Specifiche preliminari per una base dati bibliometrica italiana nelle aree umanistiche e sociali

Indice

Allegati:

Allegato 1 – Data model & architettura della banca dati

Allegato 2 – Esempi di indicatori

1. DEFINIZIONI

Database
Base di dati delle riviste italiane nelle aree umanistiche e sociali, promossa in via sperimentale dal Gruppo di lavoro Database e nuovi indicatori dell'ANVUR.

Contributor
Soggetto (persona fisica) che compare nella lista di coloro che hanno una responsabilità su un'opera rilevante ai fini della valutazione. Include l'Autore, il Curatore, il Traduttore e altri ruoli di potenziale interesse.

Soggetti strutturati
Ricercatori, professori associati e professori ordinari affiliati a università italiane; ricercatori, dirigenti e primi dirigenti affiliati a enti di ricerca.
Nel caso di soggetti strutturati al nome del contributor è legata una affiliazione. Nel caso di affiliazione ad università al nome sono associati anche un settore scientifico disciplinare e un settore concorsuale (con le date di inizio e fine relative al periodo di strutturazione).

Soggetti non strutturati
Altri soggetti (persone fisiche) diverse dai soggetti strutturati che risultino contributor.

Affiliazione
Relazione che lega un contributor a un'istituzione (dipartimento, università, ente di ricerca o simili). Si distinguono due casi: (a) affiliazione singola; (b) affiliazione multipla. Il primo caso si ha nella grande maggioranza delle situazioni, con l'affiliazione di un autore a un solo dipartimento, e attraverso di questo, ad una sola università. Il secondo caso si riscontra quando, in aggiunta all'affiliazione al dipartimento, è in essere un'affiliazione (associazione) a uno o più enti di ricerca o altri organismi di ricerca, formalizzata dai soggetti interessati. Tale relazione può dunque essere uno a uno (una sola affiliazione per autore) o uno a molti (più affiliazioni per autore). I soggetti strutturati non affiliati a un dipartimento sono classificati in una categoria residuale.

Subject category
Area scientifica omogenea per oggetto di indagine e metodo, definita all'interno di un tesauro eventualmente di struttura gerarchica, per tener conto di soggettazioni più o meno fini.
Può essere riferita alle singole pubblicazioni o alle riviste.

Dipartimento
Unità organizzativa in cui è suddivisa una università, ai sensi della legislazione vigente. Ogni dipartimento fa capo ad una sola università.

Università
Università statali e non statali, incluse le telematiche.

Ente di ricerca
Ente pubblico di ricerca. In prima applicazione il database includerà analisi relative ai 12 enti di ricerca vigilati dal MIUR, laddove le aree umanistiche e sociali siano rilevanti. In successivi sviluppi potrà essere esteso.
Per ragioni di semplicità, nel seguito del documento si userà il termine "università" a intendere sia università che enti pubblici di ricerca.

Rivista
Pubblicazione seriale contraddistinta da uno stesso titolo e in genere da un ISSN, che prevede uscite regolari, in versione cartacea e/o elettronica.

Rivista scientifica
Rivista inclusa nella lista delle riviste scientifiche pubblicata dall'ANVUR ai sensi del DM 76/2012 e dei provvedimenti relativi.

Rivista di classe A
Rivista inclusa nella lista delle riviste scientifiche di classe A pubblicata dall'ANVUR ai sensi del DM 76/2012 e dei provvedimenti relativi.

Content provider
Ai fini del database si considerano content provider gli editori, i distributori, le università, dipartimenti o enti che editano riviste scientifiche, le singole riviste scientifiche, sia che si tratti di testate ad accesso aperto che a distribuzione tradizionale.

Corpus
Insieme di riviste incluse nel database.

Anno
Anno di pubblicazione degli articoli su rivista, come evidenziato nell'annata riportata in copertina. L'anno di inizio del database è fissato, in via di prima applicazione, al 2002.

Per ogni definizione si userà nel seguito una notazione algebrica, in modo informale. Per semplicità si omette la lista delle notazioni.

2. I CONTENUTI

Il Database è uno strumento di valutazione e di ricerca, messo a disposizione delle istituzioni e delle comunità scientifiche delle aree umanistiche e sociali.
Esso include riviste italiane di area umanistica e sociale, di cui vengono pubblicati i metadati riferiti ai singoli articoli, e i cui articoli *full text* vengono conferiti allo scopo di estrarre informazioni citazionali e di uso a fini di valutazione.
Le riviste di cui ai punti successivi vengono invitate a conferire il *full text* delle ultime dieci annate di pubblicazione o, per riviste di più recente formazione, dal primo anno di pubblicazione (almeno tre), allo scopo di consentire la estrazione automatica di informazioni citazionali, ai fini della costruzione di indicatori per la valutazione.
Allo stesso tempo, le riviste conferiranno i metadati, con il risultato di costituire un repertorio bibliografico di consultazione a fini di ricerca e di rafforzamento della visibilità internazionale della ricerca umanistica e sociale italiana.

