato fenomeno quando non si**

disponga di misure affidabili. Ciascuna delle tre fonti adottate (società scientifiche, esperti, GEV) ha dei limiti, probabilmente insuperabili nel breve periodo. La procedura di triangolazione dovrebbe aver ridotto al minimo le distorsioni.
Chi invoca le esperienze straniere per chiedere che l'esercizio venga svolto in più anni, ignora che, se si fosse seguita questa strada, si sarebbe persa l'occasione straordinaria offerta dalla VQR e probabilmente non si sarebbe ottenuto alcun risultato. Con una procedura compressa nei tempi ma sostanzialmente corretta si è invece aperta una strada che potrebbe da ora solo migliorare.

2.7 Scelta del numero dei valutatori

Non esistono regole definite per la scelta del numero dei valutatori.
Data la natura esperta del giudizio, non si applicano di norma criteri di natura inferenziale in riferimento ad un campionamento dalla popolazione.
Tuttavia è opportuno che il numero dei valutatori sia considerato ai fini della misura di affidabilità (vedi oltre). **Più alto è il numero di valutatori e di categorie di assegnazione, minore è la soglia critica di consenso nel giudizio a cui corrisponde un valore accettabile**. Ciò aumenta la robustezza dei giudizi e aumenta la credibilità dell'intero esercizio.
L'intero esercizio ERIH è stato gestito da soli 140 ricercatori da 28 paesi. [19] SI tratta di un elemento di debolezza, che non a caso è spesso citato criticamente. Al contrario, gli esercizi di valutazione in Spagna si avvalgono di consultazioni online di molte centinaia di esperti, che sono tuttavia selezionati in base alla valutazione che ricevono in precedenza dalla agenzia di valutazione.
Nel caso italiano, la numerosità dovrà essere oggetto di attenta riflessione. In linea di massima l'obiettivo dovrebbe essere quello di coinvolgere nella classificazione tutti i ricercatori attivi, che hanno pubblicazioni negli ultimi cinque anni.

2.8 Missing data

Un problema tecnico che potrebbe generare notevoli difficoltà di ordine pratico è rappresentato dai **dati mancanti** (*missing data*), cioè dai casi nei quali i valutatori non esprimono il giudizio su una rivista. La presenza di dati mancanti rende difficile il calcolo della *inter-rater reliability*, perché la assenza di giudizio non può che essere considerata come una mancanza di accordo. Ciò comporta un aumento della soglia critica oltre la quale si può ritenere che i valutatori abbiano raggiunto un accordo.
La soluzione proposta in letteratura è rappresentata dalla compilazione di tabelle nelle quali la mancata risposta è considerata come una variabile aggiuntiva, e gli indicatori di affidabilità sono calcolati al netto di questa variabile.

2.9 Vincoli alla assegnazione alle categorie

In condizioni ideali si dovrebbero lasciare liberi i valutatori di esprimere un giudizio di qualità non vincolato, rispondente puramente alla definizione fornita. La appartenenza delle riviste alle categorie discenderebbe esclusivamente dalla presenza di un grado di consenso giudicato accettabile secondo regole quantificate di *reliability*.
Tuttavia occorre considerare un problema, largamente conosciuto da chi si occupa di valutazione, che **consiste nella tendenza dei valutatori a "inflazionare" il giudizio assegnato ai prodotti della propria area scientifica, quando ritengono che esso possa influenzare la distribuzione**

[19] http://www.esf.org/research-areas/humanities/erih-european-reference-index-for-the-humanities/erih-foreword.html

delle risorse. Questo fenomeno, denominato *grade inflation*, è stato identificato fin dagli anni ’60 al National Health Institute (NIH) negli Stati Uniti. Ciò che accadeva era che i referee delle singole aree medico-scientifiche assegnavano punteggi secondo verità ai progetti presentati se sapevano che le risorse erano state pre-assegnate alle aree scientifiche, in modo che i punteggi servivano solo ad allocare un budget definito ai progetti migliori. Al contrario, quando sapevano che i punteggi entravano come elementi di decisione in un panel di secondo livello dell’Istituto, il quale provvedeva ad allocare il budget complessivo in funzione delle valutazioni, gli stessi esperti si ritenevano in dovere di inflazionare i punteggi, allo scopo di difendere la propria area scientifica. In altre parole, ritenevano di doversi comportare da partigiani, pur non riscontrando alcuna contraddizione con l’etica rigorosa della valutazione. [20]
È possibile che lo stesso fenomeno possa affliggere il rating di riviste?
In linea di principio si potrebbe ritenere che la valutazione di riviste non abbia una relazione diretta con la assegnazione di risorse. Tuttavia essa di fatto contribuisce ad una particolare forma di assegnazione di risorse- risorse di prestigio e legittimazione. Non v’è dubbio che, sotto questa forma, esperti appartenenti ad una certa area scientifica possano sentire il dovere di affermare che il numero di riviste di fascia più alta siano in numero elevato, a dimostrazione della qualità complessiva della ricerca nell’area stessa, anche nel confronto con altre. Si ritiene quindi ragionevole ipotizzare la presenza di *grade inflation* anche nel rating di riviste. In particolare, ci si attende una tendenza ad assegnare un numero eccessivo di riviste alla classe più elevata.

Le soluzioni comunemente proposte per mitigare la *grade inflation* consistono in varie forme di normalizzazione dei risultati, allo scopo di sterilizzare la competizione tra aree. Nel caso in questione è possibile pensare a due soluzioni:

- imporre che alla classe più elevata possa essere assegnato non più di una quota prefissata (es. 20%) delle riviste
- imporre un numero massimo di riviste a cui può essere assegnata la classe più elevata, eventualmente calibrando tale numero in proporzione alla ampiezza dell’area scientifica sottostante.

La prima soluzione ha il pregio di costituire una distribuzione interna di qualità. Tuttavia si presta ad una agevole manipolazione, che consiste **nell’aumento indiscriminato della coda della distribuzione, attraverso l’inserimento nella lista delle riviste da valutare di titoli minori**. In questo modo si inflazione la possibilità di inserire riviste nella fascia superiore. Più precisamente, una condizione preliminare alla definizione delle classi secondo quantili è che la lista delle riviste su cui si esercita la analisi sia chiusa- ovvero, sia stato svolto un preliminare esercizio di definizione di cosa costituisce una “rivista scientifica”. Tale definizione è controversa in molti casi rilevanti, per esempio in discipline con ampie ricadute applicative, nelle quali il confine tra rivista scientifica e rivista professionale è sovente dibattuto. È appena il caso di ricordare che **la definizione di rivista scientifica non è presente nel nostro ordinamento,** ancorché sia prevista da tempo nel decreto istitutivo della Anagrafe delle pubblicazioni (ANPRePS).
La seconda soluzione elimina alla radice la possibilità di manipolazione ora descritta ed è quindi consigliabile in sede di prima applicazione. Si tratta della scelta effettuata dai GEV delle aree CUN 10-14 in sede di criteri di valutazione.
Per una messa a regime del sistema è opportuno che la numerosità delle classi, in particolare della classe A, sia rimessa ad un aggiornamento successivo, da realizzarsi con la metodologia illustrata di seguito.