Il perimetro di applicazione si riferisce a tutte le aree non bibliometriche, come definite dal DM 76/2012, dalla Delibera n. 50 dell'ANVUR e dai successivi documenti relativi alla classificazione delle riviste, inclusi gli aggiornamenti successivi alle tornate di Abilitazione e le revisioni dei giudizi.
I criteri di inclusione delle riviste italiane sono definiti come segue:

- Nella fase sperimentale e di startup vengono invitate tutte le riviste italiane classificate di fascia A ai fini della Abilitazione Scientifica Nazionale
- Qualora vi siano riviste classificate in fascia A ai fini della VQR ma non della Abilitazione si ritiene esteso il criterio
- Nel caso dell'area 13 (Economia), nella quale nessuna rivista italiana compare in fascia A ai fini della Abilitazione, verranno sviluppati criteri specifici, in coerenza con l'attività del GEV di area e sentito il Gruppo di lavoro Riviste scientifiche
- La fase sperimentale si concluderà con la produzione di indicatori, secondo le linee illustrate nel presente documento e negli Allegati
- In parallelo si provvederà allo sviluppo e alla pubblicazione di Linee Guida per la inclusione delle riviste scientifiche italiane anche non appartenenti alla fascia A
- Sulla base delle Linee Guida si procederà ad una fase di candidatura di riviste e di selezione ai fini della messa a regime del Database
- Le Linee Guida preciseranno altresì i casi di sospensione o di uscita dal Database
- Nel corso del processo si terrà conto degli sviluppi della iniziativa della Anagrafe della ricerca (ANPRePS) e degli orientamenti del MIUR circa la definizione della natura scientifica delle riviste, su proposta del CUN.

Il Database ha caratteri di neutralità e intende offrire dati e indicatori per successivi utilizzi e rielaborazioni da parte dei soggetti interessati, senza predeterminare scelte valutative o di politica della ricerca che competono ad altre istanze, quali ad esempio gli algoritmi di valutazione, i pesi da assegnare a vari indicatori etc.. Nella fase di progettazione del Database, illustrata nel presente documento, si cercherà il massimo coinvolgimento dei soggetti (riviste, comunità scientifiche, associazioni e consulte) e dei livelli istituzionali.

Nella fase iniziale le citazioni indicizzate dal Database saranno di necessità circoscritte alle fonti incluse nel corpus. Una volta definita e implementata l'ANPRePS, sarà possibile estendere la lista delle riviste di cui vengono indicizzate le citazioni in uscita utilizzando una fonte ufficiale.

Non si procederà in fase iniziale alla indicizzazione di monografie all'interno del Database.

3. IL DATABASE

3.1. Struttura generale e data model

Vedi documento allegato.

3.2. Come sottoporre metadati e dati

La piattaforma assicurerà l'interoperabilità da e verso i content provider, i metamotori e altri archivi in generale, attraverso l'adozione di metodi consolidati basati su protocolli aperti e standard.

In particolare, per l'acquisizione, possiamo prevedere:

- *Harvesting*. Agli ip abilitati sarà consentita la navigazione di link che condurranno – gerarchicamente – ai metadati di riviste, fascicoli, articoli, giungendo infine a una pagina che permetterà l'accesso al *full text*. L'acquisizione dei metadati potrà avvenire via OpenUrl in formato COinS leggendo medatati specifici secondo il livello gerarchico di riferimento, oppure in OAI-PMH. Tra i metadati da acquisire dovrà esserci la URL per il download del full text.

- *Esportazione*. Attraverso un'interfaccia REST, sarà possibile acquisire full text e metadati in formato Marc21, MODS oppure ONIX for serials. Contemporaneamente saranno formalizzate le regole per la costruzione dei dati (quali tag, tra i tanti possibili, e quali obbligatori, opzionali, ecc.).
- *Form per input manuale*. Per i content provider sarà disponibile una form online con il core dei metadati obbligatori e opzionali per rivista, fascicolo, articolo, al quale sarà possibile associare un file.
- *Formato dei file in ingresso*. Nella fase iniziale i formati accettati sono: PDF testo, HTML, XML.
- *Strumenti di* reference management *per l'acquisizione delle citazioni*, laddove il content provider sia in grado di fornirle.

Questi metodi saranno disponibili per i diversi content provider (editori, distributori, università ed enti, singole riviste).

3.3. Quanto costa all'editore?

Il costo per l'editore per aderire all'iniziativa dipende da numerose variabili i cui termini non sono ancora definiti a sufficienza per una valutazione puntuale.
Ciascun editore dovrà fornire alla banca dati i metadati e i full text degli articoli (intero fascicolo) . La variabile più significativa in questa fase è relativa all'organizzazione interna dello stesso editore e alle modalità produttive attuali e passate, per l'esigenza di fornire anche dati per annate pregresse.
Tra gli elementi da considerare è possibile elencare:

- L'esistenza di metadati sufficientemente accurati a livello di articolo all'interno dell'azienda;
- L'esistenza dei file pdf (o in altro formato accettato) dei full text, il che può non essere vero per alcuni editori, in particolare per annate pregresse, il cui ciclo produttivo – generalmente esternalizzato – può non aver previsto la consegna dei file prodotti esternamente;
- La necessità, in caso di assenza dei file digitali, di scannerizzare i file e il livello di accuratezza;
- La necessità di interfacciare il sistema produttivo interno aziendale con il servizio di input della banca dati.

Pertanto, i costi per l'editore saranno relativi alla compilazione dei metadati, al recupero dei file o alla sua scannerizzazione, e all'interfacciamento con il sistema.
Una volta definiti i dettagli tecnici, e la loro incidenza su un campione di casi, si può supporre di avere una stima di costo chiedendo un preventivo per le diverse fasi a una o più aziende specializzate, sotto l'ipotesi che il costo sia in linea di principio equivalente a quello di una lavorazione interna.

3.4. Qualità dei dati e interoperabilità con altre fonti

La banca dati deve per definizione dialogare con altri sistemi, giacché il suo uso per la valutazione lo impone. Dal punto di vista tecnico, il problema può essere posto nei seguenti termini:

a) La banca dati contiene informazioni relative a una serie di *entità*: pubblicazioni (articoli, libri, capitoli di libri...); sedi editoriali (riviste, editori…); persone fisiche (autori / ricercatori); strutture di ricerca (cui gli autori sono affiliati).
b) Una qualsiasi banca dati A dialoga con una banca dati B quando è in grado di scambiare informazioni su una determinata entità. Perché questo avvenga, l'entità di interesse deve essere identificata univocamente nelle due banche dati tramite uno stesso identificatore. È fondamentale che tali identificatori siano standard riconosciuti.

c) Per l'efficienza dello scambio di informazioni, le stesse devono essere analogamente espresse secondo schemi standard, il che implica che (i) il data model deve essere semanticamente compatibile con gli standard, nella misura richiesta dagli scambi di informazioni prevedibili; (ii) il sistema deve supportare l'export e l'import di dati secondo i protocolli standard più comuni.

L'analisi di dettaglio degli standard applicabili dovrà essere fatta in sede di definizione delle specifiche operative, ma si possono anticipare alcune considerazioni generali:

- Per gli aspetti istituzionali (autori strutturati, affiliazioni ad università ed enti, ricostruzione nel tempo delle affiliazioni nei passaggi di carriera e di sede) andrà assicurata l'interoperabilità con i database del MIUR, attraverso la importazione delle anagrafiche e l'aggiornamento dei dati relativi
- Per le pubblicazioni sono rilevanti sia gli identificatori standard dei prodotti (ISBN, ISSN), e ancor più quelli delle opere (ISTC, ISSN-L, DOI…), che tuttavia hanno ancora un uso limitato. Il data model deve tracciare le relazioni tra prodotti riferibili alla stessa opera.
- Per gli autori sembra imprescindibile l'uso di ORCID – the Open Researcher and Contributor ID –, che nella produzione scientifica si sta affermando come standard di riferimento ed è essenziale per l'interoperabilità con altre banche dati bibliometriche che lo stanno adottando (è ad esempio già adottato in Scopus e CrossRef). Va valutata l'utilità anche di ISNI (che incorpora il VIAF-id, a sua volta utilizzato dalle Biblioteche nazionali e integrato in WoS) che è ad oggi prevalente per gli autori di monografie. Vi sono progetti di integrazione tra ORCID e ISNI / VIAF che dovrebbero chiarire a breve il quadro.
- Per le strutture di ricerca si dovrà tener conto del modello di descrizione presente nei risultati di CERIF (Common European Research Information Format)..
- Per gli editori, è necessario costruire un autority file, che tenga conto del rapporto con i prefissi ISBN, e con altri standard in uso.

3.5. Valorizzazione delle citazioni di articoli non appartenenti al corpus

L'estrazione delle citazioni dagli articoli delle riviste incluse nella banca dati produce di fatto una *Banca dati delle citazioni*, i cui elementi sono le pubblicazioni citate. Queste possono essere le stesse incluse nella banca dati, o diverse. La banca dati delle pubblicazioni citate esterne al corpus potrà essere utilizzata in due modi:

- Per individuare nuove riviste da includere nella banca dati, identificate come rilevanti per il numero di citazioni ricevute;
- Per avere dati citazionali riferiti a pubblicazioni non incluse nella banca dati ma di interesse in procedure di valutazione.

Per valorizzare questo secondo aspetto, le citazioni estratte dal corpus possono essere confrontate con i record presenti nell'intera ANPRePS, che comprende le pubblicazioni potenzialmente soggette a procedure di valutazione.

Ciò ha il vantaggio di ridurre la tensione sui criteri di inclusione delle riviste. Un articolo presente in ANPRePS citato all'interno di un pacchetto significativo di riviste avrà un suo profilo citazionale anche laddove la rivista su cui è edito non è ricompresa tra quelle del corpus. In linea di principio, ciò è vero anche per un articolo pubblicato su una rivista valutata come "non scientifica", in quanto l'unità base su cui si costruiscono i dati è l'articolo, non la sede di pubblicazione.