[20] Sull’impatto della valutazione della ricerca sulle modalità di finanziamento delle università, e quindi sul comportamento individuale dei ricercatori è aperto un dibattito, su cui da ultimo vedi Hicks (2012), basata in larga parte su OECD (2010). Sull’esperienza inglese, vedi HEFCE (1997); su quelle australiana Butler (2003a; 2003b).

2.10 Aggiornamento della classificazione dei GEV e procedura di rating

Si pone adesso il problema dell'aggiornamento e della trasformazione dell'esercizio dei GEV in una risorsa permanente.
L'aggiornamento del lavoro svolto dai GEV si impone per diverse ragioni. In primo luogo perché la VQR produrrà elementi informativi preziosi circa la qualità dei singoli lavori pubblicati sulle riviste appartenenti alle varie classi di merito. Ciò potrò portare ad una conferma o ad una modifica del rating iniziale. [21]
In secondo luogo è opportuno coinvolgere un numero più ampio di studiosi in un processo collettivo.
In terzo luogo la classificazione svolta dai GEV, per sua natura, è prevalentemente orientata a catturare le fasce alte delle riviste scientifiche. Infatti la VQR, in quanto si basa sulla sottomissione volontaria di soli tre prodotti nel settennio 2004-2010, verosimilmente consente ai ricercatori di sottoporre i migliori prodotti, spingendo per così dire verso l'alto la qualità. È ragionevole assumere che i singoli atenei abbiano un problema ulteriore, che consiste nella classificazione di *tutte* le riviste su cui pubblicano i propri docenti, per esempio al fine di assegnare risorse locali di ricerca. In questo senso la classificazione VQR e quella di singoli atenei non sono concorrenziali, ma complementari.
Infine, occorre ricordare che la reputazione delle riviste, come più in generale della produzione scientifica, non può essere concepita come una acquisizione irreversibile, ma deve essere continuamente soggetta a verifica.

Come ha ricordato Ronald Rousseau:

> Christenson & Sigelman (1985) found that scholarly journals in sociology and political sciences tend to establish reputations that endure in spite of what they merit. Once a journal has been placed on a discipline's prestige ladder, it tends to retain its place because its reputation is accepted at face value. Such journals are not re-evaluated in the light of changing circumstances. Comparing prestige scores with impact scores showed that good and bad reputations tend to be exaggerations of what impact scores suggest are merited. This clearly is a form of the Matthew effect (Merton, 1968): Already famous persons (or journals) receive more credit than they actually deserve, while recognition of less prestigious scientists (or journals) is withheld (Rousseau, 2002).

Si tratta quindi di sviluppare una metodologia che consenta di rendere permanente la creazione e la gestione di un grande archivio nazionale di riviste delle aree umanistiche e sociali.
Questa attività vedrà impegnata l'ANVUR al massimo livello, i GEV delle aree 10-14, e le comunità scientifiche nazionali. Una opportunità è fornita dalla attività di rating delle riviste in corso e in programma presso vari atenei, in particolare all'interno della collaborazione tra Università di Bologna, Padova, Milano Statale e Torino, nonché a Roma La Sapienza.
L' ANVUR propone di svolgere l'esercizio tra i vari atenei secondo la metodologia ora indicata (tracciatura dei giudizi individuali), e inoltre di sviluppare una metodologia di meta-analisi per combinare giudizi provenienti da metodi diversi.
La metodologia di seguito proposta è ritenuta condizione indispensabile perché singole valutazioni di ateneo possano aspirare a essere riconosciute a livello nazionale dalla Agenzia.
È prematuro stabilire quando l'esercizio di aggiornamento verrà iniziato e chiuso, e con quale periodicità verranno pubblicati i dati aggiornati. Alla fine del 2012, a VQR largamente avviata, sarà possibile pianificare con maggiore precisione le date.

[21] Sul tema della correlazione tra peer review e indicatori citazionali esiste una letteratura specializzata, in larga parte basata sul RAE inglese: Oppenheim (1995; 1997); Norris e Oppenheim (2003; 2010); Oppenheim e Summers (2008). Sulla correlazione tra rating delle riviste nelle aree umanistiche e sociali e indicatori citazionali si veda Haddow e Genoni (2010).

Si raccomanda la adozione di una procedura uniforme e standardizzata, che consenta la messa a regime del sistema di valutazione.
Si suggeriscono i seguenti passi.

A. Composizione dei panel

Non è necessario che i panel abbiano la stessa dimensione in termini di numero dei valutatori. È tuttavia importante che le regole di reclutamento degli esperti siano esplicite e comunicate.

B. Definizione delle categorie

È necessario definire per scritto una descrizione dettagliata e non ambigua della qualità delle riviste per classi. Le classi possono essere ridotte a tre (classe A, B e C).
È opportuno che la definizione sia accompagnata dalla esemplificazione di indicatori, anche quantitativi, che si ritengono associati alla qualità.
La definizione di qualità deve essere testata per la chiarezza del testo e la non ambiguità dei significati, attraverso la richiesta di opinioni a soggetti diversi.

C. Assegnazione delle riviste

È necessario che ad ogni panel sia assegnato lo stesso set di riviste da valutare, eventualmente anche con la composizione di sub-panel per aree omogenee.
Se alcuni membri del panel non si ritengono titolati ad esprimere giudizi su particolari gruppi di riviste, allora è opportuno estrarre queste riviste dal set generale e farle valutare da un numero inferiore di valutatori.

D. Assegnazione alle riviste delle categorie

La assegnazione delle categorie alle singole riviste dovrebbe essere svolta individualmente dagli esperti e registrata in appositi formulari (Appendice 1).
Nel caso in cui il panel ritenga necessario pervenire ad una valutazione di sintesi, essa viene registrata nel formulario e archiviata. Essa entrerà come elemento di giudizio di una procedura meta-analitica.
Sui giudizi individuali si effettueranno le misure di affidabilità.

E. Calcolo dell’indice di affidabilità

Disponendo di valutazioni individuali, provenienti da esperti di vari atenei, sarà possibile calcolare indici di affidabilità (vedi oltre) al livello più dettagliato.
In assenza del giudizio individuale, si potrebbe considerare giudizio individuale quello di sintesi formulato da ogni singolo panel di ateneo, ottenendo una procedura a 4 valutatori (Bologna, Padova, Torino, Milano Statale).
Il vantaggio del calcolo su giudizi individuali consiste nella maggiore affidabilità delle misure di consenso.
In ogni caso, se i giudizi individuali (di singoli esperti e/o dei panel) producono elevati indici di affidabilità, in riferimento alle opportune tavole di valori di riferimento, essi possono essere pubblicati.