Inoltre, si possono così avere dati sulle monografie. Per queste, infatti, l'inclusione nelle banche dati è difficile per l'estrazione dei dati citazionali, molto meno per la possibilità di rilevare se la stessa monografia è citata da altri.

3.6. Accordi di collaborazione con fonti esterne

Occorre prevedere per il futuro accordi di collaborazione finalizzati all'arricchimento e alla interoperabilità dello strumento con qualificate banche dati internazionali e autorevoli repertori disciplinari.
Il controllo di qualità dei dati forniti diventa un elemento cruciale di questa fase di lavoro. La interoperabilità con altre banche dati sarà svolta in modo da non inserire distorsioni nei profili citazionali.
Per evitare che la banca dati resti un sistema chiuso è possibile rilevare citazioni degli articoli inclusi nella banca dati ricevute da pubblicazioni non presenti in essa.
Per tutte è necessario un approfondimento operativo, per valutare la fattibilità dell'operazione, che tuttavia può essere un elemento molto qualificante dell'intero servizio.

4. ESTRAZIONE DELLE CITAZIONI

Un'operazione su larga scala di estrazione dei riferimenti bibliografici dalle pubblicazioni è di fatto inattuabile mediante una pura e semplice attività manuale: di qui la necessità di disporre di strumenti automatici. Si tratta di un settore in qualche misura ancora sperimentale, considerate le oggettive difficoltà che riveste il processo di estrazione automatica delle citazioni, come tutte le forme di *text mining* avanzato. Le problematiche più rilevanti da considerare riguardano le tipologie di dati in ingresso (i testi possono essere disponibili in formati diversi ed essere stati redatti secondo stili editoriali diversi) e l'interoperabilità con altre fonti di dati citazionali.

4.1. Formati dei file in ingresso

Tra gli strumenti attualmente disponibili – considerati in generale, senza distinguere tra prodotti commerciali e progetti di ricerca accademica – si possono in primo luogo riscontrare alcune differenze in relazione al formato trattato, alcuni trattano solo il .pdf; altri solo il .doc; altri solo l'html o il .txt, eventualmente ottenuti mediante scansione OCR; altri ancora hanno invece la capacità di analizzare documenti in maniera indipendente dal formato.
Nel nostro caso, si può ipotizzare che l'esigenza sia limitata alla gestione in input di soli formati editoriali. Nella fase iniziale ipotizzata, che suppone di limitare la gestione ai soli articoli di rivista, vi sarà una larghissima prevalenza di .pdf, anche se dovrà essere approfondita l'utilità di analizzare anche l'html (per i casi in cui la lista dei riferimenti bibliografici è disponibile separatamente in questo formato[1]) o formati intermedi di lavorazione, se adottati dagli editori internamente e laddove il loro uso migliori l'efficienza delle estrazioni[2].

[1] Ad esempio, il software per la produzione delle riviste elettroniche OJS - Open Journal System contiene una funzione (non sempre utilizzata) per esporre i riferimenti bibliografici separatamente dal testo.
[2] Se l'editore aderisce al sistema "Cited by" di Crossref sottopone le citazioni in alcuni formati accettati dal sistema, che potrebbero essere accettati anche dalla banca dati in costruzione.

4.2. Stili citazionali

L'elemento più critico è tuttavia relativo alle modalità di citazione adottate in particolare nell'ambito delle scienze umane e sociali. Occorre fare innanzi tutto una distinzione tra citazioni riportate in bibliografia alla fine dell'articolo e citazioni in nota. Per ciascuna di tali categorie esistono diversi stili di formattazione (i.e. corsivo, *font-size*, ecc.), con tutta una serie di opzioni editoriali la cui variabilità incide significativamente sulle prestazioni dei sistemi di estrazione automatica.

Una procedura essenziale, e quindi comune a tutti i *software* che eseguono questo compito, è quella della segmentazione del testo, che permette di isolare dal resto del documento le porzioni di testo che contengono riferimenti bibliografici (le *reference list* e le note a piè di pagina), e successivamente di analizzarle.

I principali software esistenti sono efficienti per le analisi delle bibliografie, molto meno per le note a piè pagina, tuttora molto diffuse, se non prevalenti, nelle scienze umane e sociali. La struttura discorsiva delle note, con le citazioni inserite nel testo secondo schemi non standardizzati e senza apparente soluzione di continuità rende infatti il lavoro di identificazione delle stesse più complesso.

Sarebbe auspicabile promuovere un'uniformazione di tali stili. Si tratta quindi per il momento di acquisire un numero sufficientemente elevato di stili citazionali, e delle relative varianti, così da massimizzare l'efficienza del software utilizzato.

I diversi estrattori citazionali fanno in genere uso di algoritmi di *machine learning* che si basano sull'esistenza di un *training set*, ovvero di un insieme di documenti pre-annotati manualmente che vengono utilizzati per addestrare i relativi modelli. Alcuni *tool* completano queste metodologie di tipo quantitativo implementandone una di tipo qualitativo: si tratta di un approccio euristico, che prevede la definizione di un *set* di regole appositamente sviluppate in funzione degli stili citazionali propri di un certo ambito disciplinare.