2.11 Definizione di affidabilità (*reliability*)[22]

La assegnazione di una rivista ad una categoria costituisce un atto di giudizio (*rating*), effettuato da un valutatore (*rater*), di norma esperto nel settore.
La natura esperta della valutazione non elimina il bisogno di validazione inter-soggettiva. Diversi esperti potrebbero assegnare lo stesso oggetto a categorie diverse, con legittime motivazioni. È importante sottolineare che questo problema non può essere evitato.
Sorge dunque il problema metodologico di definire e misurare il grado in cui esiste accordo tra diverse valutazioni effettuate dallo stesso soggetto nel tempo e tra diversi valutatori.
L'accordo ha due dimensioni: accordo intra-individuale (*intra-rater agreeement*) e accordo inter-individuale (*inter-rater agreement*).
Il primo si riferisce alla coerenza nel tempo dei giudizi soggettivi. Vi sono molte buone ragioni per cui, al di là di situazioni patologiche individuali, lo stesso soggetto potrebbe assegnare lo stesso oggetto a classi diverse nel tempo.
Poiché tuttavia l'approccio che qui viene proposto si sulla affidabilità di giudizi forniti da numerosi valutatori, il problema della coerenza individuale non verrà trattato.
L'accordo inter-individuale si riferisce al grado in cui, dati N valutatori diversi nello stesso tempo t, l'oggetto x è assegnato alla categoria k da una proporzione elevata di essi. Se l'accordo è elevato, la probabilità che tutti gli oggetti valutati siano assegnati alle stesse classi dai diversi valutatori è elevato.

[22] Questa sezione si basa su Gwet K.L. (2010). Si vedano inoltre Carmines e Zeller (1979), Traub (1994), Kirk e Miller (1986) e Fleiss, Levin e Paik (2003).

Box 2
Definizioni di accordo

Intra-rater agreement
Grado di accordo che un individuo esprime rispetto a giudizi espressi sullo stesso oggetto ma in tempi differenti, sotto le stesse condizioni sperimentali. Esiste alto accordo intra-individuale se è elevata la probabilità che l'oggetto x sia assegnato alla categoria k dal valutatore j per t= 1, 2…T. Essa viene anche detta **Test-retest reliability**.

Inter-rater agreement
Grado di accordo che due o più valutatori ottengono esprimendo un giudizio sullo stesso oggetto, in modo indipendente e sotto le stesse condizioni sperimentali, nello stesso tempo.

Occorre distinguere tra consenso e misura dell'accordo. La misura dell'accordo è definita come affidabilità (*reliability*), [23] in particolare affidabilità del giudizio inter-individuale (*inter-rater reliability*) (Box 1).
Intuitivamente, si potrebbe pensare che una buona misura dell'accordo tra soggetti sia rappresentata dal numero di valutazioni condivise rapportato al numero totale di valutazioni effettuate.
Si consideri la Tabella 2. Supponiamo che due soggetti A e B si trovino a valutare l'opportunità di dare un finanziamento a progetti per i quali non si disponga di indicatori quantitativi predefiniti, ma ci si debba affidare a giudizi soggettivi. Una volta che i soggetti hanno espresso i propri giudizi, si compila una tabella che esprime gli stessi in percentuale del totale di giudizi, in modo da poter interpretare i numeri come frequenze relative, che ai fini della analisi possono essere interpretate come probabilità.
Si potrebbe pensare quindi che il grado di consenso tra A e B sia dato dalla percentuale di progetti totali per i quali hanno espresso lo stesso giudizio, accettando entrambi il progetto o rigettandolo entrambi. La coincidenza di giudizi è leggibile lungo la diagonale principale. Tale misura sarebbe quindi pari a (30+50)/100= 0.80, che denoterebbe un buon grado di accordo. In realtà questa misura non è corretta.

Tabella 2

		Valutatore A		
		Accettabile	Non accettabile	Totale
Valutatore B	Accettabile	30	15	45
	Non accettabile	5	50	55
	Totale	35	65	100

Perché l'intuizione di base che sta dietro a questa conclusione è fallace?

[23] "Reliability refers to the degree of consistency with which instances are assigned to the same category by different observers or by the same observer on different occasions. For reliability to be calculated, it is incumbent on the scientific investigator to document his or her procedure and to demonstrate that categories have been used consistently" (Silverman, 2000, 188).

La risposta è che l'effettivo grado di consenso deve essere misurato depurando la quota di casi in cui si ha coincidenza di giudizio *dai casi nei quali la coincidenza avviene per ragioni puramente casuali.* Poiché il numero delle categorie è finito, ogni valutatore deve obbligatoriamente assegnare ogni oggetto ad una delle categorie disponibili. Ciò significa che **in un certo numero di casi la coincidenza di giudizio potrebbe aver luogo indipendentemente dalla circostanza che i valutatori condividano effettivamente la decisione sottostante al giudizio.**
Quindi una buona misura dell'affidabilità del giudizio parte dalla proporzione sul totale del numero di casi di coincidenza, ma sottrae a questa una qualche misura della proporzione di coincidenze puramente casuali. È importante sottolineare che, mentre questo principio è universalmente accettato, in letteratura sono state proposte diverse misure delle coincidenze puramente casuali.

La più utilizzata è l'indice Kappa, introdotta da Cohen (1960). Essa si basa sulla definizione delle coincidenze di giudizio dovute al caso come somma dei prodotti delle probabilità marginali. Vediamo il ragionamento sottostante.
In tabella 2 la probabilità che il valutatore B consideri accettabile un progetto è pari a 0.45, ottenuto come rapporto tra il numero dei casi (numero che compare nella colonna al margine della tabella, e per questo viene definito "marginale") e il numero totale. Allo stesso modo la probabilità che il valutatore A ritenga accettabile un progetto è pari a 0.35, valore che si ottiene leggendo il numero nella riga marginale in basso e rapportandolo a 100. Quindi la probabilità che A e B diano contemporaneamente il giudizio "accettabile" è data dal prodotto delle probabilità, ovvero 0.45* 0.35= 0.1575. Questa probabilità può essere considerata una misura dell'accordo sulla valutazione "accettabile" che può avvenire per caso.
Esattamente nello stesso modo, la probabilità che entrambi i valutatori diano un giudizio "non accettabile" si ottiene moltiplicando le probabilità ai margini, ovvero 0.65*0.55 = 0.3575. Quindi la probabilità che due giudizi siano coincidenti per puro caso è data dalla somma delle probabilità così ottenute.
Si diano le seguenti definizioni

p_a = probabilità di accordo tra due giudizi (*agreement probability*)

p_e = probabilità di accordo casuale tra due giudizi (*expected chance agreement rate*)

Nell'esempio sopra riportato avremmo dunque:

$p_a = (30/100) + (50/100) = 80/100 = 0.80$

$p_e = (35/100) * (45/100) + (65/100) * (55/100) = 0.1575 + 0.3575 = 0.515$

Cohen (1960) ha proposto una semplice definizione di affidabilità, che è data dalla formula

$$Kappa = \frac{(pa - pe)}{1 - p_e}$$

Nel caso in questione avremmo

Kappa = (0.80 – 0.515) / 1 – 0.515 = 0.285 / 0.485 = 0.59

Come si vede, l'indice Kappa **restituisce una misura dell'accordo tra i due valutatori significativamente inferiore rispetto alla misura "intuitiva**" calcolata solo sulla diagonale

principale della matrice. Ciò è molto importante: se ogni osservatore sarebbe pronto a concedere che un accordo dell'80% tra due valutatori significa che i risultati sono affidabili, con un indice inferiore al 60% potrebbero sorgere dubbi sull'effettivo consenso raggiunto.