4.3. Corrispondenza con altre fonti

La disponibilità di fonti esterne di dati bibliografici strutturati consente di migliorare la risoluzione complessiva delle citazioni estratte automaticamente. In questa fase, particolare rilievo assume la disponibilità di basi dati relative ai singoli elementi citazionali (i.e. autori, editori, riviste, luoghi di pubblicazione) che possono supportare il processo di riconoscimento e validazione degli stessi elementi.

La prima banca dati con cui interfacciarsi è anche in questo caso l'ANPRePS, nell'ipotesi – che allo stadio attuale è solo un auspicio – che questa contenga dati bibliografici sufficientemente strutturati. Tuttavia, la possibilità di interfacciarsi con altre fonti (ORCID, per fare un solo esempio), può ulteriormente migliorare l'efficacia di un sistema automatico di estrazione citazionale. Particolare attenzione va posta agli Open Data, e alla disponibilità di un crescente numero (a partire da Open Library, ma non solo) di basi dati bibliografiche aperte.

Il sistema, dopo aver individuato le citazioni, dovrà effettuare un *matching* con tali basi dati bibliografiche, per verificare la corrispondenza tra il *pattern* individuato e la *reference* effettivamente depositata. Ogni potenziale citazione è formata dai suoi vari componenti (autore, titolo, rivista, anno, …): più grande è il valore x corrispondente al numero di componenti che superano il *matching* (pesato per dare più rilievo a componenti chiave quali titolo e autore), maggiore è la probabilità che il *pattern* sia stato riconosciuto correttamente.

4.4. Interventi manuali di correzione

Il contesto in cui operiamo impone l'esigenza di fornire informazioni con un livello di accuratezza quanto più possibile vicino al 100% (pur tenendo conto che anche i principali *database* come Web of Science e

Scopus, già utilizzati a scopi valutativi, non sono scevri da errori: duplicazioni, omonimie, ecc.). Considerata la complessità sopra descritta, è ragionevole ipotizzare, a valle dell'estrazione automatica delle citazioni, uno stadio successivo rappresentato dalla validazione manuale delle citazioni stesse.
Le due fasi devono essere progettate come integrate. In primo luogo, nella fase automatica di estrazione, la procedura più efficace prevede di fissare una soglia di corrispondenza α (il valore di α dovrà essere sufficientemente elevato). Le citazioni che trovano corrispondenza pressoché completa con una fonte esterna certificata ($x > \alpha$) potranno essere acquisite direttamente. Tuttavia, il mancato riconoscimento può essere dovuto a diversi fattori, tra cui il fatto che una certa opera non è presente nel *database* usato per il *matching*. Viene pertanto fissata un'altra soglia β, articolata su due valori, β_1 e β_2, che misura il livello di riconoscimento y dei singoli componenti di una citazione (autore, titolo, rivista, anno, …). Con $y > \beta_2$, la *reference* può essere acquisita anche in mancanza di *matching* esterno. Con $y < \beta_1$, la *reference* può essere scartata. La rimanente frazione, collocata nella fascia compresa tra i due valori-soglia ($\beta_1 < y < \beta_2$), può essere inviata a un operatore per la validazione manuale.
La validazione manuale richiede un controllo redazionale interno a garanzia della qualità, ma potrebbe in parte basarsi anche sulla collaborazione volontaria degli autori o degli stessi editori (*crowdsourcing*), ai quali richiedere di verificare l'esattezza delle citazioni in uscita (= citazioni effettuate) o – meglio ancora perché più incentivante – in entrata (= citazioni ricevute).

5. LIVELLI DI AGGREGAZIONE E SUBJECT CATEGORIES

5.1. Livelli di aggregazione

La banca dati deve permettere almeno i seguenti livelli di aggregazione:

a) Autore
b) Dipartimento
c) Università o ente di ricerca.

La base di dati dovrà garantire interoperabilità con archivi di natura amministrativa.
Le regole di aggregazione sono governate dai principi di affiliazione:

- Nel caso di affiliazione singola, un autore è univocamente affiliato a un dipartimento e quindi ad una università
- Nel caso di affiliazione multipla, un autore è affiliato ad un dipartimento e a uno o più enti di ricerca (associazione ad enti di ricerca)

Nel caso di affiliazione singola, l'intera collezione di soggetti strutturati costituisce una partizione, nel senso che ogni soggetto è affiliato ad uno e un solo dipartimento, e ogni dipartimento afferisce ad una e una sola università. In questo caso gli indicatori si ottengono per aggregazioni successive, per somma senza duplicazioni.
Nel caso di affiliazione multipla, viene assegnata una unità dell'indicatore a ciascuna delle affiliazioni inserite dall'autore. In questo caso la aggregazione avviene per somma e il conteggio finale conterrà delle duplicazioni (es. la stessa pubblicazione potrà comparire nel Dipartimento X, quindi nell'Università K, ma anche negli enti di ricerca A e B).

5.2. Subject categories

Le pubblicazioni incluse nel database dovranno essere classificate secondo categorie omogenee di tipo scientifico e bibliografico.

Il tema della soggettazione è di primaria importanza e allo stesso tempo implica un dispendioso lavoro redazionale. Anche in questo caso si può cercare di acquisire soggettazioni di terzi pre-esistenti o di invitare autori o editori a forme di collaborazione. Non si può immaginare, tuttavia, un procedimento che prescinda da un significativo lavoro redazionale del gestore della banca dati.

Relativamente ai livelli di soggettazione, è auspicabile mantenere una classificazione, ad uso prevalentemente concorsuale e amministrativo, basata sulla normativa in essere ai fini della Abilitazione scientifica nazionale:

i) Area CUN (per la validità scientifica)
ii) Settore concorsuale (per la classe A).

Nell'esemplificazione degli indicatori discussa sotto si farà esclusivamente riferimento ai settori concorsuali in quanto unità relativamente omogenee dal punto di vista scientifico. Sarà in ogni caso possibile in futuro sviluppare indicatori per le intere aree CUN, se opportuno, o – all'estremo opposto – per ambiti di ricerca più di dettaglio, per evitare che, per ciascun settore concorsuale siano privilegiate le pubblicazioni che affrontano i temi di ricerca più "popolari"

Circa la classificazione scientifica e bibliografica saranno da esplorare le seguenti soluzioni, anche tra loro complementari:

- adozione delle subject categories presenti nei database generalisti (ISI e Scopus), e/o in database specialistici delle aree umanistiche e sociali (es. EBSCO, EconLit, EUROVOC etc.) o anche in ERIH;
- classificazione secondo le categorie scientifiche adottate in istituzioni internazionali (es. ESF, ERC);
- ulteriori classificazioni a grana più fine, anche tenendo conto delle specificità nazionali e disciplinari, ad esempio attraverso analisi di parole chiave presenti negli articoli o negli abstract.

Una ipotesi preliminare di soggettazione, da sottoporre a ulteriori verifiche, è basata su una struttura gerarchica organizzata nel modo seguente:

a) macro-categoria: viene assegnata in modo univoco ad ogni rivista dall'editore, sulla base della classificazione delle aree scientifiche dell'European Research Council (ERC);
b) categoria: viene assegnata in modo automatico ad ogni articolo sulla base del settore scientifico-disciplinare dell'autore; nel caso di più autori si considera una regola di maggioranza (da definire) oppure regole frazionarie;
c) topics: viene assegnato sulla base di parole chiave e/o parole incluse nel titolo e nell'abstract, sulla base di una soggettazione bibliotecaria a grana intermedia (da definire in dettaglio in seguito);
d) sub-topics: viene assegnato autonomamente dall'autore attraverso la messa a disposizione di una procedura del database, sulla base di un menu predefinito basato su soggettazioni a grana fine.

Nei livelli c) e d) possono rientrare classificazioni operate da istituzioni bibliotecarie nazionali.

Sarà opportuno chiedere la collaborazione delle riviste e il parere delle società scientifiche al fine di identificare gli schemi di soggettazione più appropriati.

6. ANALISI DEI DATI

L'analisi dei dati attiene alle applicazioni necessarie per servire i singoli processi di valutazione, che possono essere fatti a diversi livelli (VQR, ASN, valutazioni interne degli atenei, ecc.). È pertanto necessario garantire la massima flessibilità nel modo in cui i dati sono organizzati, così da consentirne un utilizzo diverso.
Dal punto di vista della struttura dei dati, è importante tuttavia definire quali siano le informazioni più importanti da raccogliere perché tale flessibilità sia effettiva.

Gli indicatori possono essere suddivisi in tre gruppi:

i) Indicatori di produzione
ii) Indicatori di citazione
iii) Indicatori di uso

6.1. Indicatori di produzione

Attraverso gli indicatori di produzione sarà possibile esaminare il numero di articoli prodotti da ogni autore, in ogni intervallo di tempo prefissato, sulle riviste incluse nel database.
L'indicatore di produzione potrà avere due declinazioni:

a) Articoli
b) Altri prodotti (note, recensioni, lettere, risposte etc.)

La distinzione tra le due tipologie di prodotti è concettualmente chiara ma empiricamente problematica. Una soluzione potrebbe essere quella di addivenire ad una definizione formale di articolo (ad esempio a valle del decreto ministeriale che dovrà definire le categorie di scientificità per la Anagrafe della ricerca) e di chiedere agli editori una auto-disciplina nel conferire in *full text* solo prodotti corrispondenti ad articoli o, meglio, di dichiarare al momento della fornitura della pubblicazione l'appartenenza ad una delle categorie individuate.
In alternativa si potrebbe:

- elaborare un criterio di definizione sulla base di parametri controllabili in modo automatico ;
- definire criteri il più possibile oggettivi ma che necessitano di un'analisi redazionale per l'assegnazione di un prodotto a una categoria.