La misura di Cohen (1960) vale per la assegnazione di giudizi a categorie nominali. Per questo caso vale la regola che "ogni disaccordo è disaccordo totale". Tale regola non vale, come si è anticipato, per le variabili ordinali o per intervalli e rapporti: in questi casi "alcuni disaccordi sono accordi parziali", nel senso che se i valutatori assegnano valori diversi ma in categorie contigue, il loro disaccordo è di fatto inferiore. Se si applicasse l'indice Kappa a queste situazioni si otterrebbe una pesante sottostima dell'effettivo accordo tra valutatori.
La determinazione dell'indice Kappa per categorie ordinali o continue richiede il calcolo delle distanze Euclidee tra vettori; per il suo calcolo si rimanda a Gwet (2010, capitolo 3).

La metrica di accordo tra valutatori di Cohen è stata all'origine di una ampia letteratura ed è correntemente usata in molte aree delle scienze sociali e mediche.
Nella prassi scientifica si confronta il valore trovato con un benchmark, che in genere viene suggerito in letteratura in riferimento a classi di problemi derivanti da varie aree (cliniche, sociali etc.). Una procedura più rigorosa si ottiene simulando, con metodo Monte Carlo, un numero molto alto di distribuzioni che si otterrebbero assegnando i valori in modo casuale, e derivando da queste la soglia critica oltre la quale deve essere rigettata la ipotesi che l'accordo osservato sia solo casuale. In Gwet (2010) sono forniti i dettagli della procedura e le tabulazioni dei valori critici in riferimento a:

- numero di valutatori
- numero di categorie
- numero di oggetti valutati.

Sono state successivamente proposte altre metriche, anche per correggere alcuni limiti dell'indice Kappa.

Ai fini dell'esercizio di rating delle riviste, si propone di studiare la seguente procedura:

- i giudizi individuali degli esperti sono registrati e tabulati
- viene effettuato il calcolo dell'indice Kappa per categorie ordinali, nonché di altri indici presenti in letteratura
- viene confrontato l'indice con le tabelle dei valori critici
- viene pubblicato il rating solo per il sottoinsieme delle riviste per i quali i giudizi dei valutatori superano la soglia critica (o in alternativa, vengono pubblicati tutti i rating con associato l'indice di affidabilità e la soglia critica).

2.12 Verso una meta-analisi

In preparazione al lancio di procedure di aggiornamento del rating delle riviste effettuato dai GEV, è utile iniziare una meta-analisi dei risultati disponibili nelle esperienze internazionali.
Ciò sarà possibile creando in sede ANVUR un modello di analisi la cui struttura potrebbe essere sintetizzata come da Tabella 3.

Una volta raccolti i dati, occorrerà sviluppare un modello meta-analitico che confronti accuratamente i sotto-insiemi di riviste per i quali sono disponibili più valutazioni e stabilisca il grado di comparabilità. Infatti i giudizi che sarà possibile raccogliere per un sotto-insieme di riviste sono stati prodotti a partire da categorie diverse, definite con assunzioni a priori differenziate, in tempi e con procedure non omogenei.

La metodologia della meta-analisi fornisce una definizione quantitativa del grado di generalizzabilità di risultati eterogenei.

Tabella 3
Modello di analisi dei rating delle riviste di area umanistica e sociale nelle esperienze internazionali

Rivista	Indicizzazione			Rating								
	IF	SJR	Altri	VQR	Università di Bologna	Università di Roma La Sapienza	Panel altri atenei	ERA 2010	AERES 2008	CRIC	ERIH	CNRS
1												
2												
….												

Le fonti dei dati sono reperibili come segue:

ERA 2010 (prima del ritiro del rating delle riviste per l'ERA 2012):
http://www.arc.gov.au/era/era_2012/journal_list_dev.htm

AERES 2008 (prima della pubblicazione delle sole liste di riviste scientifiche senza rating, attualmente in vigore):
http://www.aeres-evaluation.fr/index.php/Publications/Methodologie-de-l-evaluation/Listes-de-revues-SHS-sciences-humaines-et-sociales

CRIC http://epuc.cchs.csic.es/circ/categorias.html

ERIH: http://www.esf.org/research-areas/humanities/erih-european-reference-index-for-the-humanities/erih-foreword.html

CNRS: http://www.cnrs.fr/comitenational/sections/doc/categorisation37_0911.pdf (solo riviste di Economia e Management)

3 Pubblicazione di informazioni validate sulle procedure editoriali e di selezione dei manoscritti da parte di editori nazionali

In questa sezione del programma di attività si tratta di attivare un Gruppo di studio con la Associazione Italiana degli Editori per la creazione di una griglia di indicatori volti ad asseverare le modalità con cui gli editori gestiscono la sottomissione e la selezione dei manoscritti.
Le situazioni di base possono essere identificate a partire dalla tabella predisposta dall'ANVUR ai fini della delibera con cui ha promosso l'avvio delle procedure per la Anagrafe nominativa dei professori ordinari e associati e dei ricercatori, contenente per ciascun soggetto l'elenco delle pubblicazioni scientifiche prodotte (ANPRePS).

Variabili per la tipologia: **Monografia**	
Codice identificativo della pubblicazione (ID)	ID
Codice identificativo autore (CF)	C.F.
Anno di pubblicazione	*tttt*
Altri autori	nomi e cognomi
Titolo pubblicazione	
Lingua	selezione su apposito elenco
Paese di pubblicazione	selezione su apposito elenco
DOI (digital object identifier)	
Titolo libro	
Editore	Nome Paese Città
Formato pubblicazione	☐ Stampa ☐ Elettronico
Accettazione della pubblicazione subordinata a superamento procedura di referaggio	☐ Si ☐ No
Tipologia di procedura di referaggio utilizzata (Se SI a punto precedente):	☐ Incarico ad esperti ☐ anonimi ☐ non anonimi ☐ Presenza di un Comitato scientifico, o organismo equivalente, che effettua la procedura di referaggio su ogni prodotto sottoposto
ISBN	
N° volume	

4 Creazione di un archivio di riviste italiane disponibili in formato digitale e di metadati e referenze tratte da monografie in lingua italiana

Si tratta di iniziare uno studio di fattibilità sugli archivi digitali attualmente disponibili delle riviste italiane. Allo stato dell'arte sono disponibili due grandi piattaforme di tipo commerciale, una riferibile ad un editore, una ad un distributore. Occorre aprire un dialogo operativo con tali soggetti per verificare la disponibilità a sostenere un esercizio nazionale, di carattere pubblico (non commerciale) di messa a disposizione di riviste in formato digitale al solo scopo di svolgere analisi valutativa e bibliometrica. Potrebbero essere mobilitati anche i consorzi interuniversitari di calcolo.

Questa linea di attività si dovrà coordinare con le iniziative a livello europeo sorte sotto l'egida di un Gruppo di progetto coordinato da Ben Martin (Martin, 2010).