La scelta adottata avrebbe ovvie implicazioni in termini di costo.
Realisticamente in prima applicazione non sarà possibile distinguere tra le diverse tipologie di prodotti.

Nel prosieguo si parlerà esclusivamente di "articoli", lasciando aperta la questione della identificazione dell'oggetto.
Gli indicatori di produzione si potranno riferire alle singole *subject categories* delle riviste, ad altre classificazioni, oppure all'intero database.
Inoltre i livelli di aggregazione potranno essere diversi: singolo autore; dipartimento; università.
Come detto, l'elaborazione di specifici indicatori attiene all'uso della banca dati in fase di valutazione. Nel disegno generale della banca dati è importante valutare se il modello di dati adottato e la ricchezza delle informazioni raccolte forniscono un sufficiente grado di flessibilità per supportare analisi diverse.

La capacità della banca dati di fornire le informazioni necessarie per costruire gli indicatori dipende essenzialmente dalla completezza e accuratezza delle informazioni relative all'affiliazione e alla soggettazione.

Nell'allegato 2 sono proposti alcuni esempi – non esaustivi – di indicatori di questo genere elaborabili a partire dai dati presenti nel database.

6.2. Indicatori di citazione

Sugli indicatori citazioni occorre aprire una riflessione sui seguenti aspetti:

(a) perimetro di applicazione
(b) differenziazione delle citazioni
(c) interpretazione

Circa il primo aspetto, è evidente che il significato delle citazioni dipende dal perimetro di applicazione. Lo stesso numero di citazioni può essere considerato alto o basso a seconda che venga rapportato ad aree ampie, oppure a nicchie specialistiche, a seconda del comportamento in termini di numero medio di citazioni per articolo nella specifica area. Questo fenomeno, ben noto in bibliometria e alla base delle più adottate normalizzazioni, incontra difficoltà nelle aree umanistiche e sociali per una mancanza di classificazioni consolidate a livello di gruppi omogenei di riviste. Nel seguito si ipotizza di definire i perimetri di citazione come *subject category*, con le avvertenze sopra dette, e a livello di insieme delle riviste in classe A per il settore concorsuale. Le conseguenze di diverse scelte di perimetro saranno profonde e richiederanno approfondimenti.
Circa il secondo punto, occorre stabilire se le citazioni vengono considerate in modo uniforme, oppure se è opportuno introdurre una differenziazione. Considerando che a regime il Database includerà sia riviste di fascia A che altre riviste, e che in entrambe tali riviste possono esservi pubblicazioni indicizzate in altre banche dati, potrà essere di interesse introdurre differenziazioni del tipo:

- citazioni da riviste in fascia A appartenenti al corpus della banca dati;
- citazioni dall'insieme delle riviste scientifiche appartenenti al corpus;
- citazioni da riviste scientifiche appartenenti al corpus e indicizzate nella banca dati *i* (*i* essendo una delle banche dati prese in considerazione, per le quali v. *supra*, §3.6),

e così via. Per abilitare l'elaborazione di tali indicatori, la banca dati dovrà contenere, per ciascuna rivista, le informazioni aggiornate relative all'appartenenza riviste alle diverse categorie sopra indicate.
Analogamente, potrà essere rilevante classificare le riviste secondo altre variabili, ad esempio la loro natura internazionale, così da distinguere:

- citazioni da riviste scientifiche internazionali appartenenti al corpus;
- citazioni da riviste scientifiche internazionali in fascia A appartenenti al corpus;
- ecc.

Anche in questo caso, indipendentemente dal criterio di distinzione tra riviste nazionali e internazionali, che può essere controverso, per l'elaborazione di questi indicatori è necessario raccogliere le informazioni necessarie e sufficienti per la loro identificazione. Ad esempio, se il record descrittivo della rivista contiene indicazioni relative alla(e) lingue(e), al luogo di edizione, o alla diffusione, sarà possibile sviluppare applicazioni che consentano di identificare le riviste internazionali, nel tempo, anche in rapporto a definizioni diverse.

Laddove la banca dati sarà in grado, come auspicato, di raccogliere citazioni provenienti da pubblicazioni estranee al corpus (v. di nuovo §3.6), analoghe classificazioni andranno operate per qualificare anche queste.

Infine occorre tenere aperto il tema dell'interpretazione. La citazione nelle aree umanistiche e sociali potrebbe significare:

- adesione
- tributo assegnato ad opere ormai classiche e/o di autori defunti
- rassegna
- dissenso
- rituale accademico

o altri significati ancora. Occorre studiare attentamente le diversità di significati, tenendo presente che il livello di granularità delle informazioni che è necessario raccogliere a tal fine è molto maggiore, in quanto va riferito ad ogni singola citazione e non – come negli altri casi sopra citati – alle riviste.
Al netto di queste scelte strategiche, una possibile mappa degli indicatori è offerta in Tabella 2 dell'Allegato 2.