In parallelo a questa iniziativa sulle riviste si potrebbe anche studiare la possibilità di indicizzare sperimentalmente degli insiemi di monografie.
Come è noto, nelle aree umanistiche e in parte nelle scienze sociali ha un peso preponderante la monografia, intesa come un prodotto di ricerca complesso, frutto di una attività prolungata di ricerca, prodotto con minore frequenza e con intervalli lunghi. Vi è evidenza empirica [24] del fatto che le monografie, rispetto alle riviste scientifiche:

- citano un insieme più ampio di prodotti editoriali (non solo altri articoli e –limitatamente- monografie, ma monografie, documenti non pubblicati, letteratura grigia e altro)
- citano più frequentemente lavori di altre discipline e non solo quelli strettamente disciplinari.

Si pone l'obiettivo di introdurre i materiali monografici all'interno delle basi di dati a partire dalle quali sia possibile sviluppare indicatori bibliometrici. Attualmente questo obiettivo è reso impossibile dal fatto che le monografie, tranne rare eccezioni, non sono indicizzate nelle basi di dati.
Il problema potrebbe venire risolto in futuro, se i prodotti monografici potessero accedere a piattaforme digitali di Open Access. Questa ipotesi tuttavia comporta significativi problemi di riallocazione dei costi della produzione editoriali ed è di non immediata implementazione. [25]

Una soluzione realistica potrebbe essere basata sui seguenti passi:

- viene aperta dall'ANVUR una piattaforma software
- la piattaforma è organizzata per aree disciplinari
- vengono definite delle regole di accesso e di uso
- vengono definite delle regole di validità in riferimento alla numerosità attesa delle monografie
- i ricercatori appartenenti alle comunità scientifiche sono invitati a sottomettere volontariamente:
    - file di testo contenente tutte le referenze della monografia
    - metadati relativi alla monografia

entro un certo intervallo di tempo, in riferimento alle monografie da essi pubblicate

- i file di testo subiscono un processo di parsing e di annotazione con procedure automatiche

[24] Tra le differenze importanti nei pattern di pubblicazione vi è il ruolo delle monografie e delle collane editoriali, sui quali vedi Clements et al. (1995) e Lariviére et al. (2006).
[25] L'impatto delle tecnologie digitali e delle prospettive di pubblicazione dei risultati scientifici in Open Access sulla valutazione della ricerca è discusso in una serie di contributi della comunità italiana delle biblioteconomia: Comba (2003), De Robbio (2004; 2009), Cassella (2010; 2011), Cassella e Bozzarelli (2011).

- a partire dai dati annotati vengono costruiti
    - analisi delle strutture citazionali (natura delle pubblicazioni citate, ritardi nelle citazioni, persistenza etc.)
    - indici citazionali
    - analisi dei grafi di citazioni
- laddove le analisi conducessero a risultati affidabili, i risultati vengono pubblicati.

Nella fase di pre-fattibilità occorre:

- verificare i profili inerenti alla gestione di eventuali diritti di proprietà su porzioni delle monografie (liste di referenze)
- esaminare l'offerta di mercato di piattaforme software per la gestione di archivi bibliografici caricabili in forma distribuita
- esaminare lo stato dell'arte di software per l'analisi di grafi ad alta dimensionalità a partire da dati bibliometrici.

Con lo svolgimento congiunto di questi studi di fattibilità sarebbe possibile anche confrontare le strutture citazionali di riviste e monografie, aggiungendo elementi di grande interesse al dibattito in corso sulla valutazione bibliometrica nelle aree umanistiche e sociali.

5. Sperimentazione di indicatori non citazionali

Gran parte della bibliometria si basa su indicatori citazionali, ritenuti da una robusta letteratura la migliore approssimazione al concetto di qualità della ricerca intesa come impatto sulle comunità scientifiche.
Tuttavia da alcuni anni sono in corso sperimentazioni volte a verificare la possibilità di utilizzare indicatori che rilevano altre dimensioni della circolazione delle pubblicazioni, in particolare l'utilizzo della pubblicazione. Si ritiene in altri termini che la citazione catturi una dimensione, sicuramente la più importante, dell'impatto delle pubblicazioni, ma non sia esaustiva. Un lavoro potrebbe essere letto ma non citato, potrebbe quindi avere un impatto che va oltre la stessa citazione. [26]
Inoltre la progressiva creazione di forme di editoria elettronica consente di ottenere dati di utilizzo molto dettagliati, in due filoni principali: (a) dati di utilizzo pubblicamente disponibili su risorse elettroniche libere (es. Open Access); (b) dati di utilizzo prodotti e/o utilizzati da editori commerciali, integratori, o distributori.

East (2006) ha utilizzato i seguenti indicatori per la classificazione delle riviste nelle aree umanistiche in lingua nazionale:

- numero di sottoscrizioni da parte di grandi biblioteche accademiche all'estero
- indicizzazione in database internazionali dedicati alle aree umanistiche
- pratiche editoriali (uso di referee)
- giudizi di qualità formulati da esperti
- frequenza di citazioni nelle riviste indicizzate da ISI.

I suoi risultati sperimentali sono incoraggianti circa la significatività di indicatori non citazionali (in particolare le sottoscrizioni e la presenza nei database) e la elevata correlazione sia con indicatori citazionali che con i giudizi esperti.

Tra gli indicatori non citazionali verranno quindi esplorati:

- indicatori di utilizzo (*journal usage factor*) [27]
- indicatori di disponibilità in cataloghi [28]
- indicatori di uso basati su web [29]
- recensioni di monografie.

È opportuno predisporre uno studio di pre-fattibilità per esplorare potenzialità e limiti di tali indicatori.

---

[26] L'uso delle citazioni come indicatore esclusivo di qualità della ricerca è contestato da una parte della letteratura bibliometrica, che suggerisce un approccio integrato (vedi Linmans, 2010).
Con indicatori non citazionali ci si riferisce ad una ampia serie di indicatori di utilizzo. Bollen et al. (2005), Bollen e van de Sompel (2008), e Bollen et al. (2009) discutono le relazioni tra vari indicatori di impatto, sia citazionali che non citazionali.
[27] Per gli indicatori di utilizzo (*journal usage factor*) si veda CIBER (2011).
[28] Tsay (1998) e Torres-Salinas e Moed (2009) discutono indicatori di disponibilità in cataloghi di biblioteche e Hicks e Wang (2009) la creazione di database per le scienze umane e sociali.
[29] Con lo sviluppo della rete Internet si sono aperte nuove prospettive di valutazione basate su metriche di accesso e di utilizzo. Si vedano Björneborn e Ingwersen (2004), Vitiello (2005), Noruzi (2006), e Brody, Harnad e Carr (2006).

APPENDICE 1
Modulo di rilevazione dei giudizi individuali di rating delle riviste

Caso base: 4 valutatori, 10 riviste, 3 categorie (A, B e C)

Riviste	Valutatori			
	I	II	III	IV
1	A	A	A	C
2	A	A	B	C
3	A	A	B	C
4	A	A	C	C
5	A	B	A	A
6	B	A	A	A
7	B	B	B	B
8	B	C	B	B
9	C	C	B	B
10	C	C	C	C

APPENDICE 2
Trasformazione del modulo di rilevazione di base ai fini del calcolo del coefficiente di accordo

Caso base: 4 valutatori, 10 riviste, 3 categorie (A, B e C)

Riviste	Categorie			Totale
	A	B	B	
1	3	0	1	4
2	2	1	1	4
3	2	1	1	4
4	2	0	2	4
5	3	1	0	4
6	3	1	0	4
7	0	4	0	4
8	0	3	1	4
9	0	2	2	4
10	0	0	4	4

REFERENZE

Abramo G., D'Angelo C.A., Di Costa F. (2010) Citations versus journal impact factor as proxy of quality: could the latter ever be preferable? *Scientometrics* , vol. 84, 821–833

Amin M., Mabe M. (2000) Impact factors: Use and abuse. *Perspectives in Publishing*, no. 1, October.

Archambault E., Vignola-Gagne E. Cote G., Lariviére V., Gingras Y. (2006) Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics*, Vol. 68, no. 3, 329–342.

Archambault, E., Lariviere, V. (2009) History of the journal impact factor: Contingencies and consequences. *Scientometrics*, vol. 79 (3), 635-649.

Arts and Humanities Research Council (2011) The impact of AHRC research. First annual *Research Performance and Economic Impact* Report. London, AHRC.

Aru S., Celata F., Rondinone A., Rossi U., Santini C. (2010) L'Università che cambia, la valutazione della Ricerca, il ruolo delle riviste scientifiche - *Rivista geografica italiana*, no.1.

Baccini A. (2010) Valutare la ricerca scientifica: uso ed abuso degli indicatori bibliometrici,Bologna, Il Mulino.

Balinski M., Laraki R. (2011) Majority judgment. Measuring, ranking, and electing. Cambridge, Mass., The MIT Press.

Bartolini, S. (2007) *Come migliorare la qualità della ricerca in Italia? L'esperienza della valutazione triennale della ricerca nelle Scienze politiche e sociali (2001-2003)*, in *Sociologica*, 17 settembre.

Biolcati-Rinaldi F. (2010) Quali indicatori bibliometrici per le scienze sociali? *Working Paper 02/2010* Dipartimento di Studi Sociali e Politici Facoltà di Scienze Politiche, Università di Torino.

Björneborn L., Ingwersen P. (2004) Toward a Basic Framework for Webometrics. *Journal of the American Society for Information Science and Technology*, vol. 55 (14):1216–1227.

Bollen, J., Van de Sompel, H., Smith, J., & Luce, R. (2005). Toward alternative metrics of journal impact: A comparison of download and citation data. *Information Processing and Management*, vol. 41, no.6, 1419–1440.

Bollen J., Rodriguez M.A., van de Sompel H. (2006) Journal status. *Scientometrics,* Vol. 69, no. 3, 669–687.

Bollen J. ,Van de Sompel H. (2008) Usage Impact Factor: The effects of sample characteristics on usage-based impact metrics. *Journal of the American Society for Information Science and Technology*, vol. 59, no. 1, 136-149.

Bollen J., Van de Sompel H., Hagberg A., Chute R. (2009) A Principal Component Analysis of 39 Scientific Impact Measures. *PLoS ONE* ,1 June 2009, vol. 4, issue 6 .

Bonaccorsi A., Daraio C. (2004) Econometric approaches to the analysis of productivity of R&D systems. Production functions and production frontiers, in Moed, H.F., Glänzel, W., and Schmoch, U. (eds.) *Handbook of Quantitative Science and Technology Research. The Use of Publication and Patent Statistics in Studies of S&T Systems*. Dordrecht, Kluwer Academic Publishers.

Bonaccorsi A. (2012) La misura dei saperi. *Il Sole 24 Ore*, Domenica 26 febbraio.

Bouyssou D., Marchant T. (2011) Bibliometric ranking of journals based on Impact Factors: An axiomatic approach. *Journal of Informetrics*, vol. 5.

Braun, T., Glänzel, W., Schubert, A., (2006) A Hirsch-type index for journals. *Scientometrics* vol. 60:169-173.

Brody, T., Harnad, S., & Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, vol. 57, no 8, 1060–1072.

Bryman A. (2008) *Social research methods*. Oxford, Oxford University Press.

Burnhill, P.M., Tubby-Hille M.E. (2003) On measuring the relation between social science research activity and research publication. *Research Evaluation*, vol. 4, no. 3, 130-152.

Butler, L. (2003a) Explaining Australia's increased share of ISI publications – the effects of a funding formula based on publication counts. *Research Policy* , vol. 32, 143-155.

Butler, L. (2003b) Modifying publication practices in response to funding formulas. *Research Evaluation*, vol. 12, 39-46.

Butler, L. (2011). The devil is in the detail: Concerns about Vanclay's analysis of Australian journal rankings. *Journal of Informetrics*, vol. 5, 693–694.

Campbell K., Goodacre A., Little G. (2006) Ranking of United Kingdom Law Journals: An analysis of the Research Assessment Exercise 2001 Submissions and Results. *Journal of Law and Society*, vol. 33, 335-363.

Carayol N., Dalle J.M. (2007) Sequential problem choice and the reward system in Open Science. *Structural Change and Economic Dynamics*. Vol. 17, 167-191.

Carmines E.G., Zeller R.A. (1979) *Reliability and validity assessment*. Sage Publications.

Cassella M. (2010) Social peer-review e scienze umane, ovvero "della qualità nella Repubblica della scienza", *JLIS.it*. Vol. 1, n. 1, June 2010, 111–132.

Cassella M. (2011) Nuovi scenari per la valutazione della ricerca tra indicatori bibliometrici citazionali e metriche alternative nel contesto digitale. *Biblioteche oggi, Marzo.*

Cassella M., Bozzarelli O. (2011) Nuovi scenari per la valutazione della ricerca tra indicatori bibliometrici citazionali e metriche alternative. *Biblioteche oggi, Marzo.*

Chiesi, A.M. (2008) La valutazione della produzione sociologica, *Quaderni di Sociologia*, vol. 52, no. 47.

Christenson, J. A., Sigelman, L. (1985). Accrediting knowledge: Journal stature and citation impact in social science. *Social Science Quarterly*, 66(4), 964-975.

Chubin D. (l980) Is citation analysis a legitimate evaluation tool? *Scientometrics*, vol. 2 (1), 91-94.

CIBER Research Limited (2011) The Journal Usage Factor. Exploratory data analysis. *CIBER.*

Clements, E., Powell, W., McIlwaine, K., Okamoto D. (1995) Careers in print: Books, journals and scholarly publications. *American Journal of Sociology*, vol. 101, 433-494.

Cohen J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, vol. 20, 37-46.

Comba V. (2003) La valutazione delle pubblicazioni: dalla letteratura a stampa agli Open Archives, *Bollettino AIB*, vol. 43, no. 1, 65-76.