6.3. Indicatori d'uso e altri indicatori non citazionali

La disponibilità di riviste in formato elettronico ha reso possibile costruire indicatori di uso, che registrano l'evento di download di un articolo, e quindi presumibilmente di lettura, a prescindere dalla citazione effettiva che di esso venga fatta negli articoli successivi.
Gli indicatori d'uso tendono a fornire informazioni non sovrapposte agli indicatori citazionali e per questo sono oggetto di interesse crescente.
La costruzione del Database potrebbe avvantaggiarsi della implementazione di sistemi di conteggio automatico dei download, seguendo lo standard COUNTER dove applicabile.
Tali sistemi si possono applicare:

(a) alle riviste pubblicate da editori, distinguendo tra:
- download a pagamento
- download all'interno di abbonamenti-quadro a livello di struttura (ateneo o ente di ricerca)

(b) alle riviste in Open Access.

Esempi di indicatori di questo tipo sono riportati nella Tabella 3 dell'Allegato 2.

Nella definizione di "altri indicatori" si intende aprire una finestra di osservazione circa i rapidissimi sviluppi in corso negli indicatori non citazionali, quali le iniziative legate ad Altmetrics e gli studi sulla diffusione dei contributi scientifici (academic social networks, impact story, etc.). Le potenzialità offerte dalla analisi di dati in formato digitale relative alla utilizzazione, alla vista, alla circolazione di articoli scientifici sono molto elevate e richiederanno una attenzione nella progettazione di dettaglio del database.

7. DELIVERY: COME FORNIRE I DATI

Il principio generale è la massima apertura nel rilascio dei dati, in funzione degli obiettivi specifici della iniziativa, con modalità *open linked data*.
Il Database risponde a due esigenze fondamentali: valutazione e ricerca.
L'obiettivo di valutazione si concretizza nella fornitura di dati e di indicatori utili a vari soggetti ai fini di successive elaborazioni.
I soggetti interessati all'uso dei dati a fini di valutazione sono, in linea di massima:

- ANVUR

- MIUR
- Università
- Enti di ricerca

Ciascuno di questi soggetti potrà utilizzare i dati e gli indicatori al fine di svolgere valutazioni nei termini assegnati dalla legge (nel caso della Agenzia) o con procedure interne regolate dai rispettivi statuti, ad esempio per la allocazione di risorse di personale o finanziarie.

L'ANVUR si impegna ad utilizzare i dati e gli indicatori del Database nell'ambito delle proprie iniziative di valutazione e nei limiti stabiliti dalle norme.

Il rilascio di dati a fini di valutazione dovrà dunque seguire due canali in parte differenziati:

- pubblicazione di una serie di indicatori e/o messa a disposizione di parte dei dati con modalità aperte
- rilascio personalizzato e riservato di una serie di indicatori, da concordare con le singole Università ed Enti di ricerca, anche nell'ambito di iniziative concordate con i rispettivi livelli istituzionali.

Occorre fin da subito regolare l'utilizzo a fini valutativi da parte di altri soggetti (es. editori, soggetti esteri, fondazioni di ricerca etc.).

L'obiettivo di ricerca si sostanzia nella messa a disposizione di metadati sistematici sulle principali aree umanistiche e sociali italiane, a fini di consultazione, analisi bibliografica, visibilità internazionale. In questo caso si prevede un utilizzo integralmente aperto dei metadati, con modalità di rilascio in *open linked data*.

Per tutti gli articoli saranno resi accessibili gli indirizzi URL permanenti, in modo da indirizzare gli utilizzatori ai *full text*, nelle modalità di accesso relative.

8. BUSINESS MODEL

Il Database ha carattere istituzionale.

Esso è promosso e gestito dall'ANVUR all'interno di una iniziativa concordata con il MIUR e condivisa con i diversi livelli istituzionali del mondo della ricerca e dell'università.

La gestione tecnica del Database verrà affidata a soggetti terzi, dotati di adeguate competenze in ambito bibliografico, bibliometrico e di software, con procedure di evidenza pubblica.

L'ANVUR si impegna a sostenere la fase di startup del Database, fino alla produzione della prima serie di indicatori a fini di valutazione.

Il modello di business della fase a regime sarà definito in corso d'opera sulla base dei seguenti principi:

- La partecipazione delle riviste e dei rispettivi editori non sarà soggetta a forme di pagamento, restando inteso che sono a carico degli editori i costi di preparazione dei file di ingresso nel Database
- La fornitura di dati e indicatori per fini istituzionali e valutativi non comporterà, per livelli base di servizio da definire, oneri per gli utilizzatori

Alla luce di questi principi sono da esplorare diverse soluzioni per la messa a regime, da sole o in combinazione:

- supporto istituzionale ministeriale
- supporto istituzionale condiviso tra vari soggetti pubblici
- sponsorship privata
- fornitura di rapportistica ulteriore rispetto ai livelli base di servizio
- fornitura di servizi personalizzati a pagamento.

Si ritiene che la definizione esatta del modello di business, che resta tuttavia in cima all'agenda della fase progettuale, possa beneficiare dei risultati della fase sperimentale, alla fine della quale i vari soggetti potranno verificare in dettaglio i benefici della iniziativa.