Consiglio Nazionale delle Ricerche (2009) Criteri per la valutazione della ricerca nel campo delle scienze umane e sociali. Documento finale

Consiglio Universitario Nazionale (2009) Gruppo di lavoro CUN sulla valutazione in area umanistica (aree 10 e 11). Documento finale, 18 dicembre 2009.

Costantini C., Franceschet M. (2011) Un primo confronto tra VTR 2001-2003 e VQR 2004-2008. Mimeo, Università degli Studi di Udine.

Coyne J.G., Summers S.L., Williams B., Wood D.A. (2010) Accounting program research rankings by topical area and methodology. *Issues in Accounting Education*, vol. 25, no. 4, 631-654.

De Robbio A. (2004) Analisi citazionale e indicatori bibliometrici nel modello Open Access, *NFAIS Forum*, October 2004.

De Robbio A. (2009) L'Open Access per la valutazione. In Gargiulo P., Bogliolo P. (a cura di) *Ciber 1999-2009*. Edizioni LediPublishing

Diani, M. (2008) Indicatori bibliometrici e sociologia italiana. *Quaderni di Sociologia*, vol.52, n. 47.

East J.W. (2006) Ranking journals in the Humanities: An Australian case study. *Australian Academic & Research Libraries*, vol. 37 no. 1, March 2006.

European Commission (2010) *Assessing Europe's University Based Research*. Expert group on the Assessment of university-based research (AUBR).

Federal Ministry of Education and Research (BMBF) (2007) *Freedom for research in humanities.* Berlin.

Figà Talamanca A. (2000) L'Impact Factor nella valutazione della ricerca e nello sviluppo dell'editoria Scientifica. Atti del IV Seminario del Sistema Informativo Nazionale per la Matematica SINM 2000: *Un modello di sistema informativo nazionale per aree disciplinari*

Fleiss J.L., Levin B., Paik M.C. (2003) *Statistical methods for rates and proportions*. New York, Wiley Series in Probability and Statistics.

Franceschet, M., Costantini, A. (2009) The first Italian Research Assessment Exercise: a bibliometric perspective. *Journal of Informetrics* 5, 275–291.

Genoni, P. and Haddow, G. (2009) ERA and the Ranking of Australian Humanities Journals. *Australian Humanities Review* 46:7-26.

Gilbert G.N. (1977) Referencing as Persuasion. *Social Studies of Science*, Vol. 7, no. 1, February, 113-122.

Gimenez-Toledo E., Román- Román A., Alcain_Partearroyo D. (2007) From experimentation to coordination in the evaluation of Spanish scientific journals in the humanities and social sciences. Research Evaluation, vol. 16, no. 2, 137-148.

Gonzalez-Pereira, B., Guerrero-Bote, V.P. and Moya-Anegon, F. (2010) A new approach to the metric of journals' scientific prestige: The SJR indicator. *Journal of Informetrics* 4:379-391.

Gwet K.L. (2010) *Handbook of inter-rater reliability*. Gaithesburg, Adanced Analytics, 2nd edition.

Haddow, G. and Genoni, P. (2010) Citation analysis and peer ranking of Australian social science journals. *Scientometrics* 85:471-487.

Hardy C., Bryman A. (2004) *Handbook of data analysis*. Thousand Oaks, Sage Publications.

Hellqvist B. (2010) Referencing in the Humanities and its implications for citation analysis. *Journal of the American Society for Information Science and Technology*, vol. 61 (2), 310–318.

Hicks, D. (2004) The four literatures of social science. In: Moed, H. (Ed.) *Handbook of Quantitative Science and Technology Studies*. Dordrecht, Kluwer Academic Press, 473–496.

Hicks, D. (2012) Performance-based university research funding systems. *Research Policy* , vol. 41, 251– 261.

Hicks, D., Wang, J. (2009) Towards a bibliometric database for the Social Sciences and Humanities. A European Scoping Project. *Final Report* on Project for the European Science Foundation.

Higher Education Funding Council for England (HEFCE) (1997) The impact of the 1992 Research Assessment Exercise on Higher Education Institutions in England, No. M6/97. Higher Education Funding Council for England, Bristol.

Hofmeister R. (2011) Measuring the value of research: A Generational Accounting Approach. University of Konstanz, *Department of Economics Working Paper* 2011-07.

Huang M.H., Lin C.S. (2008) Characteristics of research output in Social Sciences and Humanities: From a research evaluation perspective. *Journal of the American Society for Information Science and Technology*, vol. 59, no. 11, 1819–1828.

Huang M.H., Lin C.S. (2010) A citation analysis of Western journals cited in Taiwan's Library and Information Science and History Research Journals: From a research evaluation perspective. *The Journal of Academic Librarianship*, vol. 37, no. 1, 34–45.

Hugher A., Kitson M., Probert J. (2011) *Hidden connections. Knowledge exchange between the arts and humanities and the private, public and third sectors*. London, Arts and Humanities Research Council (AHRC)

Hurt C.D. (1987) Conceptual citation differences in science, technology, and social sciences literature. *Information Processing and Management*, vol. 23, no. 1, 1-6.

Janis I.(1982) *Groupthink*. New York, Houghton Mifflin (1st edition, 1977).

Jarwal, S.D., Brion, A.M. and King, M.L. (2009) Measuring research quality using the journal impact factor, citations and 'Ranked Journals': blunt instruments or inspired metrics? *Journal of Higher Education Policy and Management* , vol. 31(4):289 – 300.

Kalaitzidakis, P., T. P. Mamuneas, and T. Stengos (2003) Rankings of Academic Journals and Institutions in Economics. *Journal of the European Economic Association*, 1(6), 1346–1366.

Kirk J., Miller M.L. (1986) *Reliability and validity in qualitative research*. Thousand Oaks, Sage Publications.

Lamp J.W. (2009) At the sharp end : journal ranking and the dreams of academics. *Online information review*, vol. 33, no. 4, 827-830.

Lariviére V, Archambault E., Gingras Y., Vignola-Gagne E. (2006) The place of serials in referencing practices: Comparing Natural Sciences and Engineering with Social Sciences and Humanities. *Journal of the American Society for Information Science and Technology*, vol. 57 (8), 997-1004.

Lewis-Beck M.S., Bryman A., Liao T.F. (2004) *The Sage Encyclopedia of social science research methods.* Thousand Oaks, Sage Publications.

Linmans A. J. M. (2010) Why with bibliometrics the Humanities does not need to be the weakest link Indicators for research evaluation based on citations, library holdings, and productivity measures. *Scientometrics*, vol. 83:337–354.

Marcuzzo M.C., Zacchia G. (2007) L'EconLit e gli strumenti per la valutazione della ricerca economica in Italia. *Rivista Italiana degli Economisti*, vol. 12, no. 2, Agosto.

Martin B. et al. (2010) *Towards a bibliometric database for the Social Sciences and Humanities – A European Scoping Project*. A report produced for DFG, ESRC, AHRC, NWO, ANR and ESF, 8 March 2010

Mc Roberts M.H., Mc Roberts B.R. (1989) Problems of citation analysis. A critical review. *Journal of the American Society for Information Science , vol.* 40(5), 342-349.

Merton, R. K. (1968) The Matthew effect in science. *Science*, 159(3810), 56-63.

Metris (2010) Country Report. Social Sciences and Humanities in Italy. http://www.metrisnet.eu/metris//fileUpload/Italy.pdf

Moed H. F. (2000) Bibliometric indicators reflect publication and management strategies. *Scientometrics* Vol.47, no.2, 232-346.

Moed H.F. (2002) Towards research performance in the humanities - bibliometrics in qualitative analysis of Flemish law literature - Statistical Data Included. *Library Trends*, Winter Issue.

Moed, H.F. (2005). *Citation Analysis in Research Evaluation*, Dordrecht, Springer.

Moed H.F. (2008) Research Assessment in Social Sciences and Humanities - Evaluation in the Human Sciences. Presentazione al Convegno Università di Bologna, 12-13 Dicembre.

Moed H.F., Daraio C. (2008) La valutazione dei ricercatori e delle istituzioni scientifiche in Europa. Convegno ANPRI *Un futuro per la ricerca pubblica italiana: Autonomia, valutazione, risorse*, 24 novembre.

Nederhof A.J. (2006) Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A review**.** *Scientometrics*, vol. 66, no. 1, 81–100.

Nederhof A.J., Noyons E.C.M. (1992) International comparison of departments' research performance in the humanities. *Journal of the American Society for Information Science* , 43(3):249-256.

Nederhof A.J., Zwaan R.A. (1991) Quality judgments of journals as indicators of research performance in the Humanities and the Social and Behavioral Sciences. *Journal of the American Society for Information Science*, vol. 42 (5), 332-340.

Norris, M., & Oppenheim, C. (2003). Citation counts and the Research Assessment Exercise. V: Archaeology and the 2001 RAE. *Journal of Documentation*, vol. 59(6), 709–730.

Norris, M. and Oppenheim, C. (2010) Peer review and the h-index: Two studies. *Journal of Informetrics* vol. 4, 221-232.

Noruzi A. (2006) The Web Impact Factor: A critical review. *The Electronic Library*, vol. 24.

OECD (2010), *Performance-based funding of Public Research in Tertiary Education institutions*. Paris, OECD.

Oppenheim, C. (1995). The correlation between citation counts and the 1992 Research Assessment Exercise ratings for British library and information science university departments. *Journal of Documentation*, vol. 51(1), 18–27.

Oppenheim, C. (1997). The correlation between citation counts and the 1992 research assessment exercise ratings for British research in genetics, anatomy and archaeology. *Journal of Documentation*, vol. 53(5), 477–487.

Oppenheim, C., & Summers, M. A. C. (2008). Citation counts and the Research Assessment Exercise, part IV: Unit of assessment 67 (music). *Information Research*, 13(2), 342.

Piazzini T. (2010) Gli indicatori bibliometrici: riflessioni sparse per un uso attento e consapevole - *JLIS.it.*, vol. 1, no. 1, Giugno, 63–86.

Pinski G., Narin F. (1976) Citation influence for journal aggregates of scientific publications: Theory with application to the literature of physics. *Information Processing & Management*, vol. 12, 297-312.

Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P., Stirling, A. (2011) How journal rankings can suppress interdisciplinarity. The case of innovation studies in business and management. May.

Reale E. (2010) La valutazione della ricerca nelle discipline umanistiche e sociali: riflessioni introduttive. Università di Udine, Maggio.

Ritzberger, K. (2008): A Ranking of Journals in Economics and related fields. *German Economic Review*, vol. 9, 402–430.

Rodríguez-Navarro, A. (2009) Sound research, unimportant discoveries: research, universities, and formal evaluation of research in Spain. *Journal of the American Society for Information Science and Technology*, vol. 60, 1845–1858.

Rousseau R. (2002) Journal evaluation: technical and practical issues. *Library Trends*, Winter Issue.

Royal Netherlands Academy of Arts and Sciences (2011) *Quality indicators for research in the humanities.* Interim report by the Committee on Quality Indicators in the Humanities, May.

Seglen P.O. (1997) Why the impact factor of journals should not be used for evaluating research. *British Medical Journal*, vol. 314, 15 February.

Silverman D. (2000) *Doing qualitative research*. London, Sage Publications.

So C.Y.K. (1998) Citation ranking versus expert judgment in evaluating communication scholars. Effects of research specialty size and individual prominence. *Scientometrics*, vol. 41, no. 3, 325 333.

Solow R.M., Franklin P., Jones C.C., Oakley F., D'Arms J. (2002) *Making the Humanities count: The importance of data.* Cambridge, Mass., American Academy of Arts & Sciences.

Sorensen J.R. (1994) Scholarly productivity in criminal justice: Institutional affiliation of authors in the top ten criminal justice journals. *Journal of Criminal Justice*, vol. 22, no. 6, 535-547.

Spier R. (2002) The history of the peer-review process. *Trends in Biotechnology*, vol.20, no.8, August.

Starbuck, W.H. (2005) How much better are the most-prestigious journals? The statistics of academic publication. *Organization Science*, vol. 16, 180–200.

Svantesson D.J. (2009) International Ranking of Law Journals: Can it be done and at what cost? *Legal Studies* , vol. 29, 678-691.

Tarantino E. (2005) Troppo o troppo poco? Web of science, Scopus, Google scholar: tre database a confronto (un caso di studio)**.** Paper presented to the *ICOLC Autumn meeting*, Poznan, 29 settembre-1ottobre.

Torres-Salinas D., Moed, H. F. (2009) Library Catalog. Analysis as a tool in studies of social sciences and humanities: An exploratory study of published book titles in Economics. *Journal of Informetrics*, vol 3 (1): 9-26.

Traub R.E. (1994) *Reliability for the social sciences: Theory and applications*. Beverly Hills, Sage Publications.

Tsay M. (1998) The relationship between journal use in a medical library and citation use. *Bulletin of the Medical Library Association*, vol. 86, no. 1, January.

Valdecasas A.G., Castroviejo S., Marcus L.F. (2000) Reliance on the citation index undermines the study of biodiversity. *Nature*, vol 403, 17 February 2000.

van Gestel R., Vranken J. (2011) Assessing legal research: Sense and nonsense of Peer Review versus Bibliometrics and the need for a European approach, *German Law Journal*, vol. 12, 901-929.

Vanclay, J. K. (2011). An evaluation of the Australian Research Council's journal ranking. *Journal of Informetrics*, vol. 5, 265–274.

Vanclay, J.K. (2012) What was wrong with Australia's journal ranking? *Journal of Informetrics*, vol. 6, 53– 54

Viale, R. e Cerroni, A. (a cura di) (2003) *Valutare la scienza*, Soveria Mannelli, Rubettino.

Vitiello G. (2005) Il mercato delle riviste in Scienze umane e sociali in Italia Analisi quantitativa e sua evoluzione in ambito elettronico. *Biblioteche oggi*, Gennaio-Febbraio 2005.

Wieviorka M. (2011) Evaluation, research and demonstration in the social sciences. *Social Science Information*, vol. 50, 308-316